The Journal 2023

Editorial

In the modern day, the word 'competition' has become associated, and indeed synonymous with, antagonism and opposition. The emphasis is on competing *against*. However, if we explore the origins of the word then a very different picture emerges. Competition traces its origins to the Latin *'com'*, meaning with, and *'petere'*, meaning to aim at or seek. According to this definition, to compete is to strive together. The emphasis is on competing *alongside*, as part of a community united in its pursuit of excellence and its aspiration to scale new heights.

This concept of striving together lies at the heart of our philosophy of *Scholarship for All* here at the RGS. We provide a supporting and nurturing environment where boys are encouraged to be aspirational and imaginative, engaged and reflective, and to persevere in our collective pursuit of academic excellence and development of a lifelong love of learning. These are the characteristics and values that lie at the heart of our ethos and teaching approach and are encoded in our five Learning Habits. As a new member of staff it has been a privilege and a delight to become a part of and to begin contributing to this culture of shared endeavour.

The essays contained in this year's edition of the Journal, whilst just a small selection of the outcome of this shared endeavour, do perhaps represent some of the highlights of this collective undertaking. After a COVID-enforced gap it has been a wonderful to witness the return of the Independent Learning Assignment (ILA). As a bespoke part of our curriculum here at the RGS, the ILA offers students in the final term of the Lower Sixth the opportunity to develop further the skills of independent research that are so central to undergraduate learning. Both winning ILA submissions as well as many of the shortlisted projects are included in these pages, on topics ranging from gene therapy to emoji interpretation, economic sanctions to supermassive black holes.

The *Original Research in Science* (ORIS) programme is another unique RGS offering, partnering students with university departments to complete a period of research during the summer holidays between the Lower and Upper Sixth. Extracts from the reports of three of last year's ORIS students are also included, alongside essays that have received recognition in national essay competitions.

The diversity, innovation and excellence contained within these pages are a testament to the power of a shared vision and culture of striving together. I hope that you will agree that they make fascinating reading.

Mrs H Tarasewicz
Head of Scholarship

Contents

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

Red, Yellow, and Blue: Analysing Catalan Independence through Culture

A short-listed Independent Learning Assignment (ILA)

Dev Atara, Upper Sixth

INTRODUCTION

Football fans around the world can taste the atmosphere inside the Camp Nou, even more so when FC Barcelona score to go 3-0 up. However, on the night of the 1st of October 2017, when the team scored against Las Palmas, there were no cheers from the crowd. Earlier in the day, the Catalan government had held an illegal referendum regarding the region's independence, which had led to violent clashes between protesters and police. FC Barcelona closed the stadium, and the match was played behind closed doors. Independence in Catalonia has always been a controversial topic. Since the death of Franco in 1975, the idea of an independent or autonomous Catalonia has always been considered. Years of tension culminated in the referendum of 2017, which has subsequently led to the imprisonment of Catalan political leaders such as Josep Rull and Oriol Junqueras. The referendum is usually linked to the economic crisis of 2008, or the revise State of Autonomy imposed by the governemnt in 2010. The culture movements from the 60's and 70's can help inform and describe influential reasons for independence. This can be noted through cultural, social, and political lens.

Independence in Catalonia has a complicated historical narrative with multiple factors since the era of Franco being shown through in the modern-day elements such as football, music, and literature. So, what I will attempt to show in this project is that through cultural forms like art, literature, and music we can understand critical features of the Catalan independence movement. These combine to give the idea that there is a group mentality, where the population unite around a common cause, which surrounds Catalan independence, visible today following the referendum. Combining unfamiliar cultural forms to understand concepts in greater depth, I will demonstrate how suffering, resistance, progress, and identity are central to our overall understanding of the independence movement. Overall, we will then see how a combination of these factors contribute a sense of group identity within the region. By looking at the art, music, and literature worlds individually and investigating the works of Joan Miro, Raimon, and Joan Sales, I will show you how suffering is a key theme through various cultural forms. Beyond this I will look to compare the idea of identity linking into the suffering. This uses similar forms but in particular the identity and importance of the language and how this can be represented in music and literature. Alongside these more abstract cultural concepts, I will look to use more definite and factual concepts such as resistance where we can see the difference between the development of direct and indirect resistance and how that takes many different forms for different people. However, most centrally I will look at how progress through the Catalan independence movement is represented in culture and how paintings such as The Hope of a Condemned Man show us how we can learn about independence through such a simple piece of artwork. These factors combine together to give the idea that Catalan independence is a group unity movement involving the suffering of people, creating an identity, and leading to resistance and progress with the end goal of independence. This essay will use a range of diverse sources aiming to use first hand evidence alongside my secondary interpretation and wider reading to show you how from Franco to today, Catalan independence has developed, and culture is the best mechanism to show us this with a particular focus on the 1960s and 70s.

BACKGROUND HISTORY

Before we explore the exact changes that occurred within the period, it is important to understand why these changes had such a profound effect. This can be clearly understood by analysing Franco's actions between 1939 and 1960. He fought the Civil War against the Republicans, the side who supported the founders of the Second Spanish republic (1931-36) with a left leaning ideology. After defeating the Republicans in the civil war, Franco adopted a policy of limpieza, translated as cleaning. Franco believed in Spain as a single nation, all united under the Spanish flag and by the use of Castilian, or more commonly known as "Spanish." Imposing this on regions such as Catalonia meant that the regional language of Catalan was reduced to household use as language was repressed. Newspapers such as Pueblo and television broadcasts from the region were all in Castilian and in school and Catalan textbooks were replaced by Castilian versions. Catalan culture was also eradicated as festivals and traditions such as La Diada, the Catalan National Day, were banned by the state, again being replaced by traditional Spanish ones.

Franco's policy was not just limited to cultural genocide, but he also executed over four thousand Catalans between 1938 and 1953, such as Lluís Companys i Jover, former president of Catalonia. Throughout the region there were political dissidents in concentration camps such as socialists, anarchists, and regional nationalists. Several of these prisoners were believers of Republicanism, as they had fought against Franco in the Civil War. They still believed in policies such as regional autonomy, promised by the leaders of the Second Spanish Republic (1931-36). Hence, Franco had them locked up as he felt they posed a threat. Another policy was to take a number of new-born babies from their republican mothers to give to conservative families to adopt. Franco's attempts to eliminate any foreign influence upon the population were severe, especially in Catalonia. These people had suffered so much before 1960 and this illustrates the importance of culture to them. Franco's harsh beliefs nearly eliminated their history, and they were desperate to have it back.

ART

One aspect of culture that depicted Franco's actions was art, more specifically the work of Joan Miró, the most famous Catalan artist of all time. Born in Barcelona in 1893, he lived in Paris and during the Spanish Civil War, he was unable to return home to visit his family due to the conflict. This affected him, and he sympathised heavily with the Republicans. Hence in 1937, he created El

Segador[1], translated as The Reaper for the World Exhibition which took place in Paris that year. The painting shows a Catalan peasant wearing a barretina hat, which represents Catalan identity. The hat was previously worn by Catalans in the 19th century, and children in rural areas during the 1940's and 50's continued to do so. It is regularly seen in traditional Catalan folklore dance, to illustrate the identity of both the dance and the dancer. He shows the symbol of republican resistance, the clenched fist salute, with his face shaped into a cry of despair. The peasant had often been used as a symbol of Catalan nationalism and Miró uses all these symbols to illustrate his support for his home region and his country. Overall, the Reaper is an illustration of the suffering of the Catalan people. Furthermore, Miro uses several expressions such as the pained face, and the fact he uses so many expressions portrays the extent to which his art reflected the hardship faced by the Catalan people. From this we can infer that the Catalan people were suffering under Franco, and it emphasises how his policies had a profound effect upon their way of living. Symbols such as the Barretina hat make it clear that art was being used to mirror the everyday life in the region and the suffering faced on a daily basis. culture frequently referenced the freedoms of the past and artwork served as a reminder of the hurt experienced by everyone.

Miró had an extremely distinguished art career and one of his final paintings, The Hope of a Condemned Man[2], exemplifies the Catalan social attitude of the 1970's. Miró began painting it in 1968 but finished in 1974, and it mirrors the suffering of Salvador Puig Antich, a Catalan activist who was serving time in jail after falsely being accused of robbing a bank. He would later be executed[3].

The painting depicts three spots of colours, all surrounded by thin black lines that never meet. The lines not only represent the tragic end to the young man's life, but also the journey of Catalonia as a region.

However, it is important to note here that Miro did not set out to mirror the experiences of Puig Antich. Speaking in 1974, he said, "The relationship between these paintings and the drama of Puig Antich was not sought at the beginning," Miro initially set out to denote the power of creativity despite oppression, and to resonate with the suffering of the Catalan population. This is exactly what he achieved with his incomplete lines, as we can clearly infer his underlying message. He wanted to demonstrate that whilst Catalan culture had significantly progressed in the 60's and 70's, there was still some work to be done. There were obvious signs of progress such as less censorship in the press, but there was no fully Catalan newspaper. Progress can also be seen through the very fact that Miro could paint this and send a political message. He was not pursued domestically and was able to paint freely, representing that Catalonia as a region had made progress since the early days of the dictatorship.

Overall, the work of Joan Miro clearly demonstrates two ideas in his painting. Firstly, he resonated with the Catalan suffering. The symbolism present in The Reaper exemplifies his sympathy for the Catalan population, seen through both the Barretina hat and the clenched fist. It is clear that the events of the war have upset him, and that he wishes to show solidarity and offer support. Symbols such as the clenched fist and the barretina hat illustrate this, and the extent to which he portrays this is evident. Furthermore, he references the progress made by the region, both directly and indirectly. His incomplete lines and use of colour directly reference the gains made by the region in the 60's and 70's, something which would have please him. Also, the lack of censorship and freedom offered to artists indirectly illustrates that Catalonia had further progressed. These paintings are extremely powerful in depicting the foundations of independence, and the impact they had was significant.

FOOTBALL

Football is a vital part of Spanish culture. It is said to have divided families, but it has also divided the country. The greatest rivalry within Spain is El Clasico, played between Barcelona and Real Madrid. Both are giants of the game, but Barcelona represents Catalonia and the working man whilst Real Madrid is a symbol of the elite. Their fierce rivalry nowadays actually dates back to the dictatorship when Franco made Real Madrid the club of the country, using their success as evidence of the benefits of his policies. This infuriated the Catalan population as it was a further example of suffering, hence they turned to FC Barcelona to resist the actions of Franco.

As mentioned before, language repression was a policy of Franco. A general saying in Catalonia is that during the dictatorship, there were only two places where Catalan could be spoken safely. One was in the privacy of your own home, the second was Camp Nou. During matches fans would sing in Catalan

1. Miró, J. (n.d.). El Segador. Available at: https://catalogo.artium.eus/dossieres/4/guernica-de-picasso-historia-memoria-e-interpretaciones/el-pabellon-espanol-de-la-expos-6.
2. Miró, J. (1974). L'Espoir du condamné à mort (The Hope of a Condemned Man). [Acrylic on canvas] Available at: https://theidlewoman.net/2011/07/31/miro/6d233-joan-miro-hope-of-a-condemned-man-i-ii-iii-1973/.
3. www.akpress.org. (n.d.). Salvador Puig Antich. [online] Available at: https://www.akpress.org/salvadorpuigantich.html [Accessed 7 May. 2022].

and chant for their heroes, safe in the knowledge that they could not be arrested by the authorities due to the sheer number of them. Furthermore, at 17 minutes and 14 seconds into the game, the fans chant "Independencia", a reference to the year 1714, when Catalonia lost its independence in the War of the Spanish Succession. In Catalonia, Barcelona FC acts as a symbol[4] of identity. In the Camp Nou, you can see the phrase "mes que un club" imprinted into the seats, looking down upon the turf. A Catalan phrase meaning, "more than a club," these four words alone illustrate the importance of FC Barcelona for Catalan people. It is not just a club, but represents the hurts felt by the region under the dictatorship. The ardent support from the fans exhibits the importance the club has within the region, and how it politically represents the population.

The general perception of Real Madrid is that it is an elite club and royalty. Its name gives it away, as does the use of a crown on its crest. While any fan might disagree with that sentiment, there is no doubting Franco's involvement with the club. Alfredo Di Stéfano, the greatest goal scorer in Real Madrid's history, was to sign for Barcelona in 1952. However, the Spanish Football Federation intervened and declared the transfer illegal. Shortly afterwards, Di Stéfano moved to Real Madrid. Many Catalans believe that this is an example of discrimination against them, as the state was clearly favouring Madrid over Barcelona on both a footballing level and a national level. Actions such as these antagonised the Catalan people and laid the foundations of an independence bid as they perceived further actions as increased discrimination.

Overall, it is clear that football acts as a mechanism for social tension within Spain. However, the deep-rooted history and animosity between the two clubs has only strengthened the Catalan resolve for independence. Cultural aspects such as football had a pronounced effect as people could identify with something they love so much. Football and Barcelona FC as a symbol are extremely popular with the people and their importance constantly reminds the population of what they have been through. The significance pf FC Barcelona cannot be underestimated, and it holds a clear place as both a symbol of past suffering, but future resistance for the region. The nature of FC Barcelona is such that many people associate themselves with the club, leading to a unified Catalan population with aligned desires.

MUSIC

Another cultural aspect which underwent a revolution during the 1960s was music, with the emergence of Nova Cançó[5]. This was when politics and music interacted, demonstrating the importance of a cultural identity and cultural planning for a nation. For Catalonia, Nova Cançó was extremely important in helping transform society during the last years of dictatorship. Later, I will explore that this was especially prevalent within the younger population, as they expressed their frustration and anti Franco sentiment through the use of Catalan in their songs and metaphorical language to surpass the censorship.

The movement began in the late 1950s as a group was formed by Josep Benet I de Joan and Maurici Sarrahima. They began to compose Catalan songs and invited other musicians to join them through an advert in the newspaper Germinabit. The effort led to return of former prominent singers to the region such as Raimon and Joan Manuel Serrat, both of whom had fled to France in order to escape Franco. This was a watershed moment as they illustrated that Catalan culture was beginning to re-emerge after its suppression, and the region was beginning to regain some of the lost identity.

However, the movement did not stop there and continued to grow. One specific song which resonated with the population was "al vent," written by Raimon[6]. It speaks of a young man looking for peace and finding freedom in the wind, and how the wind will soon pass. The phrase "al vent, "meaning "in the wind," is repeated over and over again in some form of cry, or a proclamation, which illustrates the extent to which this rebellious movement went. The wind is symbolic of the dictatorship and the context of the 1960s. He says that it will soon pass and suggests that the emergence of this organisation was allowing Catalonia to find some form of freedom and regain cultural autonomy. Furthermore, the songs produced clearly represented the hardship endured for the Catalan population, which only further strengthens the group outlook within the region.

It is also worth noting that the song is written and composed in Catalan, which further resonated with the youth population. Singing "al vent" in Catalan became an emblem of the Spanish youth who did not want to live and grow up in the military dictatorship. Strong lyrics and the sense of hurt illustrate that the Catalan population were expressing their desires through music. They had suffered heavily during the dictatorship and used song as a tool of ideological expression against the cultural domination imposed by Franco. The effect of Nova Cançó also extended to other regions and affected other artists. Parallel movement sprung up in Galicia, the Basque Country[7] and Castile, with artists inspired by what Catalonia had done.

Overall, musical expression in 1960's Catalonia contributed heavily to installing a desire for independence in the region. Artists such as Raimon was central to this as they referenced suffering under Franco to connect with the population, meaning that that was a stronger sense of unity in the region as a result. Movements such as Nova Canco were vital as they provided cultural liberty and freedom to the population, and their legacy can still be seen today as we can find Raimon on Spotify and other streaming platforms. This exemplifies the importance of said movements.

LANGUAGE

As referenced before, language suffered heavily under the dictatorship as Franco attempted to repress any representations of regional identity. However, alongside the cultural revival in the 1960s, we saw that language took on a greater prominence within society. One organisation which assisted with this was Òmnium Cultural[8], based in Barcelona. Omnium Cultural, a non-governmental organisation based in Barcelona, was created to promote the Catalan language, and spread Catalan culture throughout the region after the actions of Franco. It wanted to unite Catalans under the flag and language regardless of their stance in the civil war. Overall, it has been a flagship entity of Catalan civil society to defend and promote both language and culture, helping to contribute to the sense of unity within the region

Originally formed in 1961, Òmnium set up with a clear aim, they wanted to increase the use of Catalan in the public sector. However, in 1963 the authorities closed down the organisation and the founders relocated to Paris.

4. Dowling, A. (2018). Culture, language, and identity. In: The rise of Catalan independence: Spain's territorial crisis. London: Routlege, Taylor & Francis, pp.30–55.
5. Hotel Arc La Rambla. (2015). LA NOVA CANÇO (The New Song) - A CATALAN SYMBOL. [online] Available at: https://hotelarclarambla.com/blog/la-nova-canco-the-new-song-a-catalan-symbol/.
6. www.youtube.com. (n.d.). Raimon - Al Vent. [online] Available at: https://www.youtube.com/watch?v=v1hKV1xFXyc [Accessed 11 May. 2022].
7. Rate Your Music. (n.d.). Euskal kantagintza berria - Music genre - RYM/Sonemic. [online] Available at: https://rateyourmusic.com/genre/euskal-kantagintza-berria/ [Accessed 11 May. 2022].
8. Dowling, A. (2018). Culture, language, and identity. In: The rise of Catalan independence: Spain's territorial crisis. London: Routlege, Taylor & Francis, pp.36

They worked as a clandestine network in the French capital from 1963 to 1967, where they began a long legal case against the dictator. They won and were authorised to exist within the region. This is when Òmnium began to influence population and reintroduce the language. Catalan was still not used in public, left out in the media and Catalan journalists were being censored. They clearly wanted to overturn this, and they slowly began to exert a greater influence.

One way in which they helped was by creating and sponsoring various awards and literary contest for works published in Catalan, such as the Premi d'Honor de les Letres Catalanes. They also held festivals of Catalan literature where other awards were given out on behalf of the organisation. In addition, they attempted to re integrate Catalan within schools through the use of campaigns such as Catalan a l'Escola and Delegacio d'Ensenyament de Catalan[9]. By 1975 they had trained over 2000 teachers of Catalan and the language was now officially being taught within schools. In 1975, 66,000[10] students were studying Catalan courses in over 300 schools in Catalonia, with the majority in Barcelona. This represents a strong demographic shift as increased immigration to Barcelona in the 1940s had left the city with a lower level of Catalan users. However, the work of Òmnium had restored the language back into the city and the younger generation felt both Catalan and Spanish. During this period language functioned as an agent for change for Catalonia and lit the metaphorical spark that we saw throughout the period. Its reintegration to the public sphere served as a reminder of what it meant to be Catalan. 80% of the population used it under the second Spanish Republic (1931-36), but the population had been starved of it under Franco. The reintegration of Catalan into the public sector represented a restoration of the Catalan identity throughout the 60s, whilst acting as a mark of progress for the region. The population were now proud to be Catalan openly, and they showed this by speaking the language.

Within the region it was not just culture that expanded in the 1960s and 70s. The re-emergence of movements such as Òmnium also contributed to pro-independence sentiments because they acted as a reminder of what Catalonia stood for. Language provided an umbrella for the entire population to unite under in the face of the dictatorship. The political left, right and centre could join together to combat Franco and aim for goal of independence. Restoring the Catalan identity was critical in establishing a united front for the region. Organisations such as Òmnium were so important because they helped the population realise the importance of language and how they can help to form a bond between people. This social identification was really important because it pulled the population together. Furthermore, the very presence of Òmnium also highlights the fact that Catalonia was now facing less repression as a region. 20 years ago, they would not have been able to exist and the activities they conducted would have seen them persecuted. this was a sign of progress and, Òmnium was not only able to unite the people under language, but it also functioned as a symbol of how far Catalonia had come.

LITERATURE

When we think of culture and how it can be expressed, the general perception is through literature and poetry. Both of these, like several other aspects of Catalan culture, were heavily suppressed by the dictatorship but reimagined during the 1960s. I am going to analyse the works of three authors, Joan Sales, Salvador Espriu and Manuel de Pedrolo. All three achieved different things but contributed heavily to the foundations of the Catalan independence movement as they were pioneers.

JOAN SALES

Joan Sales was born in 1912 in Barcelona and was a political activist from an early age. He was imprisoned for three months at just 15 years old for protesting against the Primo de Rivera dictatorship. When the Civil War broke out, Sales joined the Catalan Military Academy and was sent to the Aragon front. He spent weeks fighting until he arrested for not reporting on family members who had not appeared for military service. He was held in a concentration camp and nearly died of typhus but was eventually acquitted of all charges and returned to the front. He was one of the last Republicans to be defeated by Franco and fled to France, where he spent nine years in exile.

All of this is especially important in understanding Sales' mindset when writing his classic, "Uncertain Glory[11]." He began writing it on his return to Barcelona after his nine years of exile, in 1948. He attempted to have it published in 1955 but the authorities rejected it, claiming it, "expressed heretical ideas often in disgusting and obscene language." Whilst a heavily shortened version was published a year later, it was not until 1971 that Sales published a definitive version. The fact that he was able to publish this in 1971, and faced no restrictions from Franco, represents a substantial change in Catalan society. Just as with Omnium, there was a lot more cultural freedom for writers and social organisations which illustrates the extent to which Catalonia had begun to gain some cultural autonomy. This is a further sign of the progress made domestically for Catalonia.

Uncertain Glory also addresses sensitive themes regarding the Spanish Civil War, which would have been unimaginable during the first part of the dictatorship. The novel narrates the experiences of four young Catalans at the front and the rear guard of the republican efforts[12]. He traces their movements during April of 1938 when the Aragon front falls, leading to the fall of Catalonia to Francoist forces. Sales was inspired by his own experiences and uses the book to illustrate the horrific actions carried out by the nationalists. Sales depicts Barcelona as an inward looking, metaphysical city which mirrors the suffering of characters such as Trini, a young woman who is converted to Christianity and attends clandestine masses at night. Sales uses the book as a political message and draws upon the suffering of Catalonians in the war to emphasise the progress of the region up to 1971.

Novels such as uncertain glory provided useful insight into Catalonia in the 1970s. The region has made considerable progress, illustrated by the lack of censorship and freedom when publishing. Furthermore, the direct condemnation of the civil war is one of the first instances where we see Catalonia resisting. Society was not only becoming more progressive, but more brazen in what they would choose to say. The increased awareness of identity and progress in certain sectors was evident, and literature is a perfect example of how the population can be inspired by culture. The work of Sales fostered a group mentality, helping to strengthen the independence drive.

Books such as Uncertain Glory are vital in analysing the progress of Catalan society. The population could now openly talk about the horrors of the regime and the war before, which leads me to think that the society as a whole had not only become more progressive but also more resistive. They understood what

9. Dowling, A. (2018). Culture, language, and identity. In: The rise of Catalan independence: Spain's territorial crisis. London: Routlege, Taylor & Francis, pp.37
10. Dowling, A. (2018). Culture, language, and identity. In: The rise of Catalan independence: Spain's territorial crisis. London: Routlege, Taylor & Francis, pp.38
11. Incerta Glòria (Uncertain Glory). (2017).
12. www.themodernnovelblog.com. (n.d.). Joan Sales: Incerta glòria (Uncertain Glory) – The Modern Novel. [online] Available at: https://www.themodernnovelblog.com/2018/03/15/joan-sales-incerta-gloria-uncertain-glory/ [Accessed 15 May. 2022].

Franco had done and how it was wrong, and Sales clearly reflects this in his novel. Anti-Francoist literature would lend itself to the independence movement as was an expression of Catalan sentiment post 1960 and represented the desire for freedom.

ESPRIU

Salvador Espriu is another writer whose literary effects resonated with the Catalan population. Like Sales, he was also born in Catalonia and had a role in the Spanish Civil War. He did not fight on any front, but instead assumed a role as a military accountant. Through his works, we can see that the War had a profound effect upon him as he questioned the integrity of man and considered whether there was hope for the future of both Spain and Catalonia.

One aspect of Espriu that is extremely important is his use of Catalan. He was one of the few writers in 20th century Spain to fully write their works in Catalan[13]. The magnitude of this achievement cannot be ignored. As we observed earlier with the work of Omnium and the music of Raimon, using Catalan both united the population and acted as a symbol was resistance against the dictatorship. Espriu Is no different but he would resist in indirectly. His sole intention was to write, and his inadvertent use of Catalan would have served as a reminder of the power of language. During the time of the dictatorship Espriu persevered against the societal bounds placed upon him and preserved the Catalan language. His desire to write in Catalan was what led him to poetry as it did not need as much space as prose. That meant it was easier to overcome obstacles of censorship from the dictatorship, letting Espriu write freely.

In his first book of poetry, Cementiri de Sinera (Sinera Cemetery, 1946), Espriu reflects on the suffering of Spain in post war Society. He speaks about "lost days and sons" and Spain destroyed by war, which he called Sinera. He references Sinera several times throughout the work and uses Cementiri de Sinera to establish themes such as hurt and the dead. Espriu reflects significantly upon the bleak landscape of Catalonia in the 1940s within the stranglehold of the dictatorship, using his literature as a means to mirror the hardship faced by the population. In 1948 he wrote the play Primera historia d'Esther, which also referenced Sinera. It has been described as a testament to the Catalan language at a high point of Catalan literature after the devastating effects of the civil war. Espriu also speaks of hope and juxtaposes the damaged world of Sinera with the potential of man. He says, "remember that the mirror of truth was shattered into tiny fragments, yet each bit holds a spark of true light." Espriu uses the post war period to reflect upon the horrors of the war, again referencing the theme of suffering. His reminder of the past would have served to unify the population due to their shared experiences.

Finally, Espriu published La pell de brau in 1960, his best-known work. It involves themes of national, racial and cultural identity which helped to inspire the younger generation of artists to speak out on social and political issues. Its frankness was exhibited by the use of "Sepharad" the Hebrew word for Spain. Espriu uses it in a context to suggest a conflict between Catalonia and Spain and its difficult relationship. He openly addresses the grievances of Catalonia and explain it suffering. This expression from Espriu is testament to how the Catalan population in the 1960s and 70s wanted to address their grievances and their difficult history and not be punished. It is also worth mentioning that Espriu's effect and legacy can be seen through the songs of popular singers such as Raimon, who based their lyrics around his work.

Overall, Salvador Espriu is one of the pioneers of Catalan literature as we know it today. His bravery to continue to write in Catalan whilst referencing both the sentiments of the population and the difficulties faced in the past was so important and he helped to transform literature around this time. His effect can be seen in other areas society such as music, which only strengthens his standing. Promoting ideas such as brotherhood was critical during this time and alongside referencing shared experiences, would have helped her create a group mentality and the idea of a Catalan identity within the region.

MANUEL DE PEDROLO

Manuel de Pedrolo is the least known writer amongst the trio. However, his literature is no less important, and continues to be taught in Catalan schools today[14]. He was born in 1918 in Catalonia and was another writer who devoted himself to writing in Catalan alone. Therefore, we can praise his bravery in the same way we praised that of Espriu, as he was able to stand up against the regime and ensure the language did not perish. This increased the relevance of the Catalan identity as more and more people identified with it. He also cultivated virtually every literary genre but is best known for his large production of prose and more specifically his sci-fi novel, Mecanoscrit del segon origen (Typescript of the Second Origin).

Pedrolo's name will always be associated with this bestseller, his most popular work. It is the bestselling novel of all time written in Catalan and has been translated into 11 other languages, just showing the extent to which it has been shared worldwide. It tells story of Alba and Didac, both of whom live in a town in Catalonia called Benaura. After extra-terrestrials destroy all of the human race except the two of them, they are forced into survival. On several occasions they barely survive attacks, leading them to mature quickly and learn survival skills. Interestingly, they learn value of books and explore the whole of the region. This helps them learn about everything that has happened to them such as illness or all the tactics of the enemy. What makes this book so remarkable is that despite what seems like a weird plot, it manages to resonate on every level with Catalans with its deeper message. The ability to overcome obstacles and preserve in such adversity gave hope to many Catalans. This contributed to the increased resistance we saw in this period, as the population felt compelled to unite as a group and act against Franco.

One such aspect of the book that appeals to many is it breaking down of barriers between races. Alba is described as," an olive skinned 14-year-old," whilst Didac is a," 9-year-old black boy." Whilst discussing their racial differences, Alba says, "we're the last white and the last black, Didac. After us, people no longer think about their skin colour." The racial harmony between the two children is well accepted by Catalans as their past had been characterised by conflicts between differing groups of people, be that ethnically with the Muslims and Christians or politically with the Republicans and the Nationalists. Therefore, this message was particularly important to Catalan society as it suggested that the future could be characterised by a sense of brotherhood and fraternity. It is also worth mentioning that the barren landscape was reflective of Catalonia after the war, another factor that resonated with the population. Pedrolo clearly understood Catalan society in the 1960's and his novel reflects theme. Preaching themes such as brotherhood and resistance would have unified Catalans, which in turn strengthened the independence drive.

13. Montañà, P. (n.d.). Salvador Espriu: one of the greatest Catalan writers of the 20th century. [online] www.catalannews.com. Available at: https://www.catalannews.com/culture/item/salvador-espriu-one-of-the-greatest-catalan-writers-of-the-20th-century [Accessed 17 May. 2022].
14. Goodreads. (n.d.). Typescript of the Second Origin. [online] Available at: https://www.goodreads.com/book/show/36567968-typescript-of-the-second-origin [Accessed 19 May. 2022].

Overall, 1960s Spanish literature underwent a notable change that clearly reflected the views of the people. Sales' direct and effective illustration of the civil war as destructive shows that society had recognised the wrongs in the past and was beginning to rebel against the dictatorship. The work of Pedrolo in creating not only a bleak landscape but a sense of teamwork between Alba and Didac reflected the more progressive and united society. Finally, Espriu painted a positive picture for the Catalan population whilst ensuring that the legacy of Catalan would not be forgotten. These three writers all come together to create a contemporary style of literature which would rejuvenate eyes the Catalan population as they fully accept what happened in the past and realise the desire for autonomy and independence.

GENDER

All of the factors above have contributed to the group mentality within Catalonia in the 1960s. By referencing ideas such as resistance, suffering, identity, and progress, it is clear that the Catalan population felt united in their desire for independence. the rise of feminism in the region perfectly exemplifies the significance of the effect these factors had had. However, feminism under Franco is split into two parts. The first wave of feminism was characterised by women attempting to rebuild and improve their lives, despite the efforts of the regime. The 1960's was a period when women managed to better their rights, after the regime had attempted to control their movements and force many into being housewives. In the 40's women were given "La Guía de la Buena Esposa", which detailed how to be a good wife. Their husbands also controlled their finances and travel, furthering emphasising the lack of rights they had.

The mid to late 1960's in Spain saw a notable change in the role of women for Spain[15]. Greater contact with foreign ideas as a result of increased immigration and tourism, stronger employment chances for women and better economic reforms all meant that women were not as oppressed as they had been. However, it is worth mentioning that feminism under Franco during this time was not unified. Despite this, all feminists felt that Spain needed greater equality and needed to defend the rights of women more. Overall, feminism moved from being about individuals to being about the collective group. This exhibits the sentiments of a wider Catalan population at this time, as despite political differences people bonded together to face up to Franco and push for independence.

Despite the united nature of feminism, women still struggled to meet freely. Feminist groups and women's organisations had been formed in the early 1960's, but the regime only made these legal in 1964. This was a significant step regarding the rights of women, but in reality, truly little changed. Groups still met clandestinely and with few numbers as the authorities would still shut down groups that had any reference to "anti-regime" propaganda. However, the changes in social attitude in the 60's had an effect on future decades as women were not portrayed in such a negative way by 1970. The comic community illustrates this. More women were seen as writers and artists, and comics did not depict women as passive or sexual beings. This rise in feminism can be linked to independence as women in Catalonia began to pursue happiness and greater rights. This increased autonomy for themselves not only allow them to reflect on what had happened previously to their rights but hope for a better future

Overall, it is clear that feminism under Franco changed during the dictatorship. Women clearly had more rights, exhibited by the greater job opportunities and economic change. However, the feminism movement was symbolic of the changes in Catalan society in the 1960s. It was influenced by factors such as art, literature, and music as these cultural elements had managed to unite the population. The sense of identity and group mentality was stronger within the region and women felt they were able to challenge stereotypes and confront oppression. The emergence of women's organisations is a clear example of this, and their very presence was a mark of progress resulting from increased confidence. Feminism was clearly affected by Catalan culture and its revolution throughout the 1960s mirrored that of Catalan society.

CONCLUSION

It is clear to see that there are several cultural elements which have contributed to Catalan independence. They all speak about similar themes, and each play a part in helping the people realise the importance of identity. By analysing various cultural expressions from the region during this time, I have demonstrated that the key themes are critical in helping form a group identity within Catalonia, which continues into the modern day.

One principal factor regarding the foundations of Catalan independence is suffering. The suffering of the Catalan population under the Franco regime were now openly addressed in society, meaning that the region was able to unite due to the shared nature of these experiences. Cultural movement such as art, music and literature all identified and resonated with suffering, furthering the group identity felt. In particular, Joan Sales describes the fall of Catalonia to Franco's forces in the late 1930's and he directly references the horrific actions of the Nationalist troops. Sensitive themes were now being discussed freely, and this is evident through the works of Raimon. His song "Al Vent" suggests that a new wind will come through Catalonia soon and that better days are coming. Not only does he provide hope, but he too reflects on the pain of the War, against illustrating that societal bounds were not as constricting now. Overall, suffering was a key part of Catalan identity, and played a crucial role in helping the Catalan population free themselves from the regime. The Civil War had affected everyone and the ability to openly address these events would have joined the population together.

Another key theme is that of progress, which can be seen through the artwork of Joan Miro in 1974, the reintegration of language into society by Omnium Cultural, and Manuel de Pedrolo's Typescript of Second Origin. Each of these three play a vital role in demonstrating the development made by Catalan society during this period. For example, both Miro and Pedrolo experienced significant freedom when publishing as artists were able to directly address suffering and writers were able to publish in Catalan. In addition, Omnium's ability to educate the younger population is further testament to the advancement made by the region in this time, as Catalan had been heavily suppressed in the past. The extent to which progress can be viewed through several cultural forms shows how it was a common theme at the time and formed a critical part of the Catalan identity. Progress was important for Catalonia as it provided hope for the population. Seeing Catalan spoken in the streets safely, reading works that promoted ideals of brotherhood and friendship all would have contributed to Catalans feeling prouder of their identity and hopeful for the future. From this, it is clear that progress had been made.

It is also important to mention the importance of resistance in Catalan society during this time period. The 60's and 70's saw greater resistance to the regime, which can be identified by looking at cultural factors such as football and literature. FC Barcelona plays a critical role in helping understand how resistance was portrayed. The popularity of football within the region means that the population identifies itself with the club and what it does. For example, the club directly opposed the regime by allowing people to chant in Catalan, knowing that they could not be arrested due to the huge volume of fans present.

15. Barcelona-Home (2019). Women throughout the history of Spain. [online] Barcelona-Home. Available at: https://barcelona-home.com/blog/position-women-spain/.

Actions like these opposed the regime and sat well with the population, who then continued to speak Catalan. Furthermore, Pedrolo references resistance through Alba and Didac as they bravely resist the alien forces and create a better world. Overall, both literature and football preach the idea of resistance. This would have been critical in illustrating to the Catalan population that it is possible to resist the regime and look to a hopeful future. This would help to link Catalans together and ensure that the population was united.

One factor that is heavily mentioned through culture is identity, and how Catalans came to feel more Catalan as a result of culture. This was important as it allowed the population to come together through identity and helped to lay the foundations of independence. Joan Miro in The Reaper explicitly references Catalan identity. He uses the Barretina hat as a symbol of Catalan identity, linking the freedom of the past to the 60's and 70's, illustrating that Catalonia possessed its own identity that people should be proud of. Another important cultural element was language, which helps to serve as a reminder of what Catalans stood for. By promoting the use of Catalan, they managed to unite the population politically and the use of this social identification was really important because it brought everyone together. It helped remind everyone of Catalanism, and what the region used to stand for. Overall, Catalan identity re-emerged during this time period and was critical in not only uniting the population but reminding society of the past. Realising what it meant to be Catalan was critical in helping lead the drive for independence.

However, all of these factors do not stand alone in helping lay the foundations of Catalan independence. Instead, they combine to help form the group mentality that has continued over several decades. the recurring theme throughout these factors is that they all draw upon ideas that the whole population could relate to. Be it suffering through the civil war, or attempting to resist the regime, these factors were critical in helping bring together the population. The importance of unity is that it allowed the population to not only face up to the regime together but define their sense of self and social identity and achieve goals that might have eluded them if they worked alone. For example, Catalonia was crucial in ensuring that after the end of the dictatorship, regional autonomy was one of the priorities for the incoming Spanish government. These cultural elements will have inspired this but also incorporated all areas of society and the population, ensuring that everyone united around a common cause.

Catalan culture and its effects clearly continue into the modern day, even inspiring the younger generation today. Artists like Raimon are frequently listened to on Spotify, and as mentioned before, Pedrolo's literature is still taught in secondary school today. Furthermore, literature continues to be taught, art continues to be viewed and the significance of these factors has not diminished I believe that the importance of this cultural movement in laying the basis for Catalan independence is heavily understated, with a general focus upon politics and economics. However, it is clearly shown that this movement has affected the overall Catalan independence movement, as its wide-ranging effects continue to inspire and provide hope to the population nearly 60 years after being released. Culture frequently reminds the population of what it meant to be Catalan, and this identity is central to the desires of Catalans. They wish for their own state, and culture clearly helped them push to achieve it.

Revolutionising the process of evolution
Genetic modification – from curing Huntington's disease to the creation of X-men

A short-listed Independent Learning Assignment (ILA)

Alexander Atkinson, Upper Sixth

INTRODUCTION

As a young child, I was enraptured by Marvel films and superheroes; "superhumans" who possessed superpowers. In some, these were genetic modifications which enhanced their capabilities to beyond those of an average human. Some of these superpowers are complete fantasy, but others, harnessing our knowledge in science, might soon be a possibility.

While the dawn of flying humans and the ability to withstand extreme conditions (e.g. space) might seem far-fetched, genetic modification is a relatively new branch of science, which has a great scope for making some of these impossibilities, eventually, conceivable. One of the branches of this developing science is gene therapy, which involves the correction of dysfunctional genes to treat or cure a disease. Gene therapy has brought us one step closer to being able to engineer a 'perfect' human with its potential to treat various diseases including the most obscure and painful genetic diseases. While there are thousands of diseases which could be potentially treated, I would like to focus on Huntington's disease, a neurodegenerative disease, which results in dementia, life-threatening depression and a loss of the capacity to control the body's movements. The most frightening aspects of this disease are that carriers do not know when the disease will begin to take action and the disease's inheritance pattern which can cause high prevalence in families.

In my Independent Learning Assignment (ILA), I hope to conclude whether or not Huntington's disease is able to be successfully treated using gene therapy techniques, taking into account the current ethics surrounding gene therapy techniques, their history, their application and the science which makes these techniques possible.

GENETIC MODIFICATION – A GROUP OF TECHNIQUES

To understand the basis of what genetic modification is, a definition is key. However, while researching I came across a broad range of definitions where terms like genetic modification and genetic engineering were sometimes used interchangeably. For example, the majority of sources suggest that both terms refer to "the introduction of new traits to an organism by making changes directly to its genetic makeup" (1). Other sources propose that the definition above holds true for genetic modification, although genetic engineering specifically "works by physically removing a gene from one organism and inserting it into another, giving it the ability to express the trait encoded by that gene" (2).

With this in mind, it was difficult for me, having covered such a broad range of definitions, to put forward my own. However, I finally concluded that genetic engineering is "a type of genetic modification, which involves the process of transgenesis – the introduction of a gene from one organism into the genome of another organism – and results in the formation of a genetically engineered organism (otherwise known as a transgenic organism), which can express this new trait beyond natural capabilities." This leads me on to my definition of genetic modification, which is used as an umbrella term for "a group of techniques which are used to deliberately alter the genetic composition of an organism."

The techniques used to genetically modify an organism include: genetic engineering, somatic gene therapy, germline gene therapy, RNA therapies and gene editing. In order to address the question regarding the treatment of Huntington's disease, I must first research the history, the mechanisms of science, and the feasibility of gene therapy which I present in my ILA.

IS GENE THERAPY AN EFFECTIVE TREATMENT FOR HUNTINGTON'S DISEASE?

GENE THERAPY – A BRIEF HISTORY

Gene therapy can be defined as "the treatment of disorder or disease through transfer of engineered genetic material into human cells, often by viral transduction" (3). It is a new form of medicine which has been fairly successful over the past 30 years, although it largely remains in experimental research in laboratories with relatively few medical applications. I shall briefly outline the most significant events in gene therapeutics history, to underline the development in this field of medicine.

The first paper regarding gene therapy was written in 1972 when Theodore Friedmann and Richard Roblin published their work titled "Gene therapy for human genetic diseases?" (4). This paper highlighted the potential for treating people with genetic disorders using genetically modified DNA, although it advised against the practice while it was still in its very early theoretical stages.

Eighteen years later, the first gene therapy trial launched on 14th September 1990 (5), when an infant girl was successfully treated using gene therapeutics. The girl, Ashanthi DeSilva, lacked an enzyme called adenosine deaminase (ADA), an essential molecule, which is generated to break down deoxyadenosine. If this is not hydrolysed, the deoxyadenosine is toxic to lymphocytes and can cripple the immune system of a person. Having extracted some of DeSilva's blood cells, working copies of the ADA gene were inserted into them by a viral vector. The blood cells were returned to DeSilva's system, where the immune system recognised the blood cells as self-cells, eliminating the risk of rejection. The success of this treatment spurred further trials of a similar kind

for severe combined immunodeficiency.

However, gene therapy does not come without its risks. Gene therapy suffered many setbacks in the first decade it was launched, including a death in 1999 and several patients developing cancer following treatment in 2000. This was caused by problems with the viral vector – which delivered the new genes to the T-cells – which had also had the side effect of activating an oncogene, triggering leukaemia.

2012 was a year of monumental advancement for gene therapy due to the discovery of a new gene-editing technology called CRISPR, which stands for "clustered regularly interspaced short palindromic repeat" (6). The relevance of CRISPR to gene therapy is that it can be used to disrupt a targeted gene, or if a template DNA strand is used, can insert a new short sequence of DNA in a precise location (6), thereby editing the harmful gene. The basic idea behind this is to extract cells from a patient, edit them using CRISPR technology, and replicate these new cells so that they can be reintroduced into the patient. However, I will go into more depth when I discuss the application of CRISPR in the treatment of cancer.

The CRISPR mechanism, which was first discovered in bacteria, allows them to recognise foreign nucleic acid sequences of invasive species (e.g. viruses), and defend against them using enzymes which target specific regions of the foreign DNA to degrade it. Essentially CRISPR DNA is repetitive DNA sequences with "spacer" DNA sequences in between the repeats that exactly match viral sequences. Along with CRISPR's discovery were several proteins, of which, Cas9 (CRISPR-associated protein 9), plays a vital role in the defence system. This enzyme, when coupled with guide RNA, which is complementary to the specific target sequence, cleaves the foreign nucleic acid sequence and saves it as a "memory" for fighting off future infections.

The mechanism of CRISPR in bacteria occurs in 3 steps as follows:

~ Adaptation – "spacer" DNA which is acquired from an invading nucleic acid sequence is integrated into the CRISPR repeating sequence.

~ Expression – the new CRISPR sequence is transcribed into crRNA (chromium), and guide RNA, which is complementary to the foreign sequence.

~ Interference – Cas-9 nuclease is guided to the targeted location on the invading nucleic acid sequence and cleaves it.

CRISPR made the headlines again in 2020 when Emmanuelle Charpentier and Jennifer Doudna were awarded the Nobel Prize in chemistry, for their discovery of the bacterial immune mechanism, and for their work in transforming it into a simple gene-editing tool which has high precision and effectiveness. This revolutionary technology effectively allows scientists to rewrite the genetic code in almost any organism, giving it a range of applications from curing genetic diseases to expressing advantageous alleles in crops for a higher yield.

THE TWO TARGETS OF GENE THERAPY

In general, there are two targets of gene therapy: somatic cell therapy and germ line cell therapy. Somatic cells refer to cells that are not destined to become a gamete, and whose genes will not be passed on to future generations (i.e. a body cell), whereas germ line cells are the cells from which gametes are derived (7). Whilst gene therapy can be performed via either of these methods, the results differ drastically. Germline editing can be performed to induce heritable genetic changes, to prevent the passing on of hereditary diseases through a family. Somatic cell therapy is more targeted to each patient who is suffering from a genetic disease and will not induce any heritable changes. Despite the enormous benefits germline cell therapy could provide, somatic cell therapy is far closer to becoming an everyday form of medicine (8). In addition to their different applications, the ethical implications surrounding both treatments differ drastically. Most scientists consider somatic cell therapy an acceptable practice; whilst changing the reproductive cells is considered unacceptable as it alters the gene pool of the human species.

THE PROCESS OF GENE THERAPY

So what does the gene therapy process entail and what are the hurdles to overcome?

There are several methods used in gene therapy, with the basic idea underlying all of them being to correct the defective gene responsible for the disease. These methods include:

- Inserting a new, functional gene into a non-specific location, to replace the defective gene.
- Using recombination – the process of exchanging genetic material between different organisms (9) – to swap a defective gene for a functional one.
- Through regulation of gene expression and transcription, thereby affecting its function.
- Using gene-editing tools such as CRISPR-Cas9 to alter the genome at a specific location, with the use of enzymes which can cut portions of the faulty DNA strand and enables the insertion of replacement functional DNA.

Once genetic diseases have been identified and diagnosed, the defective gene can be replaced with a functional copy (called a transgene) that expresses a trait correctly and suppresses the disease-causing gene's expression.

The steps taken in order for successful treatment by gene therapy include:

1. Identifying and characterising the defective gene
2. The use of an effective system to get the functional gene into the correct site in the patient. This includes identifying a suitable vector (explained below) as well as the method used to physically deliver the gene (e.g. inhalation or injection).
3. The expression and permanent integration of the corrective gene into the target cell's genome.

This can be achieved via several methods:

- Gene replacement therapy – This is the most ideal result of gene therapy in which the inserted gene replaces the defective gene and recombination between both genes occurs.
- Gene addition therapy – This method is still an effective approach and does not require the exchange of gene sequences (recombination), because the inserted gene works in parallel with the target cell's genes. While this method will still cause the defective gene to be expressed, the introduction of a new functional gene reduces its expression.
- Antisense mRNA – This involves a reversed copy of the gene which is used to produce mRNA in the antisense configuration (a complementary configuration) which is able to bind to the mRNA of the defective gene and prevent its translation. This means the defective gene cannot be read by a ribosome so the proteins that it codes for cannot be expressed.

4. Another factor which needs to be considered is the target cell itself. There are two approaches to this: in vivo therapy (which involves all of the steps above) or ex vivo therapy. In these situations, affected cells are removed from the patient's body and are treated outside the body before being reintroduced into the patient.

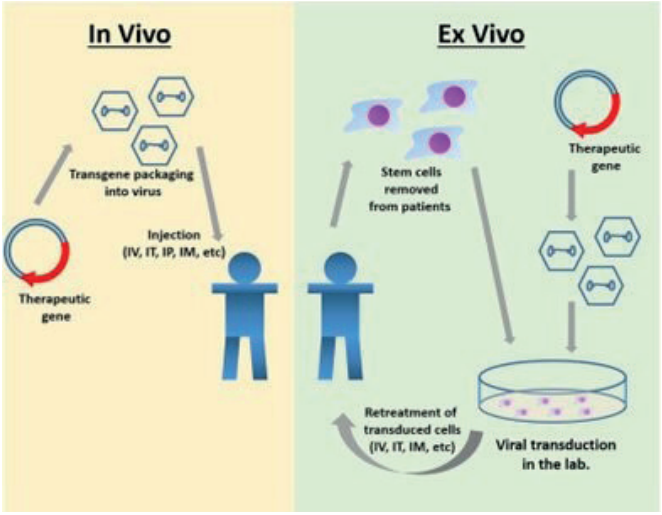


Figure 1 - Differences between In vivo and Ex vivo gene therapy https://www.researchgate.net/figure/Strategies-of-in-vivo-gene-therapy-and-ex-vivo-gene-therapy-In-vivo-gene-therapy-on-the_fig1_322970469

There are, of course, several hurdles to overcome when performing gene therapy which I must address.

1. GETTING TRANSGENES INTO PATIENTS: VECTORS – THEIR ADVANTAGES AND DISADVANTAGES

Firstly, in order to overcome the obstacle of delivering the correctional gene into the target cells, the two broad categories are via either a viral vector or a non-viral vector. A vector, in molecular biology, can be defined as "a carrier molecule which is used to deliver a specific sequence of DNA (a gene) into a targeted cell".

The main viral vectors which are associated with gene therapy and have been developed are based on retroviruses, adenoviruses and adeno-associated viruses (AAV). Scientists and clinicians have to take care that there are no operational viruses which are being cloned/developed as this could affect the patient even more. Viruses serve as powerful delivery vehicles due to their efficiency in infecting other organisms in nature. While a viral vector offers this great benefit, there are several disadvantages which must be taken into account.

Firstly, the success of any virus is dependent on how the body's immune system will respond. Any foreign material which enters the body can easily be recognised providing opportunities for the immune system to rid the body of the gene therapy. As well as this, some patients already have had previous exposure to the virus and therefore have more immunity; in response to this, scientists and clinicians have to carefully select a virus to which patients do not possess additional antibodies. More concerningly, the detection of foreign materials in the body could provoke a severe immune response which further harms the patient.

Secondly, current gene therapy treatments require the administration of large numbers of viral vectors in order for the treatment to be successful. One reason for this is the restricted number of cells that viruses target: this means that a large proportion of them must receive the functional gene. If the number of viral vectors in each dose can be lowered, the cost of gene therapy, which is very expensive (which partly explains its limited number of applications thus far), can be lowered significantly as well. This would make the treatment more accessible for patients and lowers the risk of introducing vast numbers of viruses into the body. While these viruses are harmless, the risk of a severe immune response remains, and an investigation into the deaths of three patients receiving high-dose AAV vector therapies in 2020 is being conducted (10). These unfortunate deaths highlight the need to reduce the number of viral vectors administered to patients.

The third challenge which needs to be considered is controlling the functional gene's expression once introduced into the patient's body. This determines how successful the treatment is and requires important planning and risk assessment because once gene therapies in vivo have been administered the expression of the gene cannot be controlled by clinicians. To increase success rates, viral vector cargo often also contains regulatory elements such as promoters and enhancers. Promoters are DNA regions with short regulatory base sequences that code for RNA polymerase transcriptional machinery (11), while enhancers stabilise or increase the activity of RNA polymerase by providing binding sites for proteins that help activate transcription. The binding of such proteins causes the DNA shape to change and brings about the interaction between the activator proteins and transcription factors bound to and produced by the promotors; this leads to an increased transcriptional rate and therefore more RNA is produced (12).

METHODS OF IMPROVING SUCCESS RATES OF VIRAL VECTOR THERAPIES INCLUDE:

- Improved vectors
- Improved capsids
- New types of cargo delivered by vectors

1. Improved vectors – This has two main aims: reducing the immunogenicity [the ability of foreign material to provoke an immune response in an organism (13)] of the viral vector and improving the expression of the transgenic DNA. Ways of achieving this include: localising gene transfer methods and suppression of the immune response through pharmacological means (e.g. preventing stimulation of lymphocytes or depleting B and T cell levels) (14). Furthermore, including regulatory elements is one strategy for improving the expression of the transgene. Some regulatory elements can reduce immunogenicity by limiting the expression of the transgene in specified cells, for example in cells that promote an immune response (10). One method of achieving this is through the use of microRNA-target sites which guide proteins to specific mRNA strands to repress their translation (15).

2. Improved capsids – The viral capsid is fundamental to the survival of viruses. It determines the target cells of the virus, the immunogenicity of the virus, and the efficiency of cell entry. Humans have acquired immunity against many types of capsid which affects the immunogenicity of the vectors. This is why some of the COVID-19 vaccines have been developed using adenovirus variations from chimpanzees and gorillas (10). Using capsids to which humans do not have pre-existing immunity could be revolutionary in reducing the number of viral vectors administered.

3. New types of cargo – The cargo delivered by a viral vector often refers to the transgene which it delivers to the target cell. However, it can also refer to the complementary molecules which can regulate transgene expression. These include regulatory molecules but can also refer to vectorized antibodies. The use of viral vectors filled with gene encoding antibodies which are able to cross the blood-brain barrier and deliver the contents into the central nervous system is currently being explored to treat Alzheimer's and Parkinson's (16). The expression of the transgene in the vector codes for the production of antibodies which are specific to certain diseases.

Transgenes can also be delivered by non-viral vectors. The broad types of nonviral vectors used include:

- Lipid-based vectors
- Polymer-based vectors

1. Lipids consist of a positively charged "head" group which is able to bind with the negatively charged phosphate groups in nucleic acids. This enables lipids to be effective nonviral vectors and they come in various forms such as: liposomes, solid lipid nanoparticles, or lipid emulsions. Compared with other vectors, lipids are biodegradable and less toxic, and due to their polar nature, they are able to carry hydrophilic or hydrophobic substances.

While liposomes present these advantages, they are less efficient than viral vectors and restrict the length of DNA encapsulated due to their small size.

These vesicles are composed of a lipid bilayer (that can carry lipophilic substances) which surrounds an aqueous compartment which can contain hydrophilic substances. The phospholipid bilayer has a similar structure to mammalian cell membrane which reduces its toxicity and reduces its immunogenicity when used in human drug delivery (17).

Furthermore, drugs stored in liposomes are less likely to succumb to oxidate degradation which increases the chances they will reach the target cell. The cell surface membrane properties of these vesicles can be altered to include glycoproteins which improve the targeting capacity of the vector.

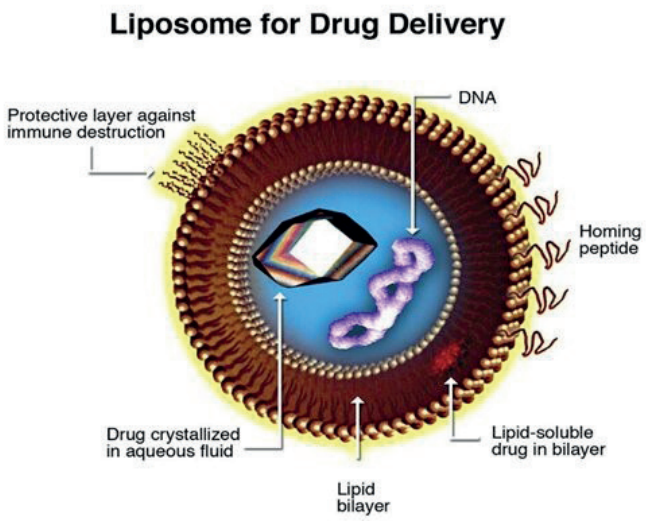


Figure 2 - Structure of a liposome used in drug delivery https://www.europeanpharmaceuticalreview.com/news/39040/high-pressure-homogenisation-the-next-generation-of-drug-delivery-liposomes/

Although they look similar to liposomes, lipid nanoparticles usually contain their cargo in a nonaqueous core and not all of them have a continuous bilayer. Attached to the phospholipid bilayer "head" groups are covalently bonded groups called PEG (polyethylene glycol) linear synthetic polymers. They help improve the nanoparticle's circulation time (and therefore the chance that they reach the targeted cell) by shielding them and preventing blood plasma proteins from absorbing into the bilayer. PEGylation (the process of covalently bonding PEG to the cell membrane/outer layer of a molecule (18)) also reduces the tendency of the nanoparticles fusing with each other. This tendency arises from the instability and large surface tension of small nanoparticles and their fusing can result in the loss of the therapeutic cargo.

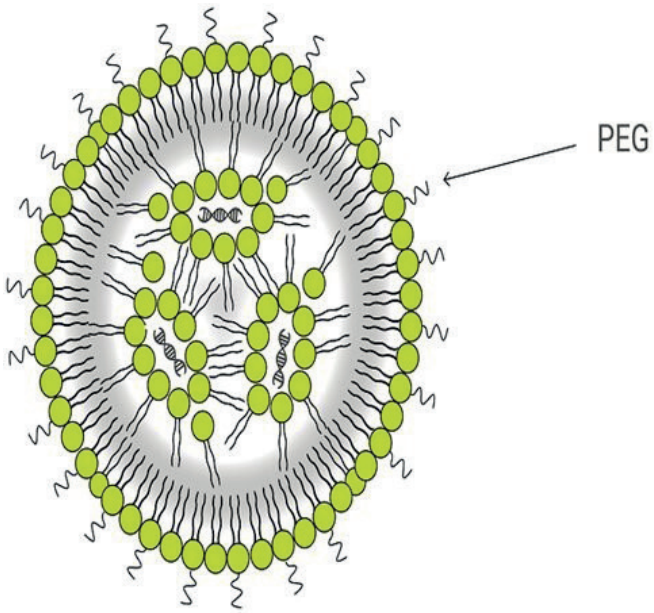


Figure 3- PEGylated lipid nanoparticle diagram https://www.biochempeg.com/article/122.html

2. Polyethyleneimine (PEI) forms polyplexes with nucleic acids. PEI complexes have a high buffering capacity in pH which is beneficial for the transfection (the process of deliberately introducing naked or purified nucleic acids into cells (19)) of the transgene. Despite this PEI is not biodegradable and aggregates in the blood which can lead to toxicity. Just as in lipid nanoparticles, the PEGylation of these complexes reduces the likelihood of aggregation occurring.

Each vector has its own advantages and disadvantages as I have explained in this section, however, deciding which vector to use in gene therapy is specific to the disease being treated. While viral vectors provide high efficiency in gene transduction, nonviral vectors are often deemed to be safer due to their reduced immunogenicity. It is imperative for clinicians and doctors to evaluate the benefits and risks associated with each vector available to treat the defect. This research has allowed me to evaluate which vector is most suitable for treating Huntington's disease which I discuss later in my ILA.

Already, scientists and clinicians are beginning to modify viral carriers through improved vectors and capsids and nonviral carriers via PEGylation. Successfully halving the immunogenicity and the number of viral particles of these therapeutics has the potential to increase the treatment success rate by over 4 times (10). Addressing the critical hurdle of the lack of feasible delivery carriers is bound to progress this field of medicine.

2. THE REMOVAL OF THE DYSFUNCTIONAL GENE AND THE EXPRESSION OF THE CORRECTIONAL GENE IN THE GENOME:

As I mentioned above, there are different ways to treat the defective gene, from knocking it out entirely to adding a functional gene working alongside the faulty one. Just as different vectors are more suitable in treating different diseases, the method used to regulate the defective gene's expression is dependent on the patient's disease. The following methods of gene therapy are contained and delivered by the vectors which I have discussed. I will give a summary of the mechanisms behind various treatments used to correct the dysfunctional gene, as well as their advantages and disadvantages before I can conclude which is most suitable to treat Huntington's disease.

1. Chimeric antigen receptor (CAR) T-cell therapy

This method is applied with the intention of getting modified T-cells to attack and destroy cancer cells. T-cells are extracted from a patient's body by adding a gene for a receptor called a CAR, which helps the T-cells bind to a specific cancer antigen. These are then reintegrated into the patient's body.

The procedure to remove blood cells from a patient's body is via IV lines using a method called leukapheresis in which white blood cells are separated from the blood sample (20). Once these have been removed, CAR genes are developed from a monoclonal antibody that binds only to complementary antigens. These genes are incorporated into the genome of the T-cells and code for the production of these receptors. These cells are grown ex vivo and cloned over several weeks to manufacture the large number of cells needed for the treatment.

To increase the effectiveness of the T-cells, the patient may be given chemotherapy to help lower the number of immune cells already in the body. Once the CAR T-cells bind to the cancer cells, they are able to multiply and destroy even more cancer cells. Their destruction is due to chemicals called cytokines released by the modified T-cells. However, this can lead to the release of toxic cytokines into the body which may harm other bodily functions. Furthermore, this process is short-term and usually only lasts for a few days after the initial procedure; this means if a patient relapses, they may require another dose of treatment which holds the same risks as the first round as well as another round of chemotherapy on an already weakened immune system.

2. Antisense oligonucleotides (ASOs)

Oligonucleotides are single-stranded DNA molecules. Many oligonucleotides have been chemically synthesised or modified so that they are more effective for use in therapeutics, and this has resulted in various oligonucleotide species. They can be divided into two categories based on their mechanism of action: the RNase H-dependent oligonucleotides which degrade mRNA and the inhibitor oligonucleotides which physically regulate the progression of splicing and translational machinery.

RNase H is an endoribonuclease that specifically hydrolyses the phosphodiester bonds between DNA-RNA heteroduplexes (21). ASOs, which bind to the mRNA by complementary base-pairing, recruit the help of this enzyme to cleave mRNA and cause its degradation. The binding between the ASO and mRNA target is such that it mimics the DNA-RNA (template strand) binding that occurs during DNA replication, in which RNAse H cleaves the RNA strand (22). This activation by the binding of ASOs-mRNA is ultimately what leads to the target's degradation.

The other mechanism of action the ASOs use, involves the alteration of splicing and the blocking of translation. Alterations in the splicing pattern of pre-mRNA can result in disease; this is because inappropriate deletions will result in a change in the reading frame of the mRNA. ASOs that block splicing machinery and stop them from "over-splicing" promote splicing that leads to the correct mRNA sequence composed solely of exons. This prevents frameshift mutations from occurring and leads to the correct expression of the gene. In addition to this, blocking translational machinery is an effective way of inhibiting the defective gene's expression. This involves chemically engineering an ASO which binds near the start codon of the spliced mRNA sequence and stop the ribosome from binding with the target sequence. This prevents the translation and folding of the target mRNA sequence.

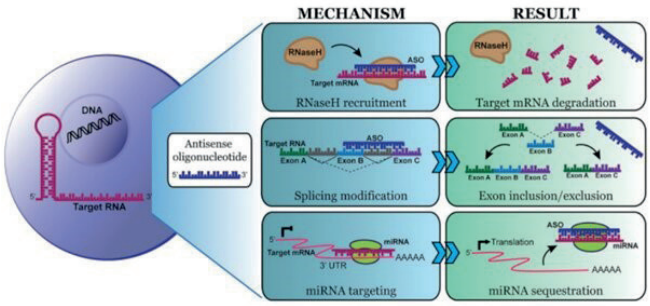


Figure 4 - Antisense oligonucleotide mechanism for regulating defective gene expression https://www.cell.com/neuron/comments/S0896-6273(17)30303-3

3. RNA INTERFERENCE (RNAI)

RNAi, which was discovered in 1998 in the roundworm Caenorhabditis elegans, is a process which is triggered by the presence of double-stranded RNA molecules (dsRNAs). The process, aptly named 'interference', involves the 'knockdown' of gene expression, also known as gene silencing, which is controlled by short RNA molecules that enable mRNAs to be regulated. There are various mechanisms through which the RNAi system can regulate mRNA molecules, each acting after transcription, but the most prevalent is through facilitating their degradation.

The degradation of mRNA molecules involves two enzymes, dicer and slicer. Double-stranded RNA is recognised by the dicer endonuclease, which 'dice' the mRNA molecules to form short double-stranded small interfering RNAs (siRNAs) or microRNAs (miRNAs). These siRNAs can also be designed exogenously to target a gene of interest. They associate with a protein to form a complex known as the RNA-induced silencing complex (RISC) which holds a nuclease called "slicer". Once bound in the complex, the double-stranded siRNAs are unwound into their single-stranded components. One of the two strands guides the protein complex whilst the other is complementary to the target mRNA's sequence. The complementary strand is also known as the antisense RNA fragment and binds with the target mRNA's complementary sequence by base pairing. The slicer nuclease then cuts the mRNA and the product is degraded by exonucleases. In some cases, the antisense strand is not always perfectly complementary to the target mRNA sequence, this leads to the inhibition of protein translation from the target as well thereby regulating the gene's expression.

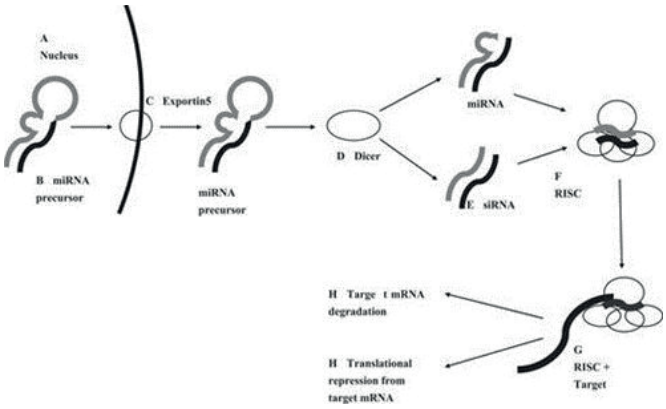


Figure 5 - Diagram showing the mechanism of interference RNA to regulate the expression of a gene

4. CRISPR-Cas9

I briefly touched upon the history of the cutting-edge CRISPR-Cas9 technology as well as a quick summary of its mechanism. Although there is more science behind this mechanism, it would be relatively easy for a new scientist exploring this field to use the technology. In fact, the CRISPR-associated proteins such as Cas9 and Cpf1 (which is used as an alternative method in the genome editing of plants, without introducing a transgene into the plant cells (23)), can be ordered alongside plasmids (to produce guide RNAs) in less than one week. On top of this, using rapidly growing organisms or cells like E.coli, genes can still be edited in less than a week (24).

In nature, CRISPR-Cas9 is used as a defensive mechanism in bacteria to fend off invading bacteriophages. Essentially, CRISPR cleaves a small part of the bacteriophage's DNA and stores it in its own genome. The bacteria, if reinfected by the same virus, can recognise the invader using the cleaved DNA stored in its genome and destroy the virus. The following steps describe the bacteria's immune response to an invading virus:

- Firstly, DNA from the invading bacteriophage is cleaved and stored in a region of the bacteria's genome called the CRISPR locus. This short sequence of the viral DNA is known as a spacer which is inserted into the repeated CRISPR sequence.
- During reinfection by the same species of virus, the cleaved DNA in the CRISPR locus, which corresponds to the invader's DNA, is transcribed into a short piece of RNA, termed CRISPR RNA (crRNA).
- crRNAs bind to proteins such as Cas9 and guides them to the invader's DNA, which has a sequence that complements the crRNA sequence. The crRNA and protein (known now as an effector complex) bind to the bacteriophage's DNA.
- The bacteriophage's DNA is cut and can be stored again to repeat this process in the event of a future invasion. Once the effector complex is bound to the target, it performs a double-stranded DNA cleavage.

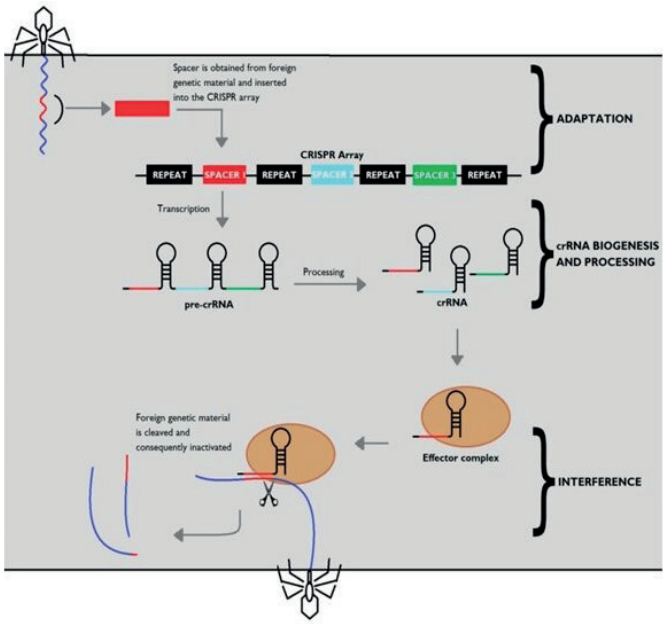


Figure 6 - Mechanism behind the CRISPR-Cas9 system in bacteria https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6466564/

The diagram to the left gives a visualisation on how the CRISPR system functions.

One term which I also refer to later is "processing". This refers to the use of a second RNA called a tracrRNA which is complementary to pre-processed crRNA. A ribonuclease cleaves ("processes") the crRNA and a hybrid double strand between this and the tracrRNA is combined with a protein to form the effector complex.

In 2012, Jennifer Doudna's group led an experiment which allowed them to fuse crRNA and tracrRNA into a single RNA sequence. This has become known as guideRNA (gRNA) and its significance in the medical application of CRISPR cannot be understated.

HOW CAN THE CRISPR-CAS9 SYSTEM BE USED IN MEDICINE?

The genomic target of CRISPR can be any nucleotide DNA sequence provided it meets these two requirements:

1. The target sequence must be unique compared to the rest of the genome

2. The chosen target must be adjacent to a Protospacer Adjacent Motif (PAM)

A PAM sequence is typically 2-6 nucleotides in length (25), and serves as a binding site for the protein-gRNA ribonucleoprotein (the effector complex). The protein will only cleave the target DNA if the gRNA is complementary to it, and if they are, a double-stranded cut is made by two catalytic domains of Cas9, HNH and RuvC (26), (or the protein in use) in the target DNA sequence at a position which is three base pairs along from the PAM on the 3' strand of the DNA. The end result is a double-stranded break (DSB) within the target DNA.

Cell-mediated repair of the DSB occurs in one of two ways: non-homologous end joining (NHEJ) or homology-directed repair (HDR). When the cell uses NHEJ, which is often a random and unpredictable process, to repair the break, often mutations occur that result in amino acid insertion, deletion or even frameshift mutations. The end result is a dysfunctional target gene which is no longer able to be expressed. This is how a dysfunctional gene can effectively be 'shut down' by the CRISPR system so that it no longer produces a defect in a patient. The mechanism is often referred to as 'down regulation' or 'knockdown' of gene expression as it causes the silencing of the defective gene's expression.

Moreover, gene repression can also be performed using the CRISPR system: scientists introduced two mutations into the catalytic domains of the Cas9 protein which inhibited its cleaving function. This 'dead' Cas9 (dCas9) could still bind to the PAM of the target sequence without any cleaving taking place, which meant that it could be used as an inhibitor to RNA polymerase to prevent transcription of the target DNA.

Another purpose of the CRISPR system is to induce or correct a point mutation which requires only a single base substitution. By inducing a single mutation in the RuvC catalytic domain of the Cas9 protein, researchers produced Cas9 nickase (Cas9n), which creates a single-stranded break in the targeted sequence. Coupling the Cas9n with a cytidine deaminase enzyme enables the deamination of cytosine bases into uracil which is converted to thymine via DNA repair methods. This induces a C to T substitution in the target sequence (or a G to A on the other strand).

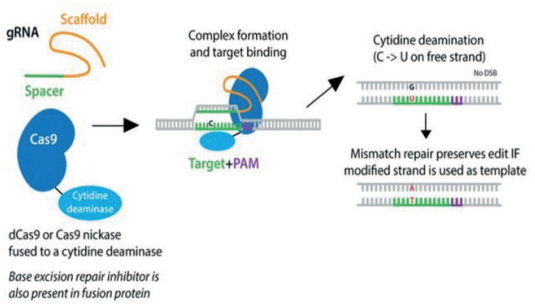


Figure 7 - Mechanism using CRISPR to induce a single base substitution in target DNA https://www.addgene.org/crispr/base-edit/

While each of the methods mentioned above have their own advantage, ultimately, the biggest disadvantage associated with all these methods, is the risk of inducing a new mutation which has an even worse effect on the patient. It is indisputable that CRISPR is the most versatile therapy due to its various functions. Having researched the pathology of Huntington's disease, I will be able to make an informed decision on which corrective method is best in its treatment and if the treatment is actually feasible.

APPLICATION OF CRISPR GENE THERAPY USED TO TREAT CANCER AND OTHER DISEASES:

The treatment of cancer presents some of the most important and innovative applications of gene therapy. To give an idea of the real-life applications of gene therapy, I will highlight some of the most exciting and successful approaches used to treat cancer and other diseases.

Using mice as models, scientists and researchers could imitate the course of cancer progression in humans. They could then use these models to experiment with the treatment of cancer using CRISPR technology. One challenge which immediately presented itself was the relatively large size of the Cas9 protein in comparison to the somatic cells of the mice which made their delivery tricky. However, several viral approaches were conceived, and these were injected into the liver. The effector complex formed by Cas9 and the gRNA was delivered to target defective tumour suppressor genes Pten and p53. Mutations occurring in these genes account for the rapid and uncontrollable cell division of liver cells. CRISPR technology caused deletion mutations in the target genes through HDR cell-mediated repair which halted the expression of the defective gene and therefore stopped tumour growth. Another example of CRISPR gene therapy was using an adenoviral vector to introduce the knockdown of mutations in the endogenous Pcsk9 gene. The Pcsk9 gene codes for a fundamental protein that helps regulating the amount of cholesterol in the bloodstream (27). The treatment of this gene was successful and resulted in decreased cholesterol levels in the bloodstream (28). Finally, a Cas9 delivery vector based on an adeno-associated virus has been used to target post-mitotic neurons in adult mice. The effector complex was targeting a single gene Mecp2 (which causes the neurodevelopmental disorder Rett syndrome (28)). The therapy was delivered by injection into the dentate gyrus, which is part of the hippocampal formation in the temporal lobe of the brain (29). The treatment was successful and has provided an exciting new outlook into the potential for using this method to treat brain cancer.

These three applications provide a general insight into the use of gene therapy and the variety of its uses. The use of gene therapy seemingly always provides new data and information alongside successful treatment. I am hopeful that the uses of gene therapy will continue to broaden so that even the most obscure genetic diseases can be treated and cured.

PATHOLOGY OF HUNTINGTON'S DISEASE:

With the process of gene therapy and its current applications in mind, we can begin to understand how these therapeutics could offer an effective treatment used to cure a genetic disease such as Huntington's disease (HD).

Huntington's disease is classed as an autosomal dominant disorder, which is caused by an inherited difference in a single gene. The "autosomal" refers to the location on one of the numbered, non-sex chromosomes, while "dominant" means that only a single copy of the mutated gene is needed to cause the inheritance of the disorder in the next generation. To explain further, every person inherits two copies of each gene; a chromosome from each parent. The term allele denotes the variant of a given gene, and this is often preceded by the word "dominant" or "recessive". If an allele is "dominant", it will always express the trait even if there is only one copy, whereas "recessive" alleles can only express themselves if there is a copy from each parent.

This means, that if one of the two parents is heterozygous – only contains one copy of the nontypical gene – there is a 50% chance that they will pass on the mutated gene.

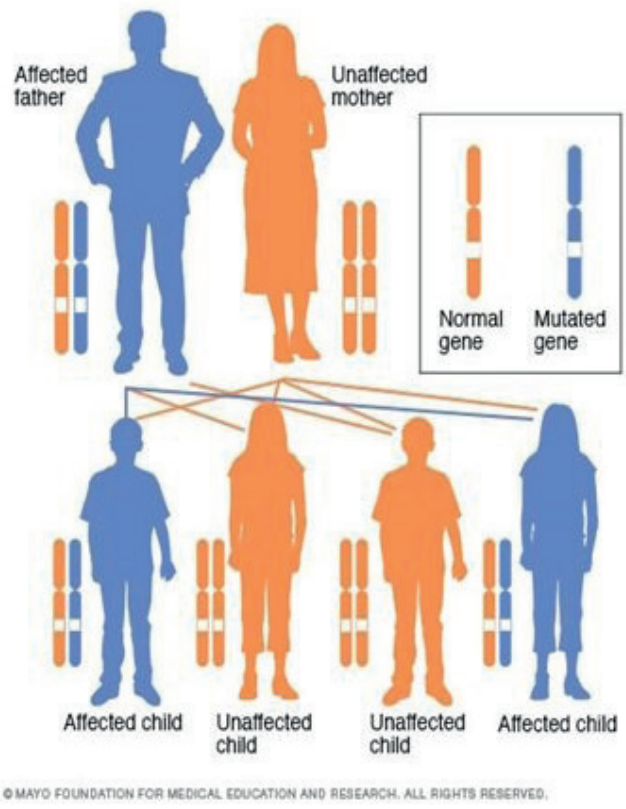


Figure 8 - the autosomal dominant inheritance pattern https://www.mayoclinic.org/-/media/kcms/gbs/patient-consumer/images/2013/11/15/17/37/r7_autosomaldominantthu_jpg.jpg

This increases to 75% if both parents are heterozygous. The diagram to the right shows the autosomal dominant inheritance pattern.

Not only are the inheritance rates shockingly high for families affected by Huntington's, but the course the disease itself runs is even more terrifying. The decline begins often with mild psychotic and behavioural symptoms, and is slowly followed by chorea – a movement disorder that causes irregular and unpredictable muscle movements (30) – and dementia. Sufferers will eventually plummet into a life affected by deep depression, hallucinations and delusions. The relentless neurodegeneration is a hallmark of Huntington's disease and

the torture of not knowing when the disease will strike make it one of the most devastating diseases.

THE IMPORTANCE OF CAG REPEATS

The defect, which is located near in the first exon of the short arm of chromosome 4, involves a type of mutation called a trinucleotide repeat. A trinucleotide repeat is a sequence of three nucleotides in consecutive succession. If repeat numbers exceed a certain threshold, there can be adverse effects on the function of the gene which often results in a genetic disease (31). In the case of Huntington's disease, the defect is caused by the number of CAG (which codes for the amino acid glutamine) codon repeats in the gene. If this number is below thirty-five repeats, the person will not be affected by the disease, but if the codon is repeated over thirty-nine times, they will slowly begin this neurodegenerative decline until they die. The age at which the disease will begin to show its effects can only be predicted by the number of repeats contained in their genome, which leaves sufferers in a state of prolonged and torturous anxiety. Such is the horrifying nature of Huntington's disease to which many scientists have been trying to find a cure.

Since the number of CAG repeats determined whether or not people developed the disease, researchers began to hypothesise whether glutaminergic overactivity (excitotoxicity) was key in the pathogenesis of Huntington's. While this has proved to be true, there are other pathways which include oxidative stress, apoptosis and abnormal protein-protein interactions that can cause transcriptional dysfunction and mutated gene expression (32).

The normal function of the Huntington gene is to code for a protein called Huntingtin. The protein is widely distributed around the central nervous system and this has led researchers to question how its mutated variant gives rise to the loss of neurons which is characteristic of HD. The normal protein is usually found in the in the cytoplasm of neurones and nerve cells (33), and is involved in protein degradation intracellularly. Researchers who investigated the effects of increasing the number of CAG repeats in mice models, found that the proteins formed insoluble aggregates in the proteosome of the nucleus. The formation of these insoluble nuclear aggregates reduces the proteosomes capacity to digest other intracellular proteins, including those related to promoting cell apoptosis. Although there is sufficient evidence to support the concept that aggregate formation is the causal agent in neuronal death, there is not enough evidence to suggest it is the only mechanism. Another potential cause of cell dysfunction in HD is the disruption of transcription by the binding of mutant Huntingtin proteins to transcriptional machinery (34). This prevents any new genes from being transcribed which ultimately leads to cell death. The neurodegeneration which is caused by Huntington's disease is characterised by this progressive loss of neurons. If there was a treatment which could alter the mutated Huntingtin gene, then the death of neurons could be prevented, and in turn, we could prevent neurodegeneration by Huntington's disease...

CONCLUSION: IS GENE THERAPY AN EFFECTIVE TREATMENT FOR HUNTINGTON'S DISEASE?

I have discussed the history, the mechanisms behind gene therapy, and the pathology of HD in my ILA, however, there is still a lot more depth to the process of gene therapy and its applications are extensive. My research has still given me a very good basis from which I can make a conclusion to my ILA question. In fact, there are multiple mechanisms which have the potential to treat HD and I will evaluate their potential to be successful and which is the most feasible treatment.

Firstly, we must consider whether the treatment should be performed in vivo or ex vivo, as this will ultimately determine which mechanism/type of gene therapy is used. Usually, ex vivo treatments are better suited for treating diseases related to the blood, such as haemophilia (35), where a person's stem cells can be engineered with a functional gene before they are returned to the body.

On the other hand, in vivo treatments are far better suited for reaching specific tissues or organs that can be accessed via the blood. In vivo methods, therefore, have a better chance of successfully treating targets in the central nervous system because they can be delivered via an intracerebellar injection directly into the brain. However, there is an extra complication when performing in vivo gene therapy in the brain: the blood-brain barrier. I discuss the solutions to this hurdle as I evaluate each treatment and their vectors, which have the potential to treat HD.

WHICH VECTOR AND TYPE OF GENE THERAPY ARE THE MOST SUITABLE FOR DELIVERING THE TRANSGENE TO THE TARGET CELLS AND FOR THE CORRECTION OF THE GENE?

The vector used to transfect or transduce a corrective gene depends on the type of cargo they are transporting. While each mechanism can use either viral or nonviral vectors, often there are benefits to using one rather than the other.

The potential of using virally expressed interference RNA to alter the mutated Huntingtin gene was explored by a team of scientists. They used short hairpin RNA (shRNA) molecules, which are sequences that have a hairpin loop in them as their name suggests. In certain situations, this molecule has a similar mechanism to both siRNAs and microRNAs, which induce the silencing of a particular gene (36). The shRNA can be incorporated into viruses such as retroviruses, adenoviruses and AAVs, which permit stable and more efficient transduction than nonviral vectors. This is especially helpful for targeting cells which are particularly difficult to access. However, delivering the transgene into the brain is especially difficult due to complications with the blood-brain barrier which protects the brain from pathogens. For this reason, the adeno-associated virus, which was successfully used to cross the blood-brain barrier (37), can be used to deliver the shRNAs. In mice model, the team was successfully able to silence the human mutant allele (38).

However, there is another complication with this method: the treatment was not specific enough for the mutant allele. This is because several other functional genes have CAG repeats which were silenced by the shRNAs resulting in adverse effects. One strategy which was employed to overcome this, has been to target a single nucleotide polymorphism (SNP), which distinguishes between the mutant allele and the other CAG repeats themselves (38). A single nucleotide polymorphism is a germline substitution of a single nucleotide at a specific locus on the genome (39). This has only been tested in mice, due to the ethics surrounding germline gene editing in humans which I shall summarise now.

The idea of germline alterations is controversial. While it could prevent future generations from inheriting Huntington's disease and other similar disorders, it has the potential to affect the development of a foetus in unpredictable ways with side effects which are potentially worse than the defect you are trying to correct. In addition to this, people who are treated using germline gene therapy do not have a choice as they are not born yet; some say it is unfair to practice this therapy on them without their consent and their choice. Scientists have only (legally) practised somatic cell therapy, and extensive research into the effects of germline cell therapy needs to be conducted before its application in animal

models and subsequently its application in humans. New laws and legislation must be implemented before this practice can be initiated and its ethics should be considered before it is made legal. The ethics surrounding germline cell therapy is a huge subject and while I have summarily discussed it in this paragraph, my ILA has primarily focused on the science behind gene therapy treatments.

Similar to the shRNA method, is the use of siRNA, although its effects are short-acting and it therefore requires continual dosing in order for successful treatment to be achieved. The mechanisms of both siRNA and shRNA are compared in the image below, and while both achieve the same end effect, there are nuances in their mechanisms of action.

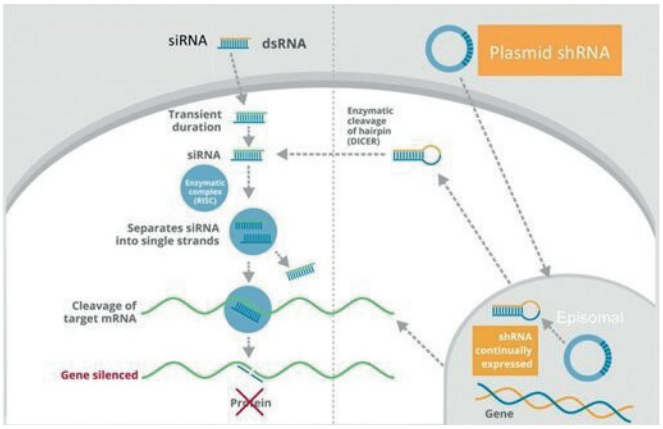


Figure 9 - The mechanisms of siRNA and shRNA when applied in the CNS https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5702971/

Another method which has the potential to treat HD is the use of antisense oligonucleotides. These are particularly helpful in conjunction with RNAse H which are advantageous when a distinct RNA dysregulation or protein accumulation is recognised as the cause of the disease. In the pathology of HD, I described the aggregation of insoluble Huntingtin proteins which gathered in the cytoplasm of neurons and led to cell death. Since these proteins, which are translated from mRNA, are the primary cause of Huntington's disease, treatment using RNAse H is a possibility. The pairing of ASOs to their complementary mRNA target activates the RNAse H enzyme (due to the mimicry of DNA-RNA binding in DNA replication), which causes the degradation of the mRNA target. This destruction of the mRNA inhibits the translation of the gene so less protein aggregates. The beauty of this method is that the ASO itself avoids degradation by the RNAse H because the enzyme specifically targets the mRNA. This means that the ASO can cause the degradation of multiple mRNA strands which causes the yield of Huntingtin to be further reduced. The results of testing in mice models was the reduction of transgenic human mRNA by 50-80% across the brain regions, and the survival rate was notably increased when applied to a more severely progressing mouse (40). However, it still must be taken into account that ASOs did not only target the mutant Huntingtin gene, but also several other wild-types; luckily there were no observable adverse effects. However, the risk remains that when applied to humans, the targeting of wild-type nonmutant genes could pose a threat to the safety of the patient.

In order for ASOs to be successful, they must remain stable and maintain efficacy over time for feasible treatment. The blood-brain barrier again proves a challenge as it restricts entry of certain molecules based on their size, charge or solubility, and since ASOs are not transported by vector, the use of AAVs is out of the question. Despite the prediction that highly charged ASOs would have difficulty crossing the blood-brain barrier, their successful distribution in the CNS was a surprising and exciting result when tested in mice using intraventricular or intrathecal injection (22).

Before their application in humans, there are certain risks and challenges which accompany the treatment by ASOs, that must be mitigated. Firstly, the on-target 'over-effectivity' may result in the ASOs reducing the protein population by too much. This can have adverse effects: a significant reduction in Huntingtin proteins resulted in toxicities when experimenting on the effects of mRNA degradation in rodents (22). Secondly, the lack of specificity of ASOs targeting can cause off-target toxicities in wild-type genes. I have already highlighted the importance of segregating CAG repeats and the actual Huntingtin gene in the application of interference RNA; the same applies here.

While CRISPR-Cas9 is indisputably one of the most effective treatments for oncogenes and certain types of cancer, it has limited applications with current technologies and with current understanding in the treatment of neurodegenerative diseases. This comes down to the fact that it is hard to insert CRISPR into the CNS as it is too large to fit in an AAV that can cross the blood-brain barrier. Furthermore, CRISPR requires a PAM site to bind to which is adjacent to the polyglutamine strand, which limits CRISPR application in treating the Huntingtin gene. Most importantly, the bacterial origin of CRISPR which I alluded to earlier has the potential to elicit an immune response in the brain. However, the major benefit which other treatments cannot provide with current technologies, is the ability of CRISPR to provide a permanent cure to diseases.

Although no treatments for Huntington's disease have been successfully trialled in humans, the emerging knowledge of how gene therapy can be applied more successfully, as well as the successful treatment of HD in mice model, provide motivation for their future application to human patients. Current gene therapy trials indicate that shRNA treatment is far more effective compared to ASOs or siRNA. While ASOs do not require the extra expense and risks associated with viral vectors, they have a short-acting effect which requires continuous dosing. This has the potential to cause adverse effects due to on-target and off-target toxicities, and while the same holds true for shRNA, constructs can be inserted in conjunction with microRNA to mitigate the potential for shRNA-induced neurotoxicity (41). The success of shRNA molecules which provided treatment for several months in primates (41), as well as the successful application of ASOs in mice models provide evidence that scientists are not far off from finding the perfect gene therapy type for the treatment of Huntington's disease. I also remain hopeful that in the future, scientists will be able to integrate a permanent cure using CRISPR to treat HD. I conclude that the transition from animal-based models to human applications is just around the corner and will hopefully be means of an end to the suffering caused by Huntington's disease.

CAN 'SUPERHUMANS' BE GENETICALLY ENGINEERED INCLUDING TRANSGENIC ORGANISMS

Finally, I return to my love of superheroes. While I have not discussed the possibilities of transgenic organisms in my ILA, the potential to fuse human DNA with the DNA of other organisms will be an exciting new chapter in the evolution of the human species. My title 'revolutionising the process of evolution' refers to the dawn of genetic modifications and its applications which may soon have the ability to create the 'perfect' human. Already, as discussed in my ILA, scientists have discovered the ingenuity of gene therapy which can be used to treat numerous diseases; the ability to cure the most vicious diseases such as cancer and Huntington's disease brings us one step closer to ridding ailments in the human species. In addition, the dawn of germline cell therapy has the potential to create a genome without any flaws or mutations, and one which could provide a superior human species with so-called 'superpowers'.

REFERENCES

1. cban. [Online] [Cited: 06 02, 2022.] https://cban.ca/gmos/faq/gmge-definition/#:~:text=Genetic%20modification%20(GM)%20is%20the,called%20genetic%20engineering%20or%20GE..
2. University of Nebraska. [Online] [Cited: 06 02, 2022.] https://agbiosafety.unl.edu/basic_genetics.shtml.
3. Pubmed. [Online] [Cited: 06 02, 2022.] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2907101/#:~:text=On%20September%2014%2C%201990%2C%20W,)%20(Anderson%2C%201990)..
4. Labiotech. [Online] [Cited: 06 02, 2022.] https://www.labiotech.eu/in-depth/gene-therapy-history/.
5. pubmed. [Online] [Cited: 06 02, 2022.] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2907101/#:~:text=On%20September%2014%2C%201990%2C%20W,)%20(Anderson%2C%201990)..
6. The Jackson Laboratory. [Online] [Cited: 06 04, 2022.] https://www.jax.org/personalized-medicine/precision-medicine-and-you/what-is-crispr#:~:text=CRISPR%20stands%20for%20Clustered%20Regularly,that%20exactly%20match%20viral%20sequences..
7. Oxford reference. [Online] [Cited: 06 04, 2022.] https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095849792.
8. University of Cambridge. [Online] [Cited: 06 04, 2022.] https://www.phgfoundation.org/briefing/somatic-genome-editing-overview#:~:text=Genome%20editing%20can%20be%20performed,received%20far%20less%20media%20attention..
9. Wikipedia. [Online] [Cited: 06 06, 2022.] https://en.wikipedia.org/wiki/Genetic_recombination.
10. mckinsey. [Online] [Cited: 06 21, 2022.] https://www.mckinsey.com/industries/life-sciences/our-insights/gene-therapy-innovation-unlocking-the-promise-of-viral-vectors#:~:text=AAV%20and%20adenovirus%20vectors%20are,outside%20of%20oncology%20and%20vaccines..
11. National Library of Medicine. [Online] [Cited: 06 21, 2022.] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5400678/#R9.
12. medicinenet. [Online] [Cited: 06 21, 2022.] https://www.medicinenet.com/how_does_an_enhancer_work/article.htm.
13. Wikipedia. [Online] [Cited: 06 21, 2022.] https://en.wikipedia.org/wiki/Immunogenicity.
14. National Library of Medicine . [Online] [Cited: 06 21, 2022.] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3584155/#:~:text=Lowering%20vector%20doses%20to%20reduce,administered%20virus%20can%20be%20reduced..
15. National Library of Medicine. [Online] [Cited: 06 21, 2022.] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7083087/.
16. Bio process international. [Online] [Cited: 06 21, 2022.] https://bioprocessintl.com/bioprocess-insider/deal-making/abbvie-ends-vectorized-antibody-accord-with-voyager/#:~:text=Vectorized%20antibodies&text=The%20idea%20is%20that%20viral,target%20the%20diseases%20in%20question..
17. Biopharma PEG. [Online] [Cited: 06 22, 2022.] https://www.biochempeg.com/article/122.html.
18. Wikipedia. [Online] [Cited: 06 22, 2022.] https://en.wikipedia.org/wiki/PEGylation.
19. Wikipedia. [Online] [Cited: 06 22, 2022.] https://en.wikipedia.org/wiki/Transfection.
20. Wikipedia . [Online] [Cited: 06 23, 2022.] https://en.wikipedia.org/wiki/Leukapheresis.
21. New England Biolabs. [Online] [Cited: 06 22, 2022.] https://www.neb.com/products/m0297-rnase-h.
22. Neuron. [Online] [Cited: 06 24, 2022.] https://www.cell.com/neuron/comments/S0896-6273(17)30303-3.
23. Frontiers in Plant Science. [Online] [Cited: 06 22, 2022.] https://www.frontiersin.org/articles/10.3389/fpls.2020.00264/full#:~:text=Cpf1%20proteins%2C%20along%20with%20guide,a%20DNA%2Dfree%20editing%20system..
24. Medium. [Online] [Cited: 06 22, 2022.] https://nikomccarty.medium.com/almost-everything-you-should-know-about-crispr-how-it-works-top-applications-and-how-to-use-it-61e27b04bea6.
25. Synthego. [Online] [Cited: 06 22, 2022.] https://www.synthego.com/guide/how-to-use-crispr/pam-sequence#:~:text=The%20PAM%20is%20about%202,4%20nucleotides%20upstream%20of%20it..
26. National Library of Medicine. [Online] [Cited: 06 22, 2022.] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6466564/.
27. Medline Plus. [Online] [Cited: 06 23, 2022.] https://medlineplus.gov/genetics/gene/pcsk9/#:~:text=The%20PCSK9%20gene%20provides%20instructions,foods%20that%20come%20from%20animals..
28. Genome medicine. [Online] [Cited: 06 23, 2022.] https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-015-0178-7.
29. Wikipedia. [Online] [Cited: 06 23, 2022.] https://en.wikipedia.org/wiki/Dentate_gyrus.
30. Cleveland Clinic. [Online] [Cited: 06 19, 2022.] https://my.clevelandclinic.org/health/diseases/21192-chorea.
31. National Cancer Institute. [Online] [Cited: 06 23, 2022.] https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/trinucleotide-repeat.
32. BMJ journals. [Online] [Cited: 06 23, 2022.] https://mp.bmj.com/content/54/6/409.
33. Frontiers. [Online] [Cited: 06 24, 2022.] https://www.frontiersin.org/articles/10.3389/fmolb.2021.769184/full.
34. BMJ Journals. [Online] [Cited: 06 24, 2022.] https://mp.bmj.com/content/54/6/409.
35. Genehome. [Online] [Cited: 06 24, 2022.] https://www.thegenehome.com/how-does-gene-therapy-work/techniques#:~:text=Selecting%20an%20ex%20vivo%20treatment,then%20delivered%20into%20their%20body..
36. horizon. [Online] [Cited: 06 24, 2022.] https://horizondiscovery.com/en/applications/rnai/shrna-applications.
37. UNC School of Medicine. [Online] [Cited: 06 24, 2022.] med.unc.edu/genetherapy/gene-therapy-researchers-find-viral-barcode-to-cross-the-blood-brain-barrier/.
38. Oxford academic . [Online] [Cited: 06 24, 2022.] https://academic.oup.com/bfg/article/6/1/40/272507.
39. Wikipedia. [Online] [Cited: 06 24, 2022.] https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism.
40. National Library of Medicine. [Online] [Cited: 06 24, 2022.] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5971383/.
41. National Library of Medicine. [Online] [Cited: 06 24, 2022.] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5702971/.

How the criticisms of Utilitarianism underline a fundamental error in our approach to ethical discourse

Winner of the University of Nottingham's David Garlick prize for the best A level essay in Religious Studies

Stuart Brown, Lower Sixth

Utilitarianism as a normative ethical theory is attacked in a number of different ways, however I hope to show how these criticisms demonstrate a fundamental mistake in the way in which we go about breaking down an ethical theory.

The first criticism which is often asserted is the impracticality of Utilitarianism when it comes to decision making in our daily lives. Even if we accept the idea that we must act in the way that best tends to produce happiness it is impossible to know which actions will cause this. We cannot predict the vast and unforeseeable consequences of our actions and hence Utilitarianism seemingly fails as we cannot effectively and accurately fulfil the task of promoting happiness in the real world. Mill strives to object to this in his book 'Utilitarianism' writing 'that there has been ample time, namely, the whole past duration of the human species.' His point here is that humans know basically which actions tend to produce more happiness as a result of the cultivated experience of humanity and the general attitudes that we have formed over time to specific actions due to such experience. Therefore, we know which actions to undertake to produce overall greater happiness. However, one must question whether Mill is even obligated to respond to the challenge of impracticality. The truth of the principle of utility and the very ethical theory itself is unaffected and detached from the question of whether it can be usefully applied in the real world. If it is true to seek the happiness of the greatest number, then this remains the case whether or not we able to do so. Hence, we see that when discussing the validity of normative ethical theories, the issue of practicality is unimportant as it has no bearing on the actual truth of the theory. The question of practicality is however not useless but rather misplaced. It should come later once a base ethical theory has been established and we look to how it can be applied.

Another popular yet erroneous approach is to argue from the starting point of a known ethical truth to try and establish or dismiss an ethical theory. To say for example, that murder is always wrong, and then to identify a specific case where Utilitarianism justifies murder is not necessarily a valid argument that Utilitarianism fails as an ethical theory because it appears to justify a wrong action. Whilst this argument may seem logical at first it presupposes that murder, or another action is simply inherently wrong. This is to fall into the fallacy of question begging as it assumes that Utilitarianism is incorrect and that some actions must have inherent value to prove that Utilitarianism is in fact incorrect. This structure of reasoning is common and often used especially in the case of Utilitarianism, but it fails crucially in all cases because it cannot without using circular reasoning establish that any given action is wrong. This problem illustrates a common mistake in how we approach ethics in that we try and find a theory to cohere with our current values. This is problematic as our self-held beliefs cannot act as a firm groundwork for an ethical theory. Instead, we must build up an ethical theory from its very foundation and derive attitudes towards specific actions later.

The trolley problem and how it is discussed often shows our disposition to starting from judgements of specific actions and then working towards an ethical theory to match such assumptions. This is a common introductory thought experiment to the topic of ethics and is one where most start with an opinion on whether it can be right to pull the lever to kill one and save five and work backwards to an ethical position. However, this is foolish as the point of an ethical theory is not to justify our previously held beliefs and judgements but rather to provide a starting framework to build our ethical perspectives anew.

Whilst many of the criticisms of Utilitarianism fail, there is one which is very difficult to overcome and demonstrates the correct way to go about analysing an ethical theory. This criticism is that Utilitarianism fails to successfully establish happiness as having inherent value. Bentham falls victim to the naturalistic fallacy when trying to establish the value of pleasure. This is the fallacy outlined by David Hume that we cannot derive an ought from an is (in this case it is Bentham's argument that we naturally pursue pain and avoid pleasure and hence we ought to do so). In 'Introduction to the principle of morals, legislation' Bentham writes on pleasure and pain 'it is for them alone to point out what we ought to do, as well as to determine what we shall do' showing how his assertion of the principle of utility is fallacious. Most however, accept the inherent value of happiness as a brute fact and do not seek to break down Bentham's starting assertion although this is exactly what must be done. We must adapt our philosophical approach to examine the foundational assertions of ethical theories and hence decide their merit rather than focusing on the practical application of the theory. This is the key point in the failure of our approach to ethics as it is the starting value assumptions (such as the value of happiness in Utilitarianism) of ethical theories that must be examined as these are the foundations of ethical theories and hence their success is entirely dependent on their truth.

In conclusion, as seen in the mishandled approach to the criticisms of Utilitarianism, we must adapt our approach to the analysis of ethics and shift our focus from the practicalities and repercussions of accepting normative ethical theories. Instead, we must judge their validity on the surety of their foundational claims as only then can we properly assess the truth of an ethical theory.

From bunnies to Bitcoin trading
The ubiquity of Fibonacci

An extract of this short-listed Independent Learning Assignment (ILA)

Sam Hinton, Upper Sixth

PART ONE: INTRODUCTION AND CORE MATHEMATICS

SCOPE

This paper looks at the prevalence of the Fibonacci Sequence in many areas, including nature, architecture, stock exchange trading strategies, economics, optics... and rabbit breeding. We explore some of the more interesting mathematics behind the patterns.

FIBONACCI SEQUENCE

The Fibonacci Sequence ("the next number is the sum of the previous two") is formally defined by the relationship

$$F_n = F_n + F_{n-2} \text{ (for } n>1) \text{ and } F_1 = 1,\ F_0 = 0 \qquad \text{equation [1]}$$

This gives the sequence of integers,

0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, 2584, 4181, 6765 and so on.

It is a sequence not a series, since a series in mathematics is where terms are summed as the series continues.

BINET'S FORMULA

Rather surprisingly, there is an exact formula, known as Binet's Formula, for the nth Fibonacci number. To derive this, note that the formula in equation [1] is a difference equation. We shall use a well-known technique (e.g. one source for this is math.umass.edu Chapter 3: Linear Difference equations) for solving difference equations by assuming that a solution can be expressed in the general form of

$$F_n = A^n \text{ for some real number } A \qquad \text{equation [2]}$$

Then equation [1] gives $A^n = A^{n-1} + A^{n-2}$

And dividing both sides by A^{n-2} gives $\frac{A^n}{A^{n-2}} - \frac{A^{n-1}}{A^{n-2}} + \frac{A^{n-2}}{A^{n-2}}$

$$\Rightarrow A^2 = A^{-1} + A^0$$

$$\Rightarrow A^2 - A = 1 = 0$$

This is a quadratic equation and using the quadratic formula $\frac{-b\pm\sqrt{b^2-4ac}}{2a}$ as the solution(s) to $ax^2+bx+c=0$ (here $a=1$, $b=-1$, $c=-1$ and we use A instead of x):

$$A = \frac{1\pm\sqrt{(-1)^2 - 4.1.(-1)}}{2.1} = \frac{1\pm\sqrt{5}}{2}$$

So we have two solutions that satisfy equation [2], namely

$\varphi = \frac{1+\sqrt{5}}{2}$ which is the well known golden ratio (known as phi), and

$\psi = \frac{1+\sqrt{5}}{2}$ which is the conjugate of the golden ratio

As is well known, the golden ratio has some beautiful properties. We can produce its square simply by adding 1, and we can produce its reciprocal by deducting 1.

Thus the powers of ψ and φ satisfy the Fibonacci recursion.

In other words, $\varphi^n = \varphi^{n-1} + \varphi^{n-2}$ and $\psi^{n-1} = \psi^{n-1} + \psi^{n-2}$ and it follows that a linear combination of φ and ψ satisfies the same recursion. That is, for any real values of a and b the sequence S_n defined by $S_n = a\varphi^n + b\psi^n$ satisfies the recursion.

This is the formula to satisfy the general recurrence condition, and we can use the starting conditions of the Fibonacci Sequence to determine a and b.
Using $S_1 = 1$, $S_0 = 0$ we obtain the following:

$S_0 = a + b = 0$ and $S_0 = a\frac{1+\sqrt{5}}{2} + b\frac{1+\sqrt{5}}{2} = 1$

Substituting b with $-a$ gives $a(\frac{1+\sqrt{5}}{2}) - a(\frac{1+\sqrt{5}}{2}) = 1$

$$\Rightarrow a\sqrt{5} = 1$$

$$\Rightarrow a = \tfrac{1}{\sqrt{5}} \text{ and } b = -\tfrac{1}{\sqrt{5}}$$

This gives the amazing Binet formula for the nth Fibonacci number:

$$F_n = \frac{\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1+\sqrt{5}}{2}\right)^n}{\sqrt{5}} \qquad \text{equation [3]}$$

We can test this.

Let's try $n=6$, then $F_6 = \frac{\left(\frac{1+\sqrt{5}}{2}\right)^6 - \left(\frac{1-\sqrt{5}}{2}\right)^6}{\sqrt{5}} = \frac{17.94427... - 0.055728}{2.23607...} = 8.0000$, which is the 6th Fibonacci number.

GOLDEN RATIO OF FIBONACCI NUMBERS

Using equation [3], we can look at the ratio of consecutive Fibonacci numbers. We have,

$$\frac{Fn+1}{Fn} = \frac{\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}}{\left(\frac{1+\sqrt{5}}{2}\right)^{n} - \left(\frac{1-\sqrt{5}}{2}\right)^{n}}$$

Since $\left|\left(\frac{1+\sqrt{5}}{2}\right)\right| < 1$, as n gets very large, $\left(\frac{1-\sqrt{5}}{2}\right)^n$ gets very small, and we find that

$$\lim_{(n\to\infty)} \frac{\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}}{\left(\frac{1+\sqrt{5}}{2}\right)^{n} - \left(\frac{1-\sqrt{5}}{2}\right)^{n}} = \lim_{(n\to\infty)} \frac{\left(\frac{1+\sqrt{5}}{2}\right)^{n+1}}{\left(\frac{1+\sqrt{5}}{2}\right)^{n}} \left(\frac{1+\sqrt{5}}{2}\right)$$

In other words, the ratio of consecutive Fibonacci numbers tends to the golden ratio φ as the sequence grows.

PART TWO: NATURE AND HISTORY

RABBITS

In 1202, Fibonacci considered the growth of an idealized rabbit population, assuming that a newly born pair of rabbits are put in a field, each breeding pair mates at the age of one month, and at the end of their second month they always produce another pair of rabbits. For the purposes of his model, he assumed that the rabbits never die, but continue breeding forever.

Fibonacci was interested in working out the population size of the rabbits during a year.

His calculations went as follows:

End of Month	Breeding	Adult pairs	Baby pairs	Total pairs
1	The two rabbits mate, but there is still only 1 pair.	1	0	1
2	They produce a new pair, so there are 2 pairs in the field.	1	1	2
3	The original pair produce a second pair, but the second pair only mate without yet producing offspring, so there are 3 pairs in all.	2	1	3
4	The original pair has produced another new pair, and the pair born two months ago also produces their first pair, making 5 pairs.	3	2	5
5	There are now five adult pairs together with the three baby pairs just produced	5	3	8
n	At the end of the nth month, the number of pairs is equal to the number of mature pairs (which is the number of pairs in month n–2) plus the number of pairs last month (month n–1). Voila! The number in the nth month is the nth Fibonacci number.			

Figure 2.1: Fibonacci's bunnies

Although the scenario used by Fibonacci was biologically simplistic, it was the genesis of his recognition of the pattern that we now call the Fibonacci Sequence.

GOLDEN RECTANGLES

Let's start with this geometric construction from the Fibonacci Sequence. We start with a 1x1 square, add another 1x1 square to it to make a larger rectangle. After that, we add a square to the long side of the rectangle to build a bigger rectangle.

IFigure 2.2: constructing golden rectangles

We can continue this construction in the same way.

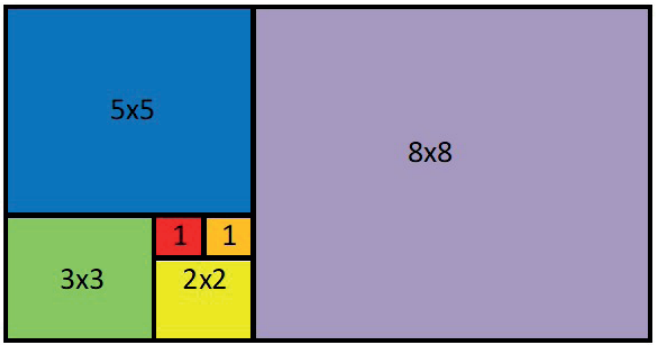


Figure 2.3: continuing constructing golden rectangles

The long side of each rectangle is a Fibonacci number, and this continues indefinitely. Note that the ratio of long side to short side is the ratio of consecutive Fibonacci numbers, which as we have seen approaches the golden ratio. These rectangles get closer and closer to being perfect golden rectangles. This suggests one property that golden rectangles have. If you add a square to the long side of a golden rectangle, you get another golden rectangle. Finally, we can inscribe a spiral onto our golden rectangles.

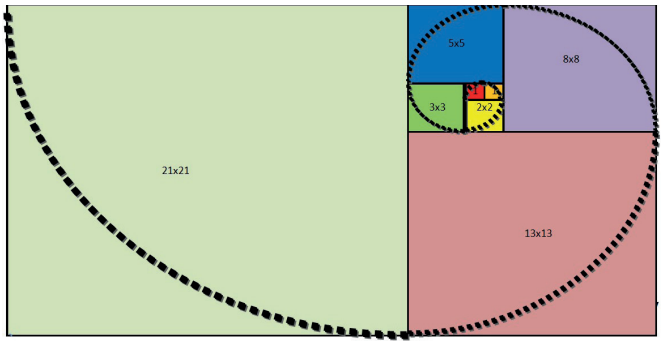
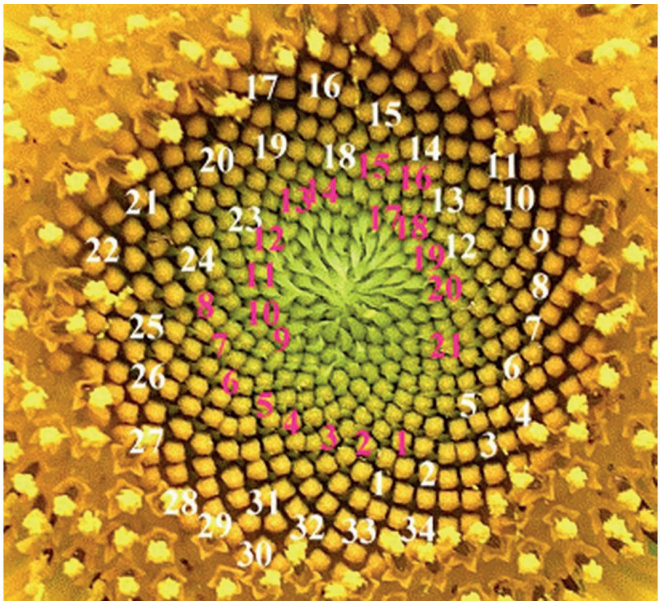


Figure 2.4: golden rectangles and the Fibonacci spiral

NATURAL PHENOMENA

This is a clue to why Fibonacci numbers appear in nature. This growth pattern is a way of getting bigger, maintaining the same basic overall shape, but without changing the old structure. Humans do not grow this way. As children grow, in general all their parts get bigger roughly equally. The length of a person's femur compared to their height has a one to four ratio, this ratio stays consistent throughout a person's life (Brabandere, 2017). Some animals and plants, however, grow this other way as we shall see. Firstly, we note that the Fibonacci Sequence appears frequently in nature. For example, you can find the Fibonacci numbers by counting the scales on a pineapple, the rows of seeds in a sunflower, etc.

Pinecones, pineapples, cauliflower, and sunflowers have Fibonacci numbers in spiral arrangements. In these cases you can count spirals clockwise and counter clockwise. The number of spirals in both directions are consecutive Fibonacci numbers. Pineapples typically have 5 and 8 spirals, or 8 and 13 spirals. Spruce cones tend to have 8 and 13 spirals. Sunflowers can have 21 and 34 , or 34 and 55 spirals – on occasions they can have as many as 233 (again in the Fibonacci Sequence).



IFigure 2.5: Fibonacci sunflower, with 21 clockwise and 34 anti-clockwise spirals

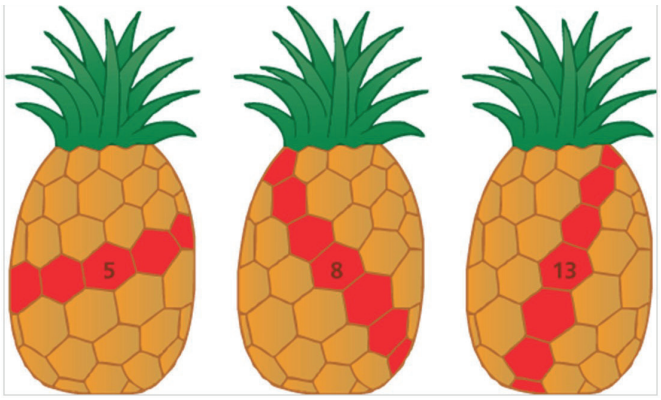


Figure 2.6: Fibonacci pineapples

Considerable work in the field of "phyllotaxis" (the study of leaf arrangement in plants) has shown that there is indeed a strong tendency in many plants to follow a true Fibonacci pattern (for example as far back as 1754 Charles Bonnet observed this phenomenon in his book "Research on the use of Leaves in Plants"). There has been extensive research producing various theories as to why this might be the case, and perhaps the strongest argument was actually made, not by botanists, but by two physicists (Douady & Couder, 1996). In an experiment involving adding tiny droplets of a magnetic fluid to silicon oil held in a magnetic field, they found that the spirals produced tended to the Fibonacci spirals. Since physical systems tend to converge to states that minimise energy, they postulated that the spirals in botany are a result of the plant growing in a way that expends minimal energy.

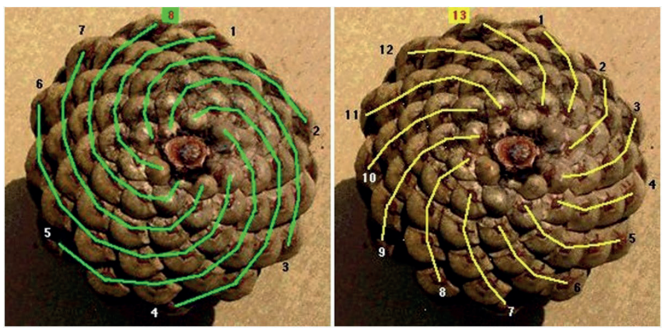


Figure 2.7: Fibonacci pinecones

GNOMONIC GROWTH

One animal that exhibits approximate Fibonacci geometry is the nautilus. As it grows, it needs a bigger shell, but since it is difficult to enlarge one's shell by stretching it all directions, the nautilus instead adds a chamber to its existing shell that's big enough for the body of the nautilus to occupy. The old part of the shell is still there, and the overall shape is the same. It's just been extended in the same way that the golden rectangles were extended above.

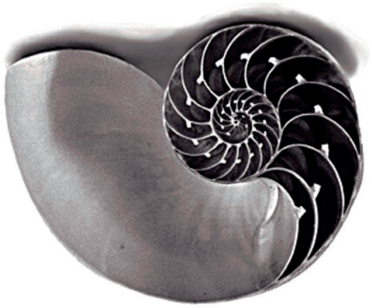


Figure 2.8: Nautilus shell showing the Fibonacci spiral

This is called gnomonic growth, where you grow by adding a piece to your old shape, but still maintain the same basic shape. Adding squares to the long side of golden rectangles is a geometric example of gnomonic growth and the nautilus's shell is a reasonable approximation to that geometry.

GALAXIES

On an even larger scale, the golden ratio can be seen when looking at galaxies. The spiral arms coming from the "bulge" (the centre of a galaxy) approximately follow the shape of a golden spiral.

Figure 2.9: Fibonacci spiral in galaxies (above is Messier 83, a galaxy located 15m light-years away from Earth)

ANCIENT GREEKS

The ancient Greeks seemed to be inspired by the golden ratio. In their buildings and monuments, for example, they used golden rectangles, in other words rectangles whose ratio between long side and short side equalled the golden ratio.

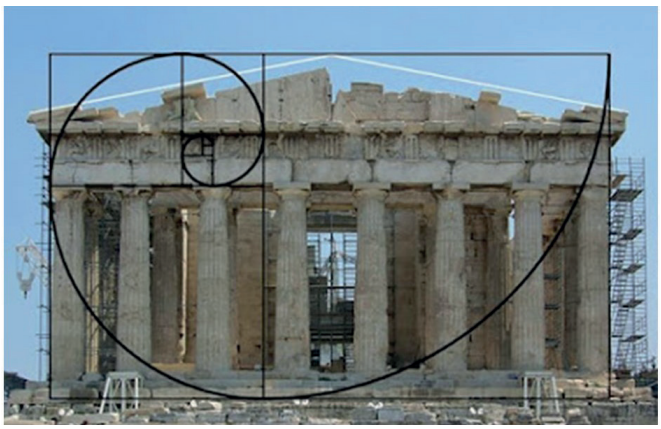


Figure 2.10: Fibonacci spiral in Greek Architecture

OBSERVATIONS

It has been demonstrated that Fibonacci is abundant in the natural world. However, it is important to distinguish between a genuine Fibonacci pattern and a coincidental approximation to Fibonacci.

The idea that rabbits breed in such a way that the population increases following the Fibonacci Sequence is an interesting hypothesis. The conditions for this were idealistic in that it was assumed that rabbits mated a limitless number of times, never died and never experienced infertility or stillbirths. Similarly the number of offspring was assumed to be one pair per couple. Using these assumptions we have shown that the rabbits do exhibit growth in accordance with the Fibonacci Sequence. However, in order to obtain this pattern the assumptions and conditions to model the population growth are simplistic and clearly unrealistic. The actual growth of a rabbit population is dependent on many factors and in reality does not follow Fibonacci. There is mathematical interest in creating the conditions for the Fibonacci Sequence to emerge, but it is largely hypothetical and unrelated to the real-life situation.

Similarly, we have seen that the Fibonacci spiral seems to appear in the study of the geometry of galaxies. However, while the nature of the galaxy seen in figure 2.9 forms a spiral, in reality there are a numerous different shapes for galaxies. As well as spiral galaxies such as Messier 83, there are barred spiral, elliptical and lenticular galaxies. These galaxies do not show any signs of Fibonacci spirals. For example, figure 2.11 shows a lenticular galaxy that appears to have circular features, however there are no noticeable Fibonacci spirals. Moreover, the galaxies that do show spirals make up only 31% of galaxies in the distant universe (NASA, 2013) and while some appear to be similar in shape to the Fibonacci spiral we have seen before, many are merely spirals with a loose resemblance to the true Fibonacci spiral. It seems likely that the general notion that galaxies follow Fibonacci is partly exaggerated based on a confirmation bias. Analysts find galaxies that approximately fit within the model of the spiral and are tempted to apply it to the entire cohort of galaxies.

Figure 2.11: Lenticular Galaxy NGC 524

On the other hand, there is no doubt that in nature and botany in particular, the examples used in this chapter exhibit true Fibonacci characteristics. Considerable scientific work has been carried out over the centuries to confirm these patterns and indeed to establish plausible theories that explain why such patterns occur.

PART THREE: FURTHER FIBONACCI MATHEMATICS

"NON-FIBONACCI SEQUENCE" FIBONACCI NUMBERS

We derived equation [3] to give us the nth Fibonacci number.

$$F_n = \frac{\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1+\sqrt{5}}{2}\right)^n}{\sqrt{5}}$$

So we usually find F_n for a given integer n. But what if we treated this simply as a continuous function? In other words, let us consider the function f(x) for all x:

$$F(x) = \frac{\left(\frac{1+\sqrt{5}}{2}\right)^x - \left(\frac{1+\sqrt{5}}{2}\right)^x}{\sqrt{5}}$$

This function makes sense when x is simply a positive integer, and it produces the Fibonacci Sequence.

Now let us consider negative integers. The formula indeed still works and we can produce a table of results as follows:

x	$f(x)$	x	$f(x)$
-20	-6765.0000	-9	34.0000
-19	4181.0000	-8	-21.0000
-18	-2584.0000	-7	13.0000
-17	1597.0000	-6	-8.0000
-16	-987.0000	-5	5.0000
-15	610.0000	-4	-3.0000
-14	-377.0000	-3	2.0000
-13	233.0000	-2	-1.0000
-12	-144.0000	-1	1.0000
-11	89.0000	0	0.0000
-10	-55.0000		

Figure 3.1: table of x and $f(x)$ values for negative integers x

This starts to look interesting, since although the absolute values of $f(x)$ match the Fibonacci numbers that we obtain using positive integers, we note that the sign alternates between negative and positive. How this happens becomes clearer when we move on to consider non-integer values of x.

We immediately encounter a complication when x is not an integer, since $\left(\frac{1-\sqrt{5}}{2}\right)$ is a negative number. Raising a negative number to a fractional power gives a non-real answer. For example, when $x=0.5$, raising to the power of x is the same as taking the square root. The square root of a negative number is not a real number.

But this should not deter us, since we can use complex numbers. Complex numbers are written in the form of $a+bi$, where a and b are real numbers and $i^2=-1$. They can be plotted on the complex plane, where we use the real portion along the horizontal axis and the complex portion on the vertical axis.

So, we can calculate $f(x)$ for all values of x. What does this look like?

Firstly let's produce a table of values for a range of x values. A fuller table is in Appendix I, but here is an extract. This is a table of values of x and $f(x)$ where $f(x)=\frac{\left(\frac{1+\sqrt{5}}{2}\right)^x-\left(\frac{1-\sqrt{5}}{2}\right)^x}{\sqrt{5}}$.

Here x ranges from -2 to $+3$ and $f(x)$ is over the complex domain.

x		$f(x)$		
-2	-1.0000	+	0.0000	i
-1.9	-0.8820	+	-0.3448	i
-1.8	-0.6722	+	-0.6250	i
-1.7	-0.3983	+	-0.8199	i
-1.6	-0.0914	+	-0.9185	i
-1.5	0.2173	+	-0.9204	i
-1.4	0.4991	+	-0.8343	i
-1.3	0.7306	+	-0.6763	i
-1.2	0.8956	+	-0.4683	i
-1.1	0.9855	+	-0.2346	i
-1	1.0000	+	0.0000	i
-0.9	0.9459	+	0.2131	i
-0.8	0.8360	+	0.3863	i
-0.7	0.6875	+	0.5067	i
-0.6	0.5195	+	0.5677	i
-0.5	0.3516	+	0.5689	i
-0.4	0.2014	+	0.5156	i
-0.3	0.0834	+	0.4180	i
-0.2	0.0078	+	0.2894	i
-0.1	-0.0201	+	0.1450	i
0	0.0000	+	0.0000	i
0.1	0.0639	+	-0.1317	i
0.2	0.1638	+	-0.2387	i
0.3	0.2891	+	-0.3132	i
0.4	0.4281	+	-0.3509	i
0.5	0.5689	+	-0.3516	i
0.6	0.7004	+	-0.3187	i
0.7	0.8140	+	-0.2583	i
0.8	0.9034	+	-0.1789	i
0.9	0.9654	+	-0.0896	i
1	1.0000	+	0.0000	i
1.1	1.0098	+	0.0814	i
1.2	0.9998	+	0.1476	i
1.3	0.9766	+	0.1935	i
1.4	0.9477	+	0.2168	i
1.5	0.9204	+	0.2173	i
1.6	0.9018	+	0.1969	i
1.7	0.8974	+	0.1597	i
1.8	0.9112	+	0.1105	i
1.9	0.9453	+	0.0554	i
2	1.0000	+	0.0000	i
2.1	1.0737	+	-0.0503	i
2.2	1.1636	+	-0.0912	i
2.3	1.2657	+	-0.1196	i
2.4	1.3758	+	-0.1340	i
2.5	1.4893	+	-0.1343	i
2.6	1.6023	+	-0.1217	i
2.7	1.7115	+	-0.0987	i
2.8	1.8146	+	-0.0683	i
2.9	1.9108	+	-0.0342	i
3	2.0000	+	0.0000	i

Figure 3.2: table of x and $f(x)$ values over the complex plane

When we plot $f(x)$ on the complex plane, something wonderful appears:

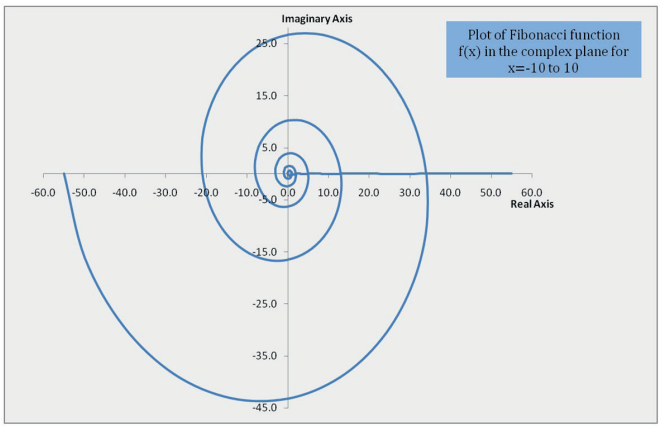


Figure 3.3: the plot of $f(x)$ on the complex plane produces a spiral pattern

This is an extremely interesting graph. Not only do we see a beautiful spiral pattern, reminiscent of the spirals we saw in golden rectangles and in the nautilus, but the plot continues along the real axis in a remarkable way. We can "zoom in" by looking at the graph that is produced when x ranges from -3 to 4:

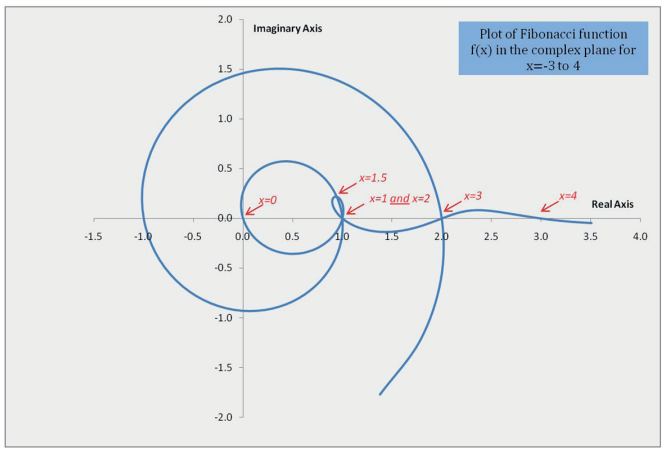


Figure 3.4: closer examination shows the Fibonacci numbers where the graph crosses the real axis

We can see how the spiral loops around and around, repeatedly crossing the real axis. If we note where it crosses the real axis we see that it is at the values of the Fibonacci Sequence.

As x increases, the graph undulates above and below the real axis – getting closer and closer to it but only ever crossing at an actual integer Fibonacci value. We can also observe that the nature of the spiral in the complex plane shows why for negative integers the sign changes – the graph crosses first the positive part of the real axis then spirals round to cross the negative part of the real axis and so on.

TRIBONACCI SEQUENCE AND BEYOND

The "Tribonacci numbers", T_n, are a generalization of the Fibonacci numbers defined by $T_1=1, T_2=1, T_3=2$ and the recurrence equation $T_n=T_{n-1}+T_{n-2}+T_{n-3}$ for integers $n>3$.

The first few terms using the above 1, 1, 2, 4, 7, 13, 24, 44, 81, 149, ...

Using a similar technique to finding Binet's Formula, let us assume the general solution is of the form $T_n=B^n$

Then the recurrence condition gives $B^n=B^{n-1}+B^{n-2}+B^{n-3}$

And dividing both sides by B^{n-3} gives $\frac{B^n}{B^{n-3}}=\frac{B^{n-1}}{B^{n-3}}+\frac{B^{n-2}}{B^{n-3}}+\frac{B^{n-3}}{B^{n-3}}$

$\Rightarrow B^3=B^2+B^1+B^0$

$\Rightarrow B^3-B^2-B-1=0$

Thus the roots of this cubic equation form the basis for the general solution of the Tribonacci sequence. The general solution for a cubic equation is well known but is extremely "messy". Thus the solutions of $ax^3+bx^2+cx+\partial=0$ are given by

$$x=\left(\frac{-\sqrt{3}i}{2}-\frac{1}{2}\right)\left(\frac{\sqrt{27a^2d^2+|4b^3-18abc|d+4ac^3-b^2c^2}}{6\sqrt{3}a^2}-\frac{27a^2d-9abc+2b^3}{54a^3}\right)^{\frac{1}{3}}-\frac{\left(\frac{\sqrt{3}i}{2}-\frac{1}{2}\right)(3ac-b^2)}{9a^2\left(\frac{\sqrt{27a^2d^2+|4b^3-18abc|d+4ac^3-b^2c^2}}{6\sqrt{3}a^2}-\frac{27a^2d-9abc+2b^3}{54a^3}\right)^{\frac{1}{3}}}-\frac{b}{3a}$$

$$x=\left(\frac{\sqrt{3}i}{2}-\frac{1}{2}\right)\left(\frac{\sqrt{27a^2d^2+|4b^3-18abc|d+4ac^3-b^2c^2}}{6\sqrt{3}a^2}-\frac{27a^2d-9abc+2b^3}{54a^3}\right)^{\frac{1}{3}}-\frac{\left(\frac{-\sqrt{3}i}{2}-\frac{1}{2}\right)(3ac-b^2)}{9a^2\left(\frac{\sqrt{27a^2d^2+|4b^3-18abc|d+4ac^3-b^2c^2}}{6\sqrt{3}a^2}-\frac{27a^2d-9abc+2b^3}{54a^3}\right)^{\frac{1}{3}}}-\frac{b}{3a}$$

$$x=\left(\frac{\sqrt{27a^2d^2+|4b^3-18abc|d+4ac^3-b^2c^2}}{6\sqrt{3}a^2}-\frac{27a^2d-9abc+2b^3}{54a^3}\right)^{\frac{1}{3}}-\frac{3ac-b^2}{9a^2\left(\frac{\sqrt{27a^2d^2+|4b^3-18abc|d+4ac^3-b^2c^2}}{6\sqrt{3}a^2}-\frac{27a^2d-9abc+2b^3}{54a^3}\right)^{\frac{1}{3}}}-\frac{b}{3a}$$

If we substitute in $a=1, b=-1, c=-1$ and $\partial=-1$ then the last of these solutions becomes

$B_1=\frac{1+\sqrt[3]{(19-3\sqrt{33})}+\sqrt[3]{(19+3\sqrt{33})}}{3}$ which is approximately 1.8393...

The other solutions B_2 and B_3 are complex numbers and the general solution to the nth Tribonacci number is a linear combination of the three solutions, namely

$$T_n=\frac{B_1^n}{-B_1^2+4B_1-1}+\frac{B_2^n}{-B_2^2+4B_2-1}+\frac{B_3^n}{-B_3^2+4B_3-1}$$

Now, as $n\rightarrow\infty$ the ratio of T_{n+1} to T_n tends to the ratio of the real solutions (we show below the reasoning for this), i.e.

$$\frac{\left(\frac{B_1^{n+1}}{-B_1^2+4B_1-1}\right)}{\left(\frac{B_1^n}{-B_1^2+4B_1-1}\right)}=B_1$$

We can test this. Let's take the ratio of 149 to 81 which is 1.8395... which is very close to the exact value for B_1 above.

Fibonacci and Tribonacci have been extended in many ways. For example, Tetranacci numbers start with four predetermined terms, each term afterwards being the sum of the preceding four terms. The first few Tetranacci numbers are:

0, 0, 0, 1, 1, 2, 4, 8, 15, 29, 56, 108, 208, 401, 773, 1490, 2872, 5536, 10671, 20569, 39648, 76424, and so on.

It can be shown (see below) using similar logic as we have applied to the Fibonacci and Tribonacci sequences, that the ratio of consecutive terms converge to the real solution of the quartic equation $x^4-x^3-x^2-x-1=0$, approximately 1.9276...

Again, let's test the ratio of 10,671 to 5536 which gives 1.9276.

Pentanacci, Hexanacci, and Heptanacci numbers can all been computed (being the sum of the last 5, 6 and 7 numbers respectively):

Pentanacci numbers: 0, 0, 0, 0, 1, 1, 2, 4, 8, 16, 31, 61, 120, 236, 464, 912, 1793, 3525, 6930, 13624, ...
(here the ratio of consecutive numbers tends to 1.966...)

Hexanacci numbers: 0, 0, 0, 0, 0, 1, 1, 2, 4, 8, 16, 32, 63, 125, 248, 492, 976, 1936, 3840, 7617, 15109, ... (ratio tends to 1.984...)

Heptanacci numbers: 0, 0, 0, 0, 0, 0, 1, 1, 2, 4, 8, 16, 32, 64, 127, 253, 504, 1004, 2000, 3984, 7936, 15808, ... (ratio tends to 1.992...)

One could go on and compute Octanacci numbers and Enneanacci numbers but we will pause here. We can plot these ratios for the "n-nacci" series:

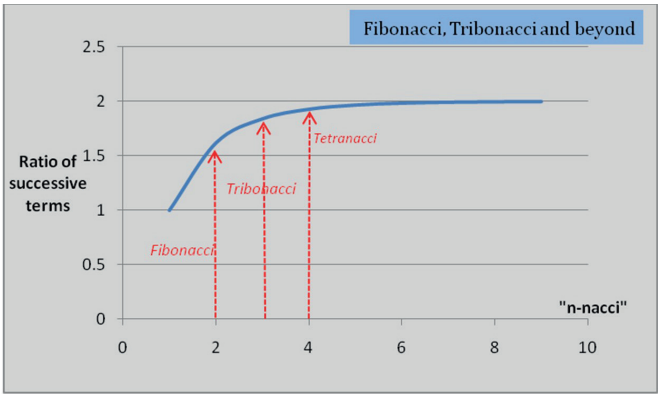


Figure 3.5: ratio of successive terms for Fibonacci, Tribonacci, Tetranacci and so on

Consider an n-nacci sequence defined by
$Z_n=Z_{n-1}+Z_{n-2}+Z_{n-3}+...+Z_{n-m}$
where m is the order of the n-nacci sequence (i.e. for the Tetranacci $m=4$, for the Pentanacci $m=5$ and so on).

So again if we try a solution of the form $Z_n=x^n$ then we get

$x^n=x^{n-1}+x^{n-2}+x^{n-3}+...+x^{n-m}$

$\Rightarrow x^m=x^{m-1}+x^{m-2}+x^{m-3}+...+1$ (dividing by x^{n-m})

$\Rightarrow x^m-x^{m-1}-x^{m-2}-x^{m-3}-...-1=0$ (rearranging) equation [4]

Firstly, we consider real solutions x.

Now if $0\le x\le 1$ then x^{m-1} will be greater than x^m since x^m is between 0 and 1 and is being raised to a higher power. Similarly x^{m-2} is greater than x^{m-1}, and so on. This however creates a contradiction since it implies $x^m-x^{m-1}-x^{m-2}-x^{m-3}-...-1$ is less than 0. Therefore x cannot be between 0 and 1.

Further, if $x\le -1$, then (i) if m is even x^m is positive and greater than 1 and the following odd power x^{m-1} is negative and greater or equal in absolute terms than the following positive term x^{m-2}, hence $x^m-x^{m-1}-x^{m-2}-x^{m-3}-...-1$ is always greater than 0; (ii) if m is odd x^m is negative and greater in absolute terms than 1 and the following odd power x^{m-1} is positive and greater or equal in absolute terms than the following negative term x^{m-2}, hence $x^m-x^{m-1}-x^{m-2}-x^{m-3}-...-1$ is always less than 0. Therefore x cannot be ≤ -1.

Therefore we have shown that any real solutions are in the range $\{-1<x<0 x>1$

According to the Fundamental Theory of Algebra (which states that a polynomial of degree n has, when counted with repeated roots, exactly n roots, real or complex), there may be complex number solutions that also satisfy

$x^m-x^{m-1}-x^{m-2}-x^{m-3}-\ldots-1=0$ For this to be the case, if the solution is in the form $re^{i\theta}$ (for $r\geq 0, 0\leq\theta\leq 2\pi$) it has been shown that $|r|\leq 1$ (Shevelev, 2014).

Now consider the general solution,

$$Z_n=a_1 x_1^n+a_2 x_2^n+\ldots+a_m x_m^n$$

where a_s are constants found from the initial starting conditions of the sequence and x_s are the solutions (real and complex) of $x^m-x^{m-1}-x^{m-2}-x^{m-3}-\ldots-1=0$

Then the limit of successive terms becomes $\lim_{n\to\infty} \frac{a_1x_1^{n+1} + a_2x_2^{n+1} +\ldots+a_{n+1}x_{n+1}^{n+1}}{a_1x_1^n + a_2x_2^n +\ldots+a_nx_n^n}$

In this limit, any complex roots x_s tend to 0 as n approaches infinity (since $|r|\leq 1$ from above), and any real roots in the range $-1<x<0$ also tend to 0. This leaves the case of the real solution where $x>1$. If we assign its coefficient as a_1 then the limit becomes $\frac{a_1x_1^{n+1}}{a_1x_1^n}=x_1$ where $x_1>1$.

In other words, the ratio of successive terms of the n-nacci sequence tends to the positive real root as n approaches infinity.

For an "n-nacci" sequence, we have shown that the limit of the ratio of successive terms approaches the real root of

$x^n-x^{n-1}-x^{n-2}-\ldots-x^n2-x-1=0$ (equation [4] above)

$\Rightarrow x^{n+1}-x^n-x^n-\ldots-x^3-x^2-x=0$ (multiplying by x)

$\Rightarrow x^{n+1}-x^n-(1+x+x^2+\ldots+x^{n-1})+1=0$ (rearranging)

$\Rightarrow x^{n+1}-x^n-(x^n)+1=0$ (since $(1+x+x^2+\ldots+x^{n-1})=x^n$ from equation [4])

$\Rightarrow x^{n+1}-2\,x^n+1=0$

$\Rightarrow x^{n+1}+1=2\,x^n$

$\Rightarrow x+x^{-n}=2$ (dividing by x^n)

Hence in general for a "n-nacci" sequence, the limit of the ratio of successive terms of an n-nacci series tends to a root of the equation $x+x^{-n}=2$, and in the limit as $n\to\infty$ it is easy to see that $x\to 2$ (since we have shown above that the positive real root satisfies $x>1$ and therefore in the limit x^{-n} tends to zero), which is consistent with the graph above.

This is only a segment of Sam's ILA. The rest discussed the use of the sequence in stock markets (specifically technical analysis).

What would have to change about 'democracy' in order to restore faith in democracy among young people?

Finalist in the 2023 Northeastern University Essay Competition

Joshua Inglesfield, Lower Sixth

Young people – who I shall class as anyone aged 16-24 (taking the 18-24 grouping used by Parliament and extending it to include those who may be enfranchised in the future) – are the future of democracy, and thus it is critical that they have faith in its operation; lest we fall into the enclave of authoritarianism. An increasing number of protests worldwide and a surge in populism signals that youth are tired of democracy's inefficiency. Populist success can be seen worldwide – from the historic city of Rome where you can find the newly elected far-right Fratelli d'Italia, to Orban in Budapest, across the Mediterranean to Syriza in Greece – the list goes on. Correlations drawn with figures showing that 55% of Italian youth no longer believe that democracy 'is the best form of government' – 7% higher than the average for European youth[1] – demonstrates that the rise of 'Fratelli d'Italia' is alongside a growing lack of faith in democracy. This is no coincidence and is happening across the globe. Thus, it is clear a solution is needed.

Direct democracy would appear to be the perfect solution to loss of faith in democracy among youth – the turnout for the 2016 Brexit referendum being 10% higher than that for the 2017 election among 18-24 year olds[2] is evidence enough that youth prefer a form of direct democracy. Not only would this give young people the impression that they could make a tangible difference, but it would also reduce this notion of 'democratic disconnect'[3] – the alienation of young people from democratic processes. Youth also have a lack of trust in governments – with such a process young people will be confident that governments will no longer be able to 'sell' policy decisions to the highest bidder through party donations to as great an extent. Further to this, Colin Crouch argues that; 'democracy requires the formal mechanisms of citizen participation but also proof of genuine political agency'- which in the eyes of young people is not being fulfilled, seeing little 'political agency' (actual actions) taking place with regards to their concerns. Consequently, we can conclude that young people would, by Crouch's argument, be seeing a failure and consequently be having a lack of faith in democracy, due to this perceived absence of 'political agency'– a situation Crouch labels a 'post-democracy'[4]. Such an implementation would deal with the perceived lack of action alongside strengthening 'citizen participation' and so increase faith in democracy. But there is a significant drawback to this suggestion. Imagine you wake up to a notification on your phone – notice of the 2nd referendum this month. Before you can even consider the proposition you must go to work, cook dinner, and go to the supermarket. 349 minutes[5] – the average amount of 'leisure time' per day for Britons – is all you have left. 349 minutes dwarfed by the amount of time Public Bill Committees spend inspecting a bill, and certainly too little time to properly understand the subject of the referendum. This is the constraint of time. The average person simply does not have enough of it to consider the wider implications of their vote, nor how the policy enacted by the referendum might fit in with existing policy. Consequently, their voting behaviour will become a lottery, an impulse on the day rather than a considered vote. So, while direct democracy may seem inviting, once realised the population would find themselves confused, overwhelmed, and not able to make a decision to benefit even themselves. Thus, if this were to take place the number of referendums would have to be strictly limited, and be on larger, more straightforward questions such as capital punishment.

The voting age is a hotly disputed topic in British politics. For years groups such as the Electoral Reform Society[6] have campaigned for the voting age to be lowered to 16 – mentioning arguments such as increasing political participation for generations to come[7] – but few cite increasing faith in democracy as the primary argument. The Electoral Reform Society's argument is a valid one – they argue that if you "don't vote, you are less likely to vote in future"[8] – and that by enfranchising these new groups we could encourage greater lifelong participation. This would have the additional benefit of increasing faith in democracy, increasing involvement and again reducing a democratic disconnect to youth– with Dr Foa and Dr Mounk writing that in the UK young people are less likely to vote for the often-anti-democratic populists when 'mobilised to vote'[9] – which here would be enfranchising 16- and 17-year-olds. An additional argument for lowering the voting age being the solution to declining faith in democracy among young people is the idea that when youth are not directly involved in democracy, they lose faith in it[10]. This action would

1. TUI Stiftung/YouGov. (2017). "Young Europe 2017: The Youth Study of the TUI Stiftung." www.tui-stiftung.de/wp-content/uploads/2017/05/All-results-TUI-Stiftung_European-Youth.pdf .
2. Stephan Mashford/89 Scotland. (2020). "Youth turnout – How does the UK compare to other European nations?" https://89initiative.com/youth-turnout-uk-europe/ .
3. Foa, R.S., Klassen, A., Wenger, D., Rand, A. and M. Slade. (2020) "Youth and Satisfaction with Democracy: Reversing the Democratic Disconnect?" https://www.bennettinstitute.cam.ac.uk/wp-content/uploads/2022/06/Youth_and_Satisfaction_with_Democracy-lite.pdf .
4. C. Crouch. (2004). Post-Democracy. Cambridge, United Kingdom: Polity Press
5. ONS. (2017). "Leisure time in the UK: 2015" https://www.ons.gov.uk/economy/nationalaccounts/satelliteaccounts/articles/leisuretimeintheuk/2015 .
6. Electoral Reform Society. (2017). "Background on Votes at 16" https://www.electoral-reform.org.uk/latest-news-and-research/parliamentary-briefings/votes-at-16/ .
7. Electoral Reform Society. (date not disclosed). "Votes at 16" https://www.electoral-reform.org.uk/campaigns/votes-at-16/ .
8. Electoral Reform Society. (date not disclosed). "Votes at 16" https://www.electoral-reform.org.uk/campaigns/votes-at-16/ .
9. R.S. Foa/Y. Mounk. (2019). "Youth and the populist wave" https://journals.sagepub.com/doi/full/10.1177/0191453719872314 .
10. A. Correia. (2021). "The necessity of youth support in sustaining democracy" https://diplomatmagazine.eu/2021/11/20/the-necessity-of-youth-support-in-sustaining-democracy/ .

therefore seem to fulfil all criteria to increase faith in democracy among youth – but there is an obvious drawback – nothing has changed for the currently enfranchised youth. Such a change would therefore do nothing to deal with the current decline in faith in democracy among the ages 18-24, a dangerous risk given that these are the ages which are already propelling extreme populists to power in nations such as Greece. Far from ameliorating the situation, this would risk escalating it. The youth ignored by such a reform may feel further alienated and see another failure of democracy to criticise, one that risks pushing the democratic disconnect to an irretrievable state of separation between democracy and young people.

First Past The Post (FPTP) – a voting system which suppresses the votes of millions. That is, from a critic's perspective – but the fact is that FPTP's nature ensures that only two large parties can ever realistically hold government, a feature which while does produce strong majoritarian governments (usually – 2010 Conservative and Liberal Democrat coalition is a notable exception), results in smaller parties receiving almost no seats. But why is this a problem regarding faith in democracy? If we take the argument that the principal reason for loss of faith is not seeing action, would not FPTP be the obvious choice, empowering a strong government to take decisive action without being hindered by Parliamentary squabbles or half-baked coalitions? Those arguments certainly hold some water; however, the issue of representation must be raised. One of the issues young people are most concerned with is climate; so many may support the Green Party; but despite getting 2.7% of the vote share across the UK in the 2019 general election, they only received approximately 0.15% of seats available[11]. Thus, many young people who voted for a party that would pioneer their beliefs have been left unrepresented. This feeling of being unrepresented will likely lead to dissatisfaction and a lack of faith in democracy, as the problem lies in the very essence of democracy, the voting system. The clearest solution would be proportional representation – as used by 40 European nations[12]. Such a system would ensure that smaller parties pioneering the views of minorities or smaller groups such as young people are heard and would allow for greater political pressure to be applied for tangible action. Critics, however, would argue that it gives opportunity to potentially dangerous populist parties such as Syriza, or even extremist ones as seen with the rise of the Nazi party under proportional representation, portraying it as a vile breeding ground for hate. However, it is necessary to note that in modern democracy this is rarely the case to such an extent, with parties such as 'Alternative for Germany'[13] being kept out of government – in fact, it almost seems as if FPTP is the envy of populists at present, leading to Trumpism and pro-Brexit populist groups rising to power and succeeding.[14]

To conclude, young people will need to see a change to the very structure of democracy to prevent further decline in faith in democracy – with it being imperative that these changes are not superficial PR stunts but tangible changes. What is needed is a two-fold implementation – With this in mind, I would suggest that what is necessary for the UK is the simpler change of increasing the number of referendums to involve youth to a greater extent in democracy, and the more structural change of shifting to proportional representation as a system to give the silenced minority parties a voice. These two implementations would allow for an increase in participation in democracy, which in turn would lead to an increase in faith in it as young people see their policy aspirations manifest into tangible change. Thus, as Aiden Correia writes; 'democracy is about providing everyone with a voice. The youth are willing to talk; governments just need to start to listen'[15] – through the measures outlined above we can fight the democratic apathy of young people before it spills over into antipathy.

11. BBC News. (2019). "Election 2019 Results" https://www.bbc.co.uk/news/election/2019/results .
12. M. Palese/Electoral Reform Society. (2018). "Which European countries use proportional representation?" https://www.electoral-reform.org.uk/which-european-countries-use-proportional-representation/ .
13. L. Drutman. (2022). "10 Ideas to Fix Democracy – Abolish Two-Party Systems" https://foreignpolicy.com/2022/01/07/10-ideas-fix-democracy/ .
14. L. Drutman. (2022). "10 Ideas to Fix Democracy – Abolish Two-Party Systems" https://foreignpolicy.com/2022/01/07/10-ideas-fix-democracy/ .
15. A. Correia. (2021). "The necessity of youth support in sustaining democracy" https://diplomatmagazine.eu/2021/11/20/the-necessity-of-youth-support-in-sustaining-democracy/ .

Establishing the Effect of MOF Particle Size on Uptake and Release of Semiochemicals

A research report written following an Original Research in Science (ORIS) placement

Alexander McDougall, Upper Sixth

ABSTRACT

Metal-organic frameworks (MOFs) such as HKUST-1 have the potential to function in sustainable agricultural practices as an alternative to pesticides by adsorption and release of semiochemicals which attract pests. This study explores the effect of particle size of the MOF on the uptake and release of semiochemicals. The varying of MOF particle size in the HKUST-1 copper (II) MOFs was achieved through altering the feedstock of copper used in synthesis. This produced different rates of uptake and release of the two semiochemicals used, 1-hexanol and 3-octanone, both of which are used as pest attractants. It was found that the larger sized particles have greater uptake of both semiochemicals. Thus, we can tune semiochemical release through particle size control in MOFs.

INTRODUCTION

With world hunger on the rise and an increase of up to 150 million undernourished people from 2019 to 2022,(1), food poverty is one of the world's most pressing problems. The current widespread use of pesticides, while broadly effective in food production, has led to a variety of unintended negative consequences. As well as the growing challenge of pesticide resistance, one such consequence is the damage to local biodiversity as seen by the 70% reduction of insect biomass in Germany over the last few decades,(2). Consequently, in the transition towards sustainable agriculture, which is one of the seventeen sustainable development goals set out by the UN in 2015 (3), the need to develop improved and sustainable crop protection techniques is of paramount importance. One proposed solution is the use of lure traps, where a pheromone (a subset of semiochemicals which trigger a response from members of the same species) evaporates from a container and attracts the pest towards the trap. However, due to the inherent volatility of pheromones, this method would be impractical as the containers would have to be replenished far too often than is practical. Consisting of metal ions coordinated to organic linker molecules in a regular 3D crystalline structure, metal organic frameworks (MOFs) offer a solution to this problem by adsorbing and then releasing pheromones over a sustained period independently. With a high porosity and a very large surface area (typically ranging from 1,000 to 10,000 m2/g (4)), these hybrid structures are ideal structures for adsorption. These unique properties of MOFs allow for large uptakes of pheromones, resulting in a longer release period and so a greater number of pests attracted towards the bait and away from the crop. Furthermore, their pore size can be altered to adsorb specific substances by changing the constituent metal salt or ligand. This versatility of MOFs offers a new mechanism with great potential for pest management since the MOF, unlike pesticides, can attract different kinds of pests over a long time period without the need for regular application. The pheromones used by organisms in the real world, which usually consist of a mixture of chemicals, can be mimicked by using just their main chemical component. A previous study used 3-octanone to recreate the effects of the ant alarm pheromone of the Atta and Acromyrmex leaf-cutting ant species, which cause an estimated $8 billion damage to Eucalyptus forestry in Brazil every year (5). To help combat the damage caused by ambrosia beetles, such as the destruction of 80% of redbay trees caused by redbay ambrosia beetles in areas of southern US (6), the pheromone attractant for ambrosia beetles 1-hexanol was used. These 2 semiochemicals were used to investigate the effect of particle size of the MOF on the inclusion and release of semiochemicals, with the overall aim of finding the most effective MOF size and structure to use as an alternative to pesticides. The Cu-MOF HKUST-1 was chosen for this study as it is highly prevalent in the literature and its particle size can be easily tuned.

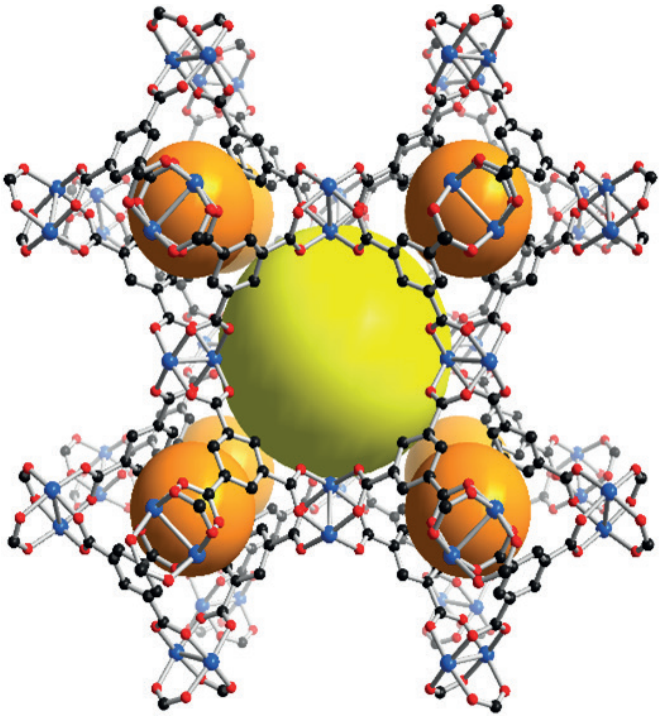


Figure 1: The structure of an HKUST-1 MOF. The large spheres indicate the pore space.

RESULTS & DISCUSSION

Five different HKUST-1 samples with different particle sizes were prepared by dissolving the linker, trimesic acid, and five different copper salts in DMF followed by solvothermal synthesis in a microwave. Full details are supplied in the experiment details. After synthesis, the 5 HKUST-1 MOFs underwent X-Ray Diffraction (XRD) analysis to determine the crystallographic structure of the samples. This technique involves incident X-rays of wavelength similar to the atomic spacing in the crystal lattice being scattered by the regularly spaced atoms. When the scattered

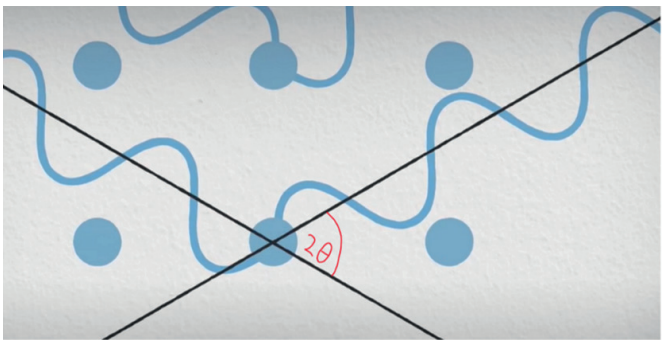


Figure 2: The incident X-ray beam from the left is absorbed by electrons surrounding the atom and then re-emitted at a new angle, 2θ , relative to the incident angle.

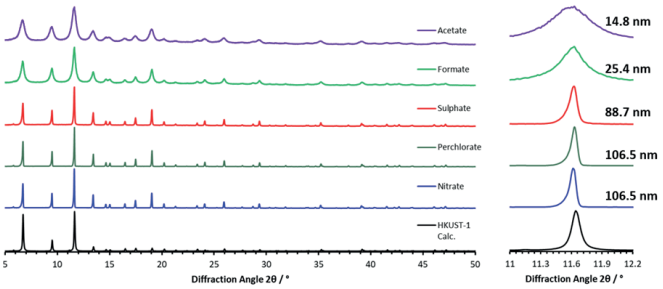


Figure 3: XRD analysis of the 5 HKUST-1 MOFs. The HKUST-1 calculated value is of the copper HKUST-1 MOF

waves constructively interfere, the displacement of each wave is added to form a new wave with greater displacement, resulting in the peaks as seen in Fig. 3. The angle between the incident and scattered beam is called 2Θ. The scattering of X-rays at certain angles allows the crystalline phases present to be identified in order to produce an XRD pattern unique to each sample. The more crystalline the structure, the more intense the peaks Before being placed into the sample holder for XRD, the MOF sample was crushed into a power to ensure total randomness of the crystallite orientations so that the whole structure of the crystal would be accurately represented.

Calculated from the Scherrer equation (see Appendix 1), the different crystallite domain sizes on the right in Figure 3 give an indication of particle size, which can be compared to the calculated value of the large sized particles of the HKUST-1 Nickel MOF. The sharpness of the peaks gradually increases in Figure 3 from acetate down to nitrate, indicating that, out of the 5 samples, the nitrate MOF ions are the most ordered to give the most crystalline structure. The relative intensity of the peaks shows the same trend as the width, showing that nitrate has the biggest particle sizes, in contrast with the broad and less intense peaks of the smallest particle acetate.

This pattern could be due to the smaller particles such as those in acetate moving around more and so not aligning in a crystal structure as much as the bigger particles which are less inclined to move around, instead tending towards the crystal structure. The paddle-wheel shape of the HKUST-1 framework, which helps give the MOF its rigidity, is shown in Figure 3, where the Copper 2+ ions are in blue, oxygen in red, carbon in grey and hydrogen in white.

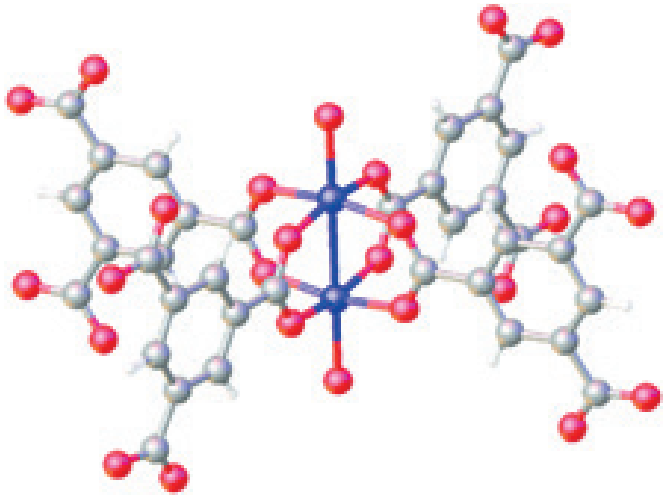


Figure 4: The paddle-wheel structure adopted by HKUST-1 MOFs with Cu2+ cations and benzene tricarboxylic acid anions (7)

To demonstrate their thermal stability, the MOFs underwent thermogravimetric analysis (TGA), shown in Figure 5, where the mass of the sample was measured over time as the temperature

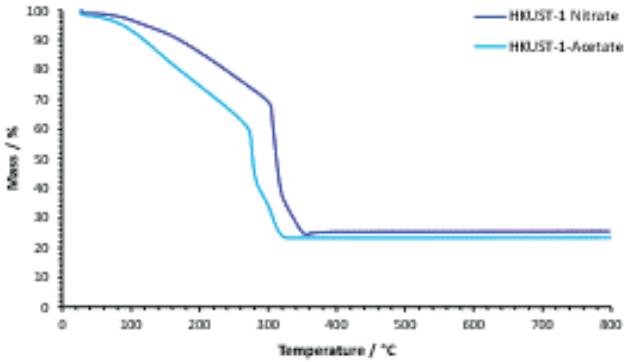


Figure 5: Thermogravimetric analysis of Nitrate and Acetate

increased. The nitrate MOF is evidently more thermally stable than the acetate MOF as it maintains a greater percentage of its mass throughout the heating period. The initial decrease in mass due to the removal of water and other chemicals used in the synthesis is followed by a steep gradient where the MOF decomposes. The nitrate MOF has a higher decomposition temperature of around 300°C than that of the acetate MOF at about 270°C. Although this shows that the nitrate MOF structure is more thermal stable, the MOFs are unlikely to be subjected to such high temperatures when used so this difference is not of huge significance. However, oven temperatures exceeding 100°C are used in the synthesis and activation processes, meaning it is likely that a greater percentage of the acetate MOFs is lost due to thermal composition. This gives the bigger particle sized nitrate MOFs an advantage in the practicality of the synthesis. Only acetate and nitrate were used because the XRD patterns of the synthesized samples, shown in Fig. 3, showed that they had the smallest and biggest particle sizes respectively so all the other MOFs would fall somewhere in between these 2 extremes and thus would not be necessary to answer the question of whether particle size affects the uptake and release of semiochemicals.

To prove that the pores of the MOFs had successfully been emptied and that the crystalline structure had been maintained after activation (the process of emptying the MOF pores through washing) XRD and 1H NMR analysis of the activated sample was done.

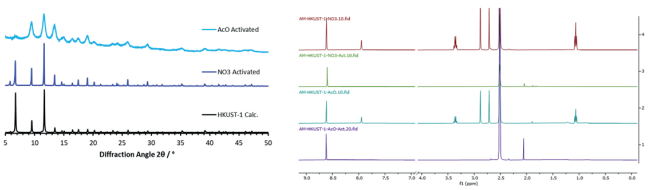


Figure 6 (a): XRD pattern of nitrate and acetate MOFs.

Figure 6 (b): 1H NMR of the nitrate and acetate MOFs before and after activation.

Figure 6(a) shows that the MOFs have maintained their structures after being washed with harsh chemicals like acetonitrile because the same intensity of peaks appear at the same diffraction angles as Figure 2(a). This is important because it demonstrates that the MOF structures are stable and will not be altered by changing conditions. The NMR in Figure 6(b) shows that the peaks of chemicals used in the synthesis (such as those of DMF between 2.5 and 3.0ppm, and diethyl ether at 3.3 and 1.1ppm) are absent in the activated samples, evidence that the pores have been successfully emptied and are ready to be loaded with semiochemicals. This means that the maximum amount of

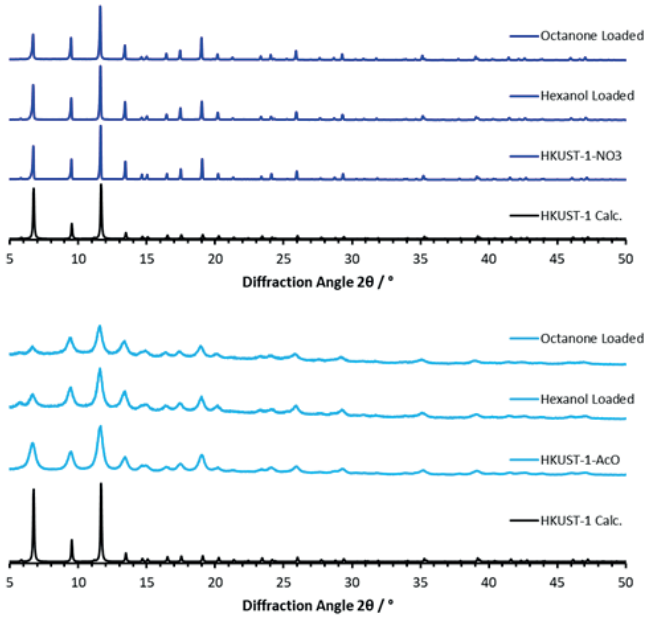


Figure 7: XRD patterns of Nitrate and Acetate MOFs before and after loading showing little to no change.

semiochemicals can be adsorbed into the MOF structure and then released, which is important when trying to attract as many pests as possible.

These activated nitrate and acetate copper MOFs were then loaded with the 2 semiochemicals 1-Hexanol and 3-octanone, and the MOFs kept their crystalline structures and were once again unchanged as shown in Figure 7.

This finding reinforces the idea that these MOF structures are very stable and indicates that they are suitable for their purpose of storing guest semiochemicals used in agriculture.

Figure 8 shows quick uptake of semiochemicals of both the nitrate and acetate MOFs as they reach their maximum capacity before plateauing and reaching equilibrium with the surroundings. However, the nitrate MOFs have a significantly larger capacity to adsorb the semiochemicals in both loadings. This suggests that the bigger particle sizes of the nitrate MOFs result in a greater.

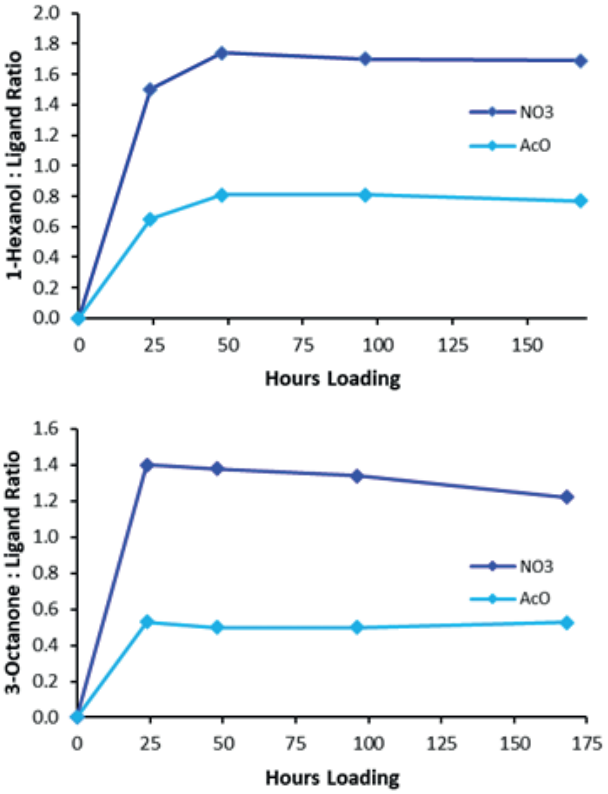
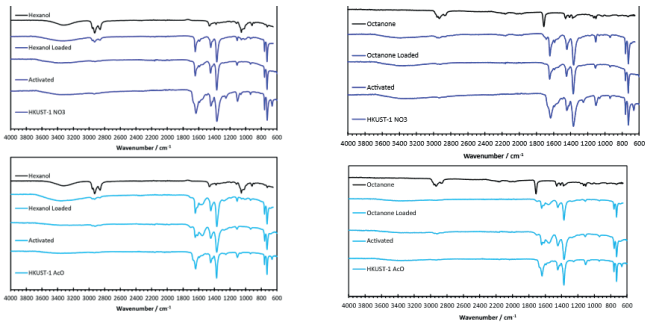


Figure 8: Loading into nitrate and acetate MOFs of semiochemicals (a) 1-Hexanol and (b) 3-Octanone

volume of semiochemicals adsorbed, thus making them more suitable as a pest attractant. They are also suitable because of their ability to store a high percentage of their maximum adsorption and maintain a stable structure with the semiochemicals incorporated into the MOF structure for several days after the initial loading.

FTIR analysis was then carried out to confirm the presence of the semiochemicals in the MOFs through the appearance of their functional groups in the spectrums of the loaded MOFs, shown in Figure 9. Each functional group contains certain bonds which stretch and vibrate at unique frequencies (and therefore unique wavelengths). When infrared radiation passes through the sample, these bonds absorb the IR of wavelength similar to the wavelength of the stretching and vibrating bonds, whilst the rest travels through unchanged. A graph of the percentage of transmittance against wavenumber (the inverse of wavelength) can then be plotted, giving peaks where the transmittance drops down because the IR has been absorbed by a functional group. In this way, the identification of different functional groups is relatively straightforward since each functional group absorbs a particular frequency of IR (though there can be some overlap between different groups).



(a) (b)

Figure 9: FTIR spectrums of the nitrate and acetate MOFs before and after loading with semiochemicals

Figure 9(a) shows a broad trough at roughly 3300cm^{-1} and a more intense peak around 2900cm-1, both representing the alcohol group of 1-hexanol, present in the spectrum of the loaded MOF, but not in the activated or original MOF spectrums. Likewise, the peak at 2900cm^{-1} in Figure 9(b) indicating the ketone group from 3-octanone appears in the loaded MOF spectrum. Interestingly, the sharp peak representing a ketone group at 1700cm^{-1} in 3-octanone turns into a less intense shoulder-type peak just to the left of the fingerprint region, a region of low wavenumber predominantly caused by C-H bending vibrations which is unique to every compound. When forming part of the paddle wheel structure of the MOF, the oxygen on the carbonyl group has a lone pair of electrons. This allows its pi orbitals to overlap with the vacant d orbitals of the copper at the centre to form a coordination bond with the copper. This shift in electron density means the carbonyl group vibrates with less energy, which means it has a smaller frequency according to the equation $E=hf$ where h is a constant. Since wavelength is inversely proportional to frequency, the wavelength absorbed by this bond will increase, and so the wavenumber will decrease and thereby shift to the right.

After loading the nitrate and acetate MOFs with semiochemicals, the MOFs were ready to start the release process. Figure 10 shows that the 2

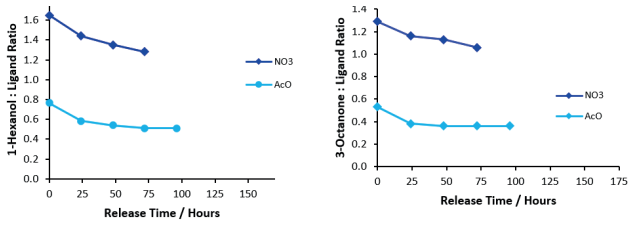


Figure 10: Release of semiochemicals 1-Hexanol and 3-Octanone from nitrate and acetate HKUST-1 MOFs

MOFs, kept at a constant temperature of 40°C, have a similar profile, except for the fact that the nitrate MOF had over double the amount of semiochemicals adsorbed within the structure to begin with. The rate of release gradually decreases over the first few days after a relatively large decrease in the first day where the semiochemical is released from the outer surface of the MOFs. Whilst the amount of semiochemicals is slowly decreasing in the nitrate MOFs, the release profiles for the acetate MOFs appear to have levelled off as the amount of semiochemical has remained unchanged. This would make the acetate MOFs far from ideal if they had to be replaced after 3 days because they stopped releasing the pest attractant. The evidence thus far indicates that the bigger particle sizes of the nitrate MOFs give a better, more consistent release profile of semiochemicals compared to the smaller particles of the acetate MOFs which appear to have stopped releasing after 3 days. This is primarily because the bigger particle sizes lead to a much greater capacity to adsorb semiochemicals, so more semiochemicals can be released.

CONCLUSION & FUTURE WORK

It was found that different copper salts, such as copper sulphate or copper nitrate, form HKUST-1 MOFs with different particle sizes. This encouraging finding opens the possibility of a huge variety of copper salt MOFs of extreme sizes to explore the effects of particle size on inclusion and release of semiochemicals. The different particle sizes, having been shown by XRD, could be confirmed by SEM imaging. MOFs of both large and small sizes were able to be loaded with semiochemicals, but larger MOFs had a much greater uptake of over double the amount. This makes the larger MOFs more practical since they have more semiochemicals to release. To verify that it is the changing of particle size that causes a difference in uptake and release, and not just the different copper salts and their different structures, the particle sizes could be controlled by modulating agents which prevent large extending crystals from forming, before measuring their uptake and release. The changing of metal could be investigated, such as using cobalt instead of copper as the metal node, to find out how particle size affects uptake and release in structures different from the paddle-wheel of the HKUST-1 Copper MOF used in this study. The rates of release for the differently sized MOFs were also different, with the bigger MOFs giving a slow but steady release unlike the smaller MOFs whose release appeared to halt after only 3 days. Overall, particle size was found to have a significant impact on the inclusion and release of semiochemicals; the larger particles showed a larger uptake and a more consistent release compared to the smaller MOFs, and were also more thermally stable. These qualities make the larger MOFs better suited to their potential role as a sustainable alternative to pesticides, and this application could be trialled in the field in the future to see how the release profiles differ in a changing environment and assess the effectiveness of the MOFs.

REFERENCES

1. (2022) World Hunger: Key Facts and Statistics 2022, Action Against Hunger, viewed on 1/8/22: https://www.actionagainsthunger.org/world-hunger-facts-statistics
2. Bruhl, C., Zaller, J., (2019) Biodiversity Decline as a Consequence of an Inappropriate Environmental Risk Assessment of Pesticides, OPINION, viewed on 1/8/22: https://www.frontiersin.org/articles/10.3389/fenvs.2019.00177/full
3. (2022) THE 17 GOALS, United Nations, viewed on 1/8/22: https://sdgs.un.org/goals
4. Furukawa, H. et al (2013) The Chemistry and Applications of Metal-Organic Frameworks, Science, viewed on 15/8/22: https://www.science.org/doi/pdf/10.1126/science.1230444
5. Hamzah, H. et al (2020) Inclusion and release of ant alarm pheromones from metal-organic frameworks, Dalton Transactions, viewed on 19/8/22: https://pubs.rsc.org/en/content/articlehtml/2020/dt/d0dt02047h
6. (2022) Ambrosia Beetles, Bartlett Tree Experts, viewed on 19/8/22: https://www.bartlett.com/resources/insects-and-pests/ambrosia-beetles
7. Worrall, D. et al, (2016) Facile fabrication of metal-organic framework HKUST-1-based rewritable data storage devices, Journal of Material Chemistry C, viewed on 20/8/22: https://www.researchgate.net/publication/307091766_Facile_fabrication_of_metal-organic_framework_HKUST-1-based_rewritable_data_storage_devices
8. Hajizadeh, Z. et al. (2022) Heterogenous Micro and Nanoscale Composites for the Catalysis of Organic Reactions, Micro and Nano Technologies, viewed on 21/8/22: https://doi.org/10.1016/B978-0-12-824527-9.00001-0

APPENDIX

Appendix 1: Scherrer equation

$$\tau = \frac{K\lambda}{\beta \cos\theta}$$

τ = mean size of the ordered crystalline domains

K= 0.94, Scherrer's constant (dimensionless)

λ = X-ray wavelength

β = width peak at half the maximum intensity

θ = the Bragg angle, related to the angle between the incident X-ray and the family of lattice planes [8].

EXPERIMENT DETAILS

The syntheses of the 5 different HKUST-1 copper MOFs involved adding 0.9mmol of the copper salt, acting as the metal node, to 0.6mmol of the organic linker molecule trimesic acid (benzene tricarboxylic acid). After 15ml of DMF (a common solvent for chemical reactions) was added to the 2 solids and the

contents were stirred with a magnetic stirrer bar, they were put in the microwave at 120°C for 30 minutes to allow the MOFs to form. This liquid was then put into the centrifuge to separate the solid MOF from the DMF solvent. This left a blue solid precipitate containing the MOF at the bottom and a suspension above containing solid MOF particles suspended in the DMF solvent. 5 more centrifugations were carried out, each time decanting the excess solvent before filling the centrifuge tube up to 35ml with the DMF solvent. The purpose of these washes was to remove any unreacted material by using a very polar solvent in DMF to dissolve any ions and other polar compounds. After each wash the suspended liquid became clearer, eventually leaving a clear and colourless supernatant (the volume of liquid lying above the precipitate after centrifugation) which indicated that all the MOF was collected at the bottom. It was found that the bigger the particles of the MOF, the quicker a supernatant was reached, since the nitrate MOF required only 3 washes but the smallest particle-sized MOFs in the acetate needed all 5 washed to reach a supernatant. 3 further washes were then carried out, using diethyl ether which, because of its dipole moment, dissolves any remaining DMF. Although this marks the end of the synthesis process, the MOFs are not ready to be loaded with semiochemicals; some of the DMF binds to the MOF when washing, and so must be removed. This process of emptying the pores of the MOFs, called activation, is done by yet more washing, this time with acetonitrile. Although DMF binds more strongly to the MOF than acetonitrile, the excess of acetonitrile is so large that it is enough to remove the DMF from the pores of the MOF. Any leftover acetonitrile, with its low boiling point of 82°C, is removed as the MOFs are put in the

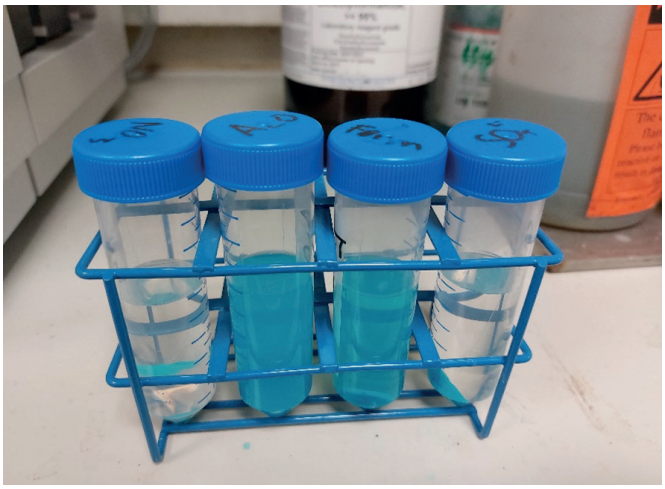


Figure 11: 4 of the HKUST-1 MOFs, during the washing process. The nitrate and sulfate MOFs have already formed a supernatant, but the acetate and sulfate MOFs are still partially suspended in the solvent.

oven at 100°C for several hours. After activation, the loading process took place for just the nitrate and acetate MOFs: the semiochemical was pipetted into several upturned centrifuge lids inside a box. 2 boxes were used, one for each semiochemical, each containing vermiculite which had been heated in the oven to expand its particles so that they were able to absorb any moisture in the box. A petri dish containing each MOF was then placed inside the box and the lid was put on to provide an environment saturated with the semiochemicals. To measure the uptake of semiochemical, a small amount of the MOFs was taken and used for 1H NMR spectroscopy by adding the solid to a DMSO solvent. Finally, once the MOFs had been loaded, the MOFs were ready to begin their release of semiochemicals. This was done by leaving the MOFs on a hot plate set to a constant temperature of 40°C and measuring the amount of semiochemical inside the MOFs by taking some of the sample for ^{1}H NMR.

Figure 12: Set-up for the loading process

Figure13: Set-up for the release process

ACKNOWLEDGMENTS

I am deeply grateful to Professor A. Burrows of the University of Bath for giving me the opportunity to undertake this project as part of his Burrows research group, and for his oversight over the 3 weeks of the project. I would like to extend my sincere thanks to the members of his research group for their encouragement throughout, giving me invaluable insight into the process of scientific research. Special thanks should go to Dr. J. Nicks for mentoring me each day, answering all the questions I could possibly think of whilst also challenging me to think beyond the task at hand.

Exploring Neural Networks

An extract of this short-listed Independent Learning Assignment (ILA)

Shaoyon Thayananthan, Upper Sixth

4 LEARNING THE PARAMETERS OF A NEURAL NETWORK

In this section we construct a simple neural network with one hidden layer (figure 4) and derive the learning algorithm from first principles. I aim to briefly introduce the mathematics used in neural networks, especially the partial derivatives of the error function that are used in updating the weights of the neural network.

4.1 MODEL

We introduce a simple neural network (figure 4) to learn continuous functions with a single variable. The neural network in this case learns values of y_i for given values of x_i. It consists of an input layer with a single input, a hidden layer with a K number of neurons (in the diagram K is set to 3), and an output layer with a single output variable. Each neuron has a weight w_k^h and bias b_k^h attached to it. The result of $w_k^h x_i + b_k^h$ is put through an activation function, g. After the activation of the neurons are weighted and have been summed, an additional bias term b^0 is added.

Though it seems redundant to have multiple neurons observing and interpreting the same inputs, the diagram shows that the presence of multiple neurons is crucial to building a successful neural network. By each neuron being able to learn a different set of weights, different functions of the input

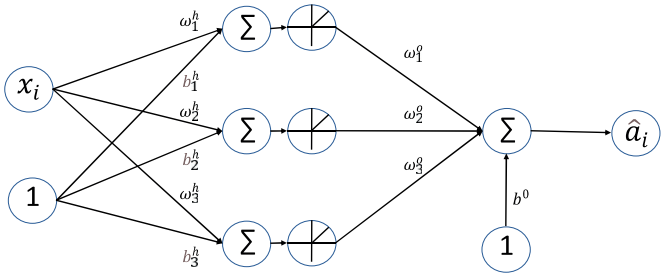


Figure 4: Neural network with one input layer with a single input, one hidden layer with three neurons and an output layer with a single output.

data can be represented. The following equation links the input and and the output of this neural network. The parameters of this neural network like w_1^h w_1^0 and b^0 need to be learnt from the training data iteratively. The predicted value $\hat{a}_i$ is shown to be

$$\hat{a}_i = \left(\sum_{k=1}^{K} w_k^o \, g\left(w_k^h \, x_i + b_k^h \right) \right) + b^o \qquad (1)$$

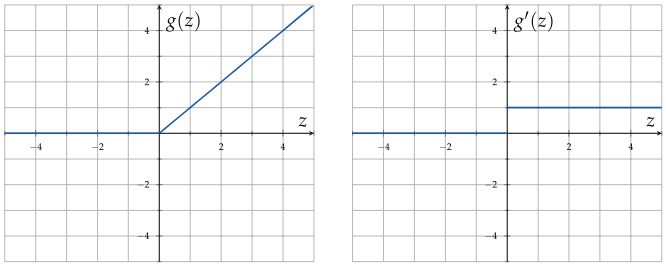
where g is the activation function of the neuron defined as

$$g(z) = \begin{cases} z, & \text{if } z \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

and its differentiation is given by

$$g'(z) = \frac{d}{dz} g(z) = \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

as shown by figure 5. The function g(z) is continuous, but its derivative g'(z) is discontinuous at z=0.



(a) $g(z)$ (b) $g'(z)$

Figure 5: Rectified linear unit function $g(z)$ and its derivative $g'(z)$

4.2 OPTIMISING WITH GRADIENT DESCENT

We need to learn the parameters of this neural network from noisy samples of a one-dimensional function. Let N be the number sample points denoted by (x_1,y_1), (x_2,y_2) ... (x_N,y_N). These points are obtained by sampling from a given one-dimensional function and adding random noise, as seen later in figure 7.

$$y_i = f(x_i) + \epsilon_i \qquad (2)$$

Our aim is to obtain an optimal set of the parameters of the neural network that minimises least square error between the sample y_1 values and the predicted $\hat{a}_i$ values. Let $\mathbf{W} = [w_1^h, b_1^h, w_1^0, \ldots w_K^h, b_K^h, w_K^0, b^0]'$ be the set that contains all the parameters of the neural network. Least square error between the samples and neural network prediction is defined as

$$E(\mathbf{W}) = \frac{1}{2} \sum_{i}^{N} (\hat{a}_i - y_i)^2 \qquad (3)$$

The difference is squared to keep it positive, as an amplitude. We would like to obtain a set of specific values of the parameters which minimises the error. By differentiating the error function, we are able to do so.

$$\mathbf{W}_{\text{opt}} = \min_{\mathbf{W}} E(\mathbf{W}) \qquad (4)$$

This error function is too complex to be optimised analytically due to the many variables involved. However, there are several algorithms which can obtain the optimal parameter set by iteratively traversing the parameter space, and using partial differentiation. Here we use a simple version of a popular algorithm used in neural networks called Gradient Descent. This method assumes that a first derivative of the error function can be calculated with respect to each of the parameter at any value. It exploits the intuition that traversing in the opposite direction of the gradient will take us closer to the value of the parameter where the minimum of the error function is obtained. This is illustrated in figure 6. The parameters are updated as follows:

$$\mathbf{W}_{\text{new}} = \mathbf{W}_{\text{old}} - \eta \left\{ \frac{\partial E}{\partial \mathbf{W}} \right\}_{\text{old}} \qquad (5)$$

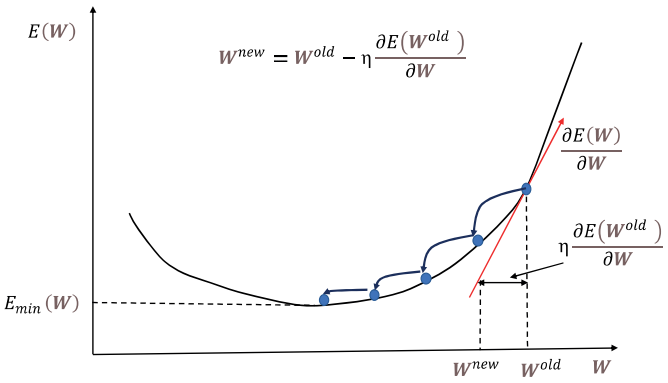


Figure 6: Illustration of gradient descent algorithm. The gradient of the error function is being brought towards zero ie the error function is being minimised.

Here η dictates how much further to travel along the opposite direction of the gradient. Now we can state the update rules for each of the parameters for our neural network as follows. Detailed derivations are provided in the appendix.

$$\begin{aligned}[b^o]_{\text{new}} &= [b^o]_{\text{old}} - \eta \left[\frac{\partial E}{\partial b^o}\right]_{\text{old}} \\ &= [b^o]_{\text{old}} - \eta \left[\sum_{i=1}^{N} (\hat{a}_i - y_i)\right]_{\text{old}}\end{aligned} \tag{6}$$

where $\{\hat{a}_i\}$ values are obtained using the old values of the neural network parameters. Similarly all other parameters are updated as follows:

$$\begin{aligned}[w_j^o]_{\text{new}} &= [w_j^o]_{\text{old}} - \eta \left[\frac{\partial E}{\partial w_j^o}\right]_{\text{old}} \\ &= [w_j^o]_{\text{old}} - \eta \left[\sum_{i=1}^{N} (\hat{a}_i - y_i) g\left(w_j^h x_i + b_j^h\right)\right]_{\text{old}}\end{aligned} \tag{7}$$

$$\begin{aligned}[b_j^h]_{\text{new}} &= [b_j^h]_{\text{old}} - \eta \left[\frac{\partial E}{\partial b_j^h}\right]_{\text{old}} \\ &= [b_j^h]_{\text{old}} - \eta \left[\sum_{i=1}^{N} (\hat{a}_i - y_i) g'\left(w_j^h x_i + b_j^h\right)\right]_{\text{old}}\end{aligned} \tag{8}$$

$$\begin{aligned}[w_j^h]_{\text{new}} &= [w_j^h]_{\text{old}} - \eta \left[\frac{\partial E}{\partial w_j^h}\right]_{\text{old}} \\ &= [w_j^h]_{\text{old}} - \eta \left[\sum_{i=1}^{N} (\hat{a}_i - y_i) g'\left(w_j^h x_i + b_j^h\right) x_i\right]_{\text{old}}\end{aligned} \tag{9}$$

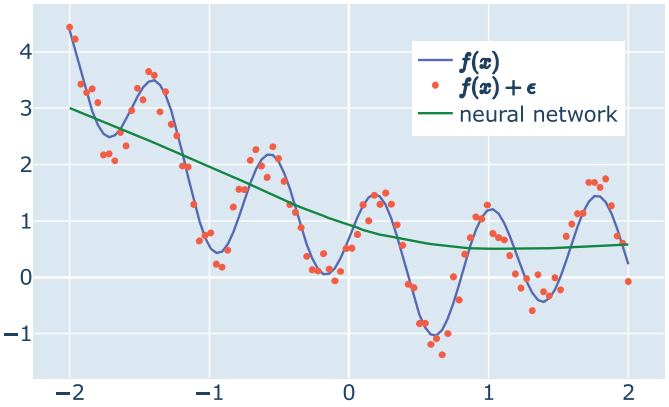
Here, the following two steps are iterated until the error $E(\mathbf{W})$ is sufficiently reduced.

- Calculate the output values of the network $\hat{a}_i$ using current set of weights and biases of the neural network.
- Calculate new values for weights and biases using the error values $(\hat{a}_i - y_i)$ and the update rules defined above.

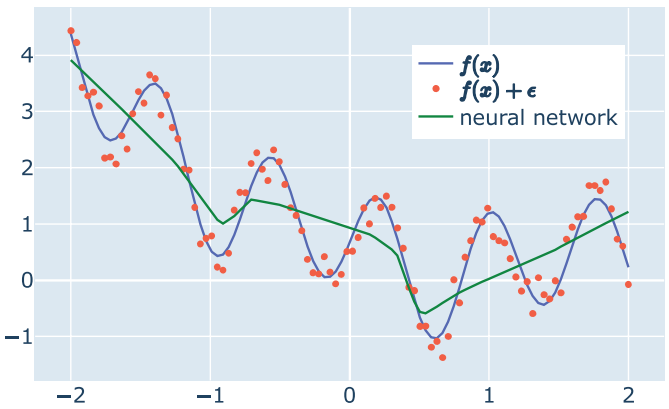
4.3 EXAMPLES

An example of a function being learnt by the derived algorithm above is shown in figure 7 below. The blue line represents the real values of the onedimensional function in a small range, whilst the orange values are the data points (which contain noise) given to the algorithm to learn from. The noise is created by offsetting the y value of a point by a random amount each time, so that the amounts by which points are offset each time form a Gaussian distribution. The green line is the interpretation of the function by the algorithm as it predicts values based off the given points. The green line does not always go through some of the orange points as it is trying to find a function that best fits all of the data, not some of it.

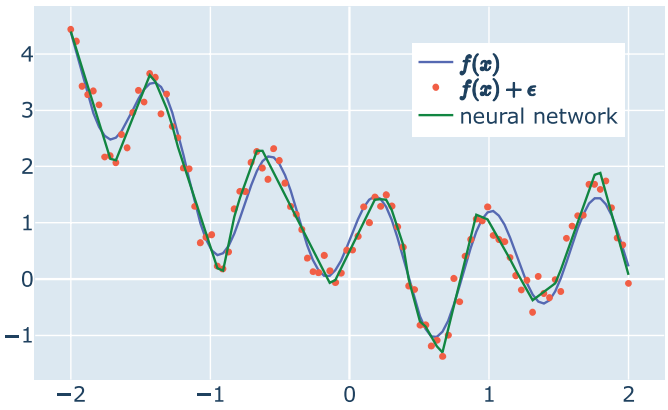
After a single iteration (the weights and other parameters have been updated once) as shown in (a), the algorithm has clearly not learnt the function very well. However, as the number of iterations are increased, the approximated function gets more and more accurate. By the 50th iteration, the predicted shape starts to take on the shape of f(x). At the 1000th iteration, the predicted graph almost matches f(x). Due to there not being infinite data points, it is impossible to predict every value perfectly, but the algorithm is able to produce an excellent approximation, as shown. Moreover, the approximated function can only be used to predict values in the range of the data points (x=-2 to x=2); extrapolation is not possible. By increasing the number of neurons, the algorithm can learn a more accurate function with less iterations, but only through requiring more computing power. This also comes with the danger of overfitting the function. Furthermore, the neurons learn f(x) by approximating it as a linear function between small intervals : the algorithm at no point is able to define a mathematical function for the graph as we have in the caption ($f(x) = 0.2 + 0.4(x-1)^2 + 0.3\sin(4(x-1)) + \cos(8(x-1))$). Also due to this, the turning points are jagged and not smooth due to f(x) also not being smooth at these points.



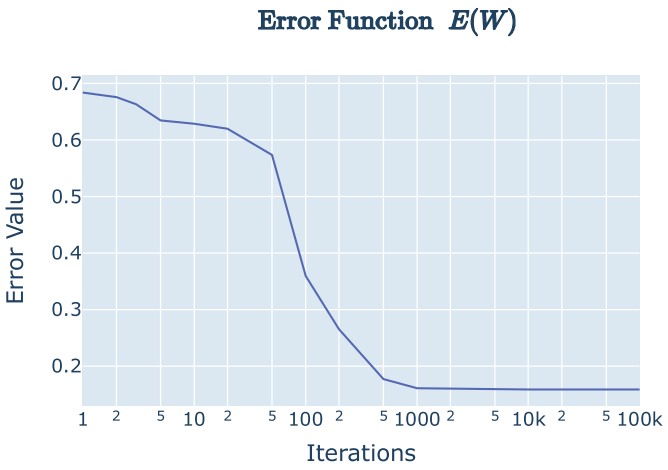
(a) After 1st iteration



(b) After the 50th iteration



(c) After the 1000th iteration



(d) Error function

Figure 7: The figures shows the one dimensional function $f(x) = 0.2 + 0.4(x-1)^2 + 0.3\sin(4(x-1)) + \cos(8(x-1))$. Hundred points are randomly sampled between -2 and 2 and a small amount of noise was added to the samples. A simple neural network described in section 4.1 was then trained with noisy samples and used to predict new values of the function. Figure (d) plots the value of the error function at each iteration, with the x axis being a log axis.

Stating the Obvious

A research report written following an Original Research in Science (ORIS) placement

Thomas Thevenon, Upper Sixth

INTRODUCTION

In this article I give an account of one part of my ORIS project, which was supervised by Professor Hutton at the University of Nottingham. The other two parts of my project involved using the Haskell programming language as a meta-language for the propositional logic [1, pp. 21–27] by using some newer language features [2, pp. 31, 264–266], and using a paper of my supervisor in order to calculate a compiler for a programming language [3], whose syntax I proved unambiguous [4, 5].

I would like to thank Mr Lau and Professor Hutton for giving me the opportunity to do this.

WHAT IS OBVIOUS?

Throughout the history of mathematics, many statements have been considered obvious and undeserving of proof. Example such statements include the fact that zero is not equal to one, that two plus three is equal to five, or that all statements are either true or false.

The view that such statements do not need to be proven rigorously is problematic for multiple reasons. Firstly, by not studying the reasons for which such statements are true, or concretising the definitions of supposedly intuitive concepts, mathematicians miss insight into proof techniques that could otherwise carry over to more complex statements, and fail to spot similarities between concepts. They also lose insight into the philosophy of mathematics itself: the process by which something may be proven true. However, ultimately, what seems obvious is not a all obvious: there have been multiple foundational crises in mathematics, wherein the entirety of mathematics has had to been reviewed when fallacious results were obtained from theories deemed obvious.

Examples of fallacious results coming from unquestioned reasoning are plentiful. Those familiar with integration by parts may want to try and spot the mistake in this proof that 0 = 1:

THEOREM 1. 0 = 1.

Proof. We consider $I = \int \frac{1}{x \ln x}\,\mathrm{d}x$ by integrating by parts, with:

$$u = \frac{1}{\ln x}$$

$$\frac{\mathrm{d}u}{\mathrm{d}x} = -\frac{1}{(\ln x)^2}\cdot\frac{1}{x}$$

$$\frac{\mathrm{d}v}{\mathrm{d}x} = \frac{1}{x}$$

$$v = \ln x$$

So

$$\begin{aligned} I &= \int \frac{1}{x \ln x}\,\mathrm{d}x \\ &= \frac{1}{\ln x}\cdot \ln x - \int \frac{-\ln x}{x(\ln x)^2}\,\mathrm{d}x \\ &= 1 + \int \frac{1}{x \ln x}\,\mathrm{d}x \end{aligned}$$

Subtracting $\int \frac{1}{x \ln x}\,\mathrm{d}x$ from the top and bottom lines of the equality gives 0 = 1.

There must be a mistake here because 0 is not equal to 1. Surprisingly, however, it lies in the very last sentence of the proof: an indefinite integral refers to a family of functions with a certain derivative. All of these functions differ by a constant, and hence, when subtracting an integral from itself, the correct result is a family of constants: not just zero alone.

True statements also often times oppose intuition. For example, even though the numbers are written differently, over the real numbers, 0.9 recurring ($0.\bar{9}$) is equal to 1. An unsatisfactory proof is often presented of this:

THEOREM 2. $0.\bar{9} = 1$.

Proof. Let $x = 0.\bar{9} = 0.9\bar{9}$. Then $10x = 9.\bar{9}$, and, by subtracting 10x and x, $9x = 9.\bar{9} - 0.\bar{9} = 9$. So $x = 1$.

Infinity is the fundamental reason for which people struggle with the idea that $0.\bar{9} = 1$: what does a number with an infinite number of decimal places represent? Infinity is not once mentioned in this proof however: the nature by which operations may be done on an infinite sequence of digits is simply glossed over, when, other times, such reasoning leads to paradox: the infinite sum $S = 1 - 1 + 1 - 1 + 1 - 1 + \cdots$ can be fallaciously shown to be equal to both 0 and 1 by separate rewritings:

$$\begin{aligned} S &= (1-1) + (1-1) + \cdots \\ &= 0 + 0 + \cdots = 0 \end{aligned}$$

$$\begin{aligned} S &= 1 - (1-1) - (1-1) - \cdots \\ &= 1 - 0 - 0 - \cdots = 1 \end{aligned}$$

A more rigorous approach to infinity is needed to give a satisfactory proof that $0.\bar{9} = 1$.

On a deeper philosophical level, even if mathematical concepts are well-defined, it is not immediately obvious whether a chain of reasoning is correct, what it means for a chain of reasoning to be correct, or whether a statement even corresponds to any mathematical meaning. Imagine that you are walking through a forest, and find a box. On the box the words "a million dollars shall lie in this box if and only if a true statement is written on it" are written. You reason that this statement must either be true or false: if the statement is true, a million dollars should lie in the box. However, if the statement were false, and a million dollars did not lie in the box, a contradiction would arise: the statement on the box would end up being true. So you reason that, no matter what, a million dollars must be in the box. You open the box and are surprised to find nothing.

"Of course!" you reason. Why would any words, written by someone you have no trust in, have any bearings on the content of a box? But you're still wondering whether the statement on the box was true or false. Does it need to be true or false? Classical mathematicians happily work with the idea that any proposition is either true or false: the law of excluded middle. The fallacy here is far more subtle: the statement on the box isn't even a proposition and the ideas of true and false cannot be applied to it.

On June 16 1902, Bertrand Russel, a British mathematician, wrote a letter in

German to Gottlob Frege, a German philosopher, questioning him on a similar paradox, which is now called Russel's paradox: does the set of all sets that don't contain themselves contain itself [6, pp. 130–133]? Frege replied that this discovery had "surprised him beyond words", and "left [him] thunderstruck, because it [had] rocked the ground on which [he] meant to build arithmetic".

One such solution to this paradox

is type theory, a field that is being worked on actively today. A type theory is a set of inference rules by which all terms within the language of maths may be assigned types. In such a theory, the term 5 + 3 = 2 would have the proposition type: 5+ 3 = 2 : Prop. Self-referential terms can not be constructed in most type theories, avoiding this paradox altogether.

Pedantic, niche, impractical are words often used to refute the need for rigour in mathematics: after all, the examples provided above are contrived, and excessive rigour hinders other mathematical progress. Most mathematicians aren't writing their proofs in terms of formal inference rules. But, with proofs becoming increasingly abstract, mathematicians need to be aware of what makes logic valid. Maths also doesn't just need to be done merely because it is useful: "Real mathematics must be justified as art if it can be justified at all".

Unfortunately, as is often the case with mathematics, the development of type theory as a result of this indomitable rigour has been immensely useful. Type theory is nowadays used to check computer programs for correctness: to prove that they don't crash. Statically typed programming languages, such as Java, C and Haskell, were all implemented using ideas from type theory. The type theories these languages use are of various strength: the most powerful type theories will be able to prove that the program itself has the correct properties. Avoiding bugs in computer programs is imperative, as they can lead to extreme monetary threat and even loss of life:

1. Heartbleed, a bug in a cryptography library, lead to the mass revocation of certificates used for encrypting online communications. Cloudflare estimated that these changes cost one certificate authority US $1 million a month, just in bandwidth [7]. Only a few lines of code had to be changed to fix the bug [8].
2. Knight Capital Group lost US $440 million in 45 minutes owing to a bug in their trading software [9, pp. 1–2].
3. The Ariane flight V88 launch failure, which was the result of bugs present in rocket software, lead to a loss of approximately US $370 million [10].
4. On February 25, 1991, timing issues within the software of an air defence system lead to the otherwise preventable death of 28 soldiers and injuries of numerous others [11].
5. The estimated cost of software defects in the US is $59 billion a year [12].

A NEW LANGUAGE

In the subsequent sections of this article the implementation details of a programming language based on a type theory used in mathematical proof is briefly described. By using such a type theory, the programming language becomes capable of verifying proofs of mathematical statements. The programming language is named after a genus of bird studied by a famous logician: Calidris [13].

Programming languages themselves have to be programmed. Sometimes programming languages are even programmed in their own language, through a process called bootstrapping, where an older version of the programming language is used to create the newer version of the programming language. Creating a programming language is a complex process.

Firstly, the programming language must be given a syntax: a set of rules used to create human-readable sentences in the language, which may then be parsed by a computer into a structure that it may process. For example, given the following syntax:

<sentence> → I like *<list>*
<list> → *<item> <list1>* | *<item>*

<list1> →, <item> <list1> | and <item>

<item> → apples | cheese | spoons

A computer could then parse sentences such as "I like apples", "I like apples and spoons", "I like cheese, cheese and cheese" into lists of items. But it would not be able to parse "I like cheese apples and spoons" or "I like, apples" because these do not satisfy the grammar. For the programming language I created, the grammar was slightly more complicated. The types that exist within the language are provided in Figure 1.

```
Prop; Set; Type(i); var; f a

forall a : T, b;

fun a : T, b;
```

Figure 1: Types of expressions that exist within *Calidris*.

The colon is used to annotate the type of an expression. `Prop` is the type of propositions, which are things which may be proven, and `Set` is the type of small types, such as numbers [14]. `Type(0)` is the type of `Set` and `Prop`, and `Type(1)` is the type of `Type(0)` and so on; this is done such that all terms may be assigned a type, and to avoid Girard's paradox, which is the typetheoretic equivalent of Russel's paradox [15, p. 8]. The term `f a` represents function application: `square 4` is the application of 4 to a function called `square`, and could be equal to 16. Functions have type `forall a : T, b`. For example, `square : forall a : number, number = fun a : number => mult a a`.

These expressions (which may be written across multiple modules of the program) are parsed using the megaparsec libary [16]: this uses a common method of parsing nowadays, called parser combinators. My supervisor's early work influenced this field [17].

Afterwards, the program is finally processed. The names of variables bound within expressions (such as the `a` in `fun a : number => mult a a`) are replaced with numbers representing the order in which they appear [18, p. 1], and the proofs present are then processed such that all proofs a proof depends on have been processed before that proof. Various errors will be shown to the user if any of the steps are not possible: some of these errors are shown in Figure 2. The final error shown is automatically generated by the parsing library.

```
  |
1 | o = k; k = o;
  | ^
could not form type for name owing to cyclic dependencies

  |
2 | a = fun a : Set => a;
  |          ^
the name a can not be redefined

  |
3 | k = a = b;
  |       ^
unexpected '='
```

Figure 2: Example errors which may be shown to the user in *Calidris*.

It is then checked that the expressions are valid within the language according to the inference rules. The properties of confluence and strong normalisation of this language can be used to deduce that a certain expression may not be formed [14]. An example inference rule, implemented from those of the Calculus of Constructions [14], is presented below (with the assumptions split over two lines):

$$\frac{E(\Gamma) \vdash T : s \quad s \in S \quad E[\Gamma ::(x : T)] \vdash U : Prop}{E(\Gamma) \vdash \forall x : T, U : \text{Prop}}$$

This rule states that `forall x : T, U` is a well-formed proposition if `T` is a well-formed expression of type sort (`Set`, `Prop`, or `Type(i)`) and `U` (given that `x` is of type `T`) is a well-formed proposition. Using these rules, the programming language can determine whether all the expressions in the language are well-formed, and hence verify all the proofs present.

BEHOLD AND LO

With the language now implemented, basic proofs can be written. I will prove that the proposition $A \wedge B$ implies $B \wedge A$. For example, if it is raining and it is cold, it is also cold and raining. Under the Curry-Howard isomorphism [19], a proof of $A \wedge B$ may be seen as the construction of a tuple type: a type containing two objects, with one object proving A and the other object proving B. The tuple type is defined using the Church encoding [20, Chapter 6.2] in Figure 3:

```
imp
  =   fun a : Prop
  =>  fun b : Prop
  =>  forall x : a, b;
and
  =   fun a : Prop
  =>  fun b : Prop
  =>  forall c : Prop
  ,   imp (imp a (imp b c)) c;
```

Figure 3: The definition of logical conjunction in *Calidris*.

In this code, implication is first defined: the function imp in the code takes in two propositions (things which may be proven), and creates a new type. A proof that one proposition implies another proposition will be a member of this type. This type is the type of a function which, given a proof of a statement, produces a proof of the implication. For example, the proof that a proposition implies itself is written in Figure 4.

```
simple_proof
  :   forall a : Prop, imp a a
  =   fun a : Prop
  =>  fun a_proof : a
  =>  a_proof;
```

Figure 4: A proof that a proposition implies itself.

This function takes in a proposition and then returns something of type `imp a a`: a function that re-turns a proof of a given a proof of a.

The and function defined in Figure 3 behaves similarly. A proof of two propositions can be defined in terms of implications: if a and b are both true, one should be able to write a function which takes in an expression of the form $a \Rightarrow (b \Rightarrow c)$ (where implies is written $\Rightarrow$) and return a proof of c.

```
final_proof
  :   forall a : Prop
  ,   forall b : Prop
  ,   imp (and a b) (and b a)
  =   fun a : Prop
  =>  fun b : Prop
  =>  fun a_and_b : and a b
  =>  fun c : Prop
  =>  fun gen : imp b (imp a c)
  =>  a_and_b c (
      fun a_proof : a
  =>  fun b_proof : b
  =>  gen b_proof a_proof
  );
```

Figure 5: A proof that $(a \wedge b) \Rightarrow (b \wedge a)$.

Using these definitions, we can now write a proof that $(a \wedge b) \Rightarrow (b \wedge a)$, shown in Figure 5, by writing a function that accepts a proof that $a \wedge b$ and uses the definition of and to convert this to a proof $b \wedge a$. The essence of it is that the proof that $a \wedge b$ is used to apply b and then a to an implication of the type $b \Rightarrow (a \Rightarrow c)$, hence proving $b \wedge a$.

REFERENCES

[1] Simon Thompson. *Type Theory & Functional Programming*. University of Kent, 1999.

[2] GHC Team. *GHC User's Guide Documentation: Release 9.2.1. 2021*. url: https://downloads.haskell.org/ghc/9.2.1/docs/users_guide.pdf.

[3] Patrick Bahr and Graham Hutton. "Calculating correct compilers". In: *Journal of Functional Programming 25 (2015)*, e14. doi: 10.1017/S0956796815000180.

[4] Thorsten Altenkirch. *What is a LL(1) grammar*. Accessed: 2022-09-03. University of Nottingham. 2001. url: https://www.cs.nott.ac.uk/~psztxa/g51mal.2001/notes/node36.html.

[5] Thorsten Altenkirch. *How to calculate First and Follow*. Accessed: 2022-09-04. University of Nottingham. 2001. url: http://www.cs.nott.ac.uk/~psztxa/g51mal.2001/notes/node37.html.

[6] Gottlob Frege. *Philosophical and Mathematical Correspondence*. Translated by Hans Kaal. Basil Blackwell, Oxford, 1980. isbn: 9780631196204.

[7] Matthew Price. The Hidden Costs of Heartbleed. Accessed: 2023-02-04. 2014. url: https://blog.cloudflare.com/the-hard-costs-of-heartbleed/.

[8] Steve Henson. Add heartbeat extension bounds check. Accessed: 2023-02-04. 2014. url: https ://github.com/openssl/openssl/commit/96db9023b881d7cd9f379b0c154650d6c108e9a3.

[9] In the Matter of Knight Capital Americas LLC (Release No. 70694). Accessed: 2023-02-04. Securities and Exchange Commission. 2013. url: https://www.sec.gov/litigation/admin/2013/34-70694.pdf.

[10] Mark Dowson. "The Ariane 5 Software Failure". In: SIGSOFT Softw. Eng. Notes 22.2 (Mar. 1997), p. 84. issn: 0163- 5948. doi: 10.1145/251880.251992. url: https://doi.org/10.1145/251880.251992.

[11] PATRIOT MISSILE DEFENSE: Software Problem Led to System Failure at Dhahran, Saudi Arabia. Tech. rep. GAO/IMTEC-92-26. United States General Accounting Office, 1992.

[12] RTI. The Economic Impacts of Inadequate Infrastructure for Software Testing. Tech. rep. 7007.011. "Prepared for Gregory Tassey". National Institute of Standards & Technology, May 2002, ES.11.

[13] Per Erik Rutger Martin-Löf. "Mortality rate calculations on ringed birds with special reference to the Dunlin Calidris alpina". In: Arkiv för Zoologi. Serie 2 21 (1961).

[14] Calculus of Inductive Constructions. INRIA. 2018. url: https://coq.github.io/doc/v8.9/refman/language/cic.html.

[15] Jean-Yves Girard. "Interprétation fonctionnelle et élimination des coupures de l'arithmétique d'ordre supérieur". Thèse d'État. Paris VII, 1972.

[16] Mark Karpov. megaparsec: Monadic parser combinators. Accessed: 2022-09-05. 2022. url: https ://hackage.haskell.org/package/megaparsec-9.2.1.

[17] Graham Hutton and Erik Meijer. Monadic Parser Combinators. Tech. rep. NOTTCS-TR-96-4. Department of Computer Science, University of Nottingham, 1996.

[18] Thierry Coquand and Gérard Huet. "The calculus of constructions". In: Information and Computation 76.2 (1988), pp. 95–120. issn: 0890-5401. doi: https://doi.org/10.1016/0890-5401(88)90005-3. url: https://core.ac.uk/download/pdf/82038778.pdf.

[19] William Alvin Howard. "The Formulae-as-Types Notion of Construction". In: To H. B. Curry: Essays on Combinatory Logic, Lambda Calculus, and Formalism. Ed. by J.R. Hindley J.P Seldin. Academic Press, 1980.

[20] Henk Barendregt. The Lambda Calculus. Its Syntax and Semantics. Elsevier Science Publishers, 1984. isbn: 0444867481.

How is inter-generational communication impacted by the difference in the interpretation of emojis?

A short-listed Independent Learning Assignment (ILA)

Aaron Luke Venter, Upper Sixth

INTRODUCTION

Emojis are defined as "any of various small images, symbols, or icons used [...] to express the emotional attitude of the writer, convey information succinctly, communicate a message playfully without using words, etc." ("Emoji", n.d.). Emojis are a form of writing system known as pictograms (From Hieroglyphics to Emojis: Evolution of Logographic Languages, 2021), a separate type of writing system compared to the Roman alphabet this paper is written in. Emojis were first invented in Japan in 1999 by Shigetaka Kurita, their name coming from the portmanteau of three Japanese words: "e" (picture) + "mo" (write) + "ji" (character). Kurita tasked himself with condensing 176 unique ideas into 12-bit symbols within the span of only 5 weeks (Shigetaka Kurita: The man who invented emoji, 2018). The next biggest leap that really brought emoji to the mainstream was their addition into Unicode, the universal character standard system, in 2010. From there, they were subsequently introduced to iOS devices in Japan and then spread worldwide.

RESEARCH

LITERATURE REVIEW

INTERPRETATION OF EMOJIS

According to an article by the Wall Street Journal in 2021 (Ishmael, 2021), Generation Z (Gen. Z) assigns meaning to emojis that are different from the "tradition" that the older generations tend to use, resulting in "a lot of confusing interactions". The article details how in situations like the workplace and even within a family, communication is affected or might require a lot more thought than normal for the younger generations, as they have to accommodate the older, more "standard" interpretations of certain emojis. This shows the effect that the apparently novel discrepancy in emoji interpretation across generations has on the online communication between these generations. One example is given when a new intern joined a media company, the welcome email appeared disingenuous and cold before the reader took a second to remember the age of the sender. Especially in these kinds of professional situations, it is evident that different interpretations of emojis may seriously affect communication.

CONTEXT OF EMOJIS

A researcher at Halmstad University conducted a study on Swedish senior school students from 16-19 years old to determine the importance of context regarding the interpretation of emojis and emoticons (Kelly, 2015). The results of the study revealed that, for the person receiving a message – the "interpreter" – to understand what the meaning of the emoji is, it is important that a "textual context is established". The study also demonstrated that lone emojis often don't have one singular meaning by themselves, and they require proper context to determine which one of their commonly associated meanings applies in the ongoing conversation. Their meanings can often vary depending on the situation being discussed, the mood of the sender, the intended recipient etc.

These findings are extremely important to consider, as they demonstrate that emojis often don't have one, single, definitive meaning, and that they are subject to the context of a conversation. This means that certain emojis in certain situations may cause more confusion than others, as they may have multiple, divergent meanings.

PURPOSE OF EMOJIS

A research article by Cramer et al. on the functions of emojis in US messaging investigated the primary purposes for the use of emojis in digital communication (Cramer, de Juan, & Tetreault, 2016). They returned the ideas that emojis are used to provide additional emotional or situational context to a conversation; they are used to adjust the tone of a message, make it more engaging; they can be used to manage the flow of the conversation or be used to maintain the personal relationship between the sender and the recipient. This establishes that emojis are not just used for one single purpose and that since their invention, they have filled a number of important roles in the medium of text communication, facilitating important aspects of conversation that were previously hard to translate from audible communication to instant text.

ADOPTION OF EMOJIS

A paper by Hamza Alshenqeeti discusses the important, often overlooked in generalisation, background to how most people experience emoji usage. The paper details that "age is less of an indicator of usage than technological awareness and capability" (Alshenqeeti, 2016). This is extremely important to realise. In the context of the differences in interpretation between different generations, some may argue that older generations are simply less acquainted with emojis, but this would be wrong to assume as a definite. Older generations might actually have similar levels of acquaintance with emojis, but just use them in different ways.

MEANING OF EMOJIS

Scheffler et al. discovered that when emojis are read by the brain in digital text, they are processed the same way that pictures are processed, not words. Overall sentences are also not necessarily affected in meaning when certain emojis replace words. Emojis can also trigger entire lexical entries, allowing for homophonous nouns to be represented by their corresponding emojis but still be interpreted entirely correctly with the relevant context (Scheffler, Brandt, de la

Fuente, & Nenchev, 2022). This is important for the later discussion in this paper on the proposition surrounding whether emojis are allowing western humans to return to more logographic systems of writing.

FIELD WORK

PROCESS OF CREATING THE FORM

In the following investigation, the questionnaire used to get the results was formulated from a combination of research from the literature review, along with data collected from Reddit and interviews with various people of different age groups. The specific emojis were curated from multiple Reddit posts in different communities and data collected from questioning individuals on what their smartphones' keyboards stated were their most frequently used emojis.

The data was collected and formulated into a Microsoft Form, populated with three main sections. The first brief section was one that simply asked for the name and email address of the respondent in order to potentially follow up on any interesting answers. The respondents were also asked to identify which age generation they belonged to, from a choice of: "Baby Boomers (1946-1964)", "Generation X (1965-1980)", "Millennials (1981-1996)", "Generation Z (1997-2012)", "Generation Alpha (2013-Present)".

The second section was a 4-point Likert scale consisting of 19 emojis queried about the frequency of their use for the respondent. The options available to select were: "Never", "Rarely", "Sometimes", "Often".

The last section of the questionnaire was broken down into a question regarding each individual emoji about the meanings that the respondent associated with each emoji. The options given for the meanings the respondent could choose were composed from mainly from the data collected from Reddit and the interviews with people from different age groups.

Below each question about the meanings associated with each particular emoji, there was an empty box left as an option for people to fill in any other meanings they thought were not covered by the initial options available.

HYPOTHESES

1) There are generational differences in the interpretation of emojis
2) Older generations perceive emojis more literally
3) The above assumptions account for a difficulty in inter-generational communication

OBJECTIVE

To test the above hypotheses through the use of an online form.

METHODOLOGY

The Microsoft Form mentioned above was sent out to friends, family and teachers. The same form was also posted onto multiple social media accounts to spread awareness. In order to get a range of responses, a form of snowball sampling was used where participants were asked to identify other participants.

These responses were cross-checked with a market research study which listed the top five most popular emojis per generation in order to confirm they aligned with previously collected data from other sources.

At the time of processing the data, a total of 108 responses had been received. For Generation Alpha, only one response was received, and this was omitted as it would have been statistically invalid.

EMOJIS INCLUDED IN THE STUDY

RESULTS / FINDINGS

Below are a set of three graphs and individual analysis on the data that they present. They analyse the top emojis used per generation, the frequency at which each generation uses emojis as a part of their communication in general, and the popularity of various meanings associated with a selection of emojis from the study.

Overall, the study appears to confirm the initial findings in the literature review and the hypotheses. It indicates that there are obvious intergenerational differences in the interpretation of emojis, particularly when comparing Gen. Z to the other three generations when looking at fig. 3. Older generations do appear to perceive emojis more literally, as seen by the far more consistently literal interpretations of the emojis seen in fig. 3. A good example of this is the "skull" emoji. Additionally, it can be seen that emoji meaning can vary based on textual context, as seen by the multiple meanings associated with emojis within generations.

Other notably interesting findings not necessarily directly relevant to the hypotheses were that the range of emojis used by Baby Boomers and Gen. X was identical, just in a slightly different hierarchy. Also, Gen. Z used a much greater variety of emojis than the other three generations, coupled with a very different set of meanings associated with those emojis.

MOST POPULAR EMOJIS

Figure 1 - The top 7 emojis per generation

This table demonstrates several interesting results. One result of the data collected was that the Baby Boomer generation and Gen. X share identical sets of emojis, with the only difference being the order of their popularity. Another is that across all generations, only three emojis were of common ground. These were the "Heart", "Laughing-Crying", and "Thumbs-up" emojis.

Gen. Z had the most dramatic shift with the order of popularity of their emojis, with an emoji that was either low or absent on the rankings of the other three generations displacing the previously most popular emoji.

USE OF EMOJIS

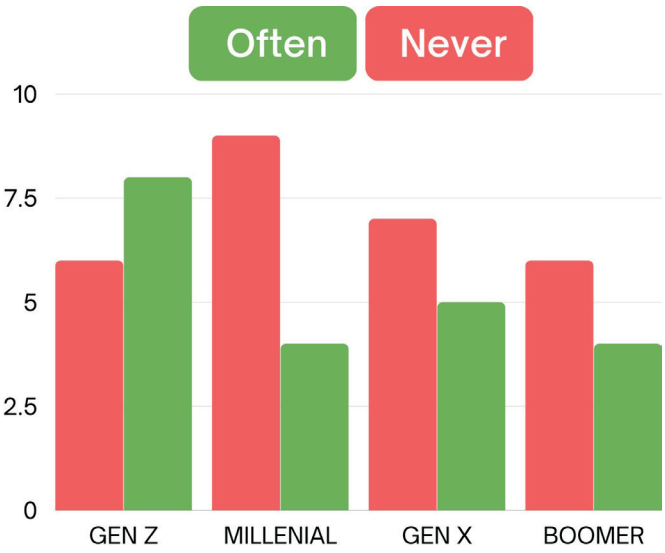


Figure 2 - Count of usage of emojis "often" and "never" across generations

In fig. 2, the data is less conclusive. It shows that Gen. Z uses nearly twice as many emojis "often" on average compared to other generations. However, there was no significant difference between the other three generations.

This section of the field study would greatly benefit from a larger sample size in order to draw more statistically significant data.

MOST COMMON MEANINGS ASSOCIATED WITH EMOJIS

	Boomer	Gen. X	Millennial	Gen. Z
❤️	Love	Love	Love Sarcasm / P-A	Love Sarcasm / P-A
💀	Death	Death Laughter Shock	Death Laughter Embarassment	Laughter Embarassment Shock
🤠	Cowboys Confidence Joy	Adventure Cowboys Joy	Cowboys I don't know / other Adventure	Sarcasm Cowboys I don't know / other
😀	Joy / Happiness Cheerfulness	Joy / Happiness Cheerfulness Dead inside	Joy / Happiness Cheerfulness Dead inside	Dead inside Sarcasm Joy / Happiness
😩	Frustration Sadness	Frustration Sadness	Frustration Sadness Desire / Lust	Ironic desire Desire / Lust Frustration
🥰	Love Happiness Embarrassed Affection	Love Happiness Embarrassed Affection	Love Happiness Embarrassed Affection	Love Happiness Passive-aggressiveness

Figure 3 - The top three meanings associated with seven emojis

The table above represents a selection of emojis that best demonstrated the trend seen in the data as a whole. It demonstrates that there are generational differences in the interpretation of and the meanings associated with emojis.

The most drastic difference from the other three generations was Gen. Z, where meanings are entirely separate from the literal, or more sarcastic associations were made with the emojis. In comparison with Gen. Z, and to a lesser extent the Millennial generation, the two older generations made much more literal associations with the given emoji based on what it actually displayed, this being particularly evident in the "Cowboy" and "Skull" emojis.

FUTURE RESEARCH

In order to confirm whether there are in fact no differences between the Baby Boomers, Gen. X, and Millennials regarding their interpretation of emojis, a larger sample size would be required. This is because, in this study, the majority of the respondents were from Gen. Z.

Ideally, the list of emojis wouldn't be defined before the respondents took part; it would be better if the respondents were able to freely identify the emojis they used. This could then be compiled more easily into a list of common emojis that is more realistic and representative of a wide range of uses.

The results of this study bring into question what the reason for the difference in interpretation and use of emojis by Gen. Z is. Did Gen. Z start off using these emojis with these subversive connotations? Have they developed this "language" of emojis over time through regular use and contact over the internet? With this fascinating use of emojis by Gen. Z, and Scheffler et al.'s findings on the processing of emojis stating that sentence meaning is not in fact impacted when certain words are replaced by emojis (Scheffler, Brandt, de la Fuente, & Nenchev, 2022), are Gen. Z bringing back elements of logographic writing systems to western alphabets?

DOES IT IMPACT INTER-GENERATIONAL COMMUNICATION?

ANALYSIS

The findings in the above field study are significant as they demonstrate a tangible difference in the use and interpretation of emojis, particularly in regard to the meanings associated with certain emojis by the different generational groups. Fig.3 in particular demonstrates an objective discrepancy in interpretations, showing the sometimes polar-opposite connotations.

Fig. 1 demonstrates a common baseline set of three emojis identified across the respondents. This could potentially show that there may be a few certain emojis that make up some type of baseline across human interpretation. A further interesting way of analysing this assumption would be to do a further study on the use of emojis across different cultures, to see if it is a generic human baseline or one that is based upon cultural similarities.

Fig. 3 represents a visible difference in interpretation. This difference, particularly in regard to Gen. Z in comparison to the other three generations, would account for the results of the literature review that stated that there were impacts on inter-generational communication due to these differences (Ishmael, 2021).

CONCLUSION

From the data collected, it can be concluded that the difference in interpretation of emojis in written communication leads to a significant enough inter-generational impact to make it relatively difficult in certain situations to convey the true meaning behind messages. The data overall supports the idea that Gen. Z in particular has its own way of interpreting emojis and therefore conveying their thoughts and ideas over the medium of online text communication in general. This unique interpretation in comparison to the other generations can be assumed to be the main chafing point for difficulty in communicating across different generations.

REFERENCES

"Emoji". (n.d.). Retrieved May 5, 2022, from Merriam-Webster Dictionary: https://www.merriam-webster.com/dictionary/emoji

Alshenqeeti, H. (2016). Are Emojis Creating a New or Old Visual Language for New Generations? A Socio-semiotic Study. 7(6).

Cramer, H., de Juan, P., & Tetreault, J. (2016, Sept 6). Sender-intended functions of emojis in US messaging. Retrieved May 5, 2022, from ACM Digital Library: https://dl.acm.org/doi/abs/10.1145/2935334.2935370

From Hieroglyphics to Emojis: Evolution of Logographic Languages. (2021, May 11). Retrieved May 5, 2022, from Akorbi: https://akorbi.com/from-hieroglyphics-to-emojis-evolution-of-logographic-languages/

Ishmael, A. (2021, Aug 9). Sending Smiley Emojis? They Now Mean Different Things to Different People. Retrieved May 5, 2022, from The Wall Street Journal: https://www.wsj.com/amp/articles/sending-a-smiley-face-make-sure-you-know-what-youre-saying-11628522840

Kelly, C. (2015). Do you know what I mean > :(: A linguistic study of the understanding ofemoticons and emojis in text messages. Retrieved from Digitala Vetenskapliga Arkivet.

Scheffler, T., Brandt, L., de la Fuente, M., & Nenchev, I. (2022). The processing of emoji-word substitutions: A self-paced-reading study. 127.

Shigetaka Kurita: The man who invented emoji. (2018, May 23). Retrieved May 5, 2022, from CNN: https://edition.cnn.com/style/article/emoji-shigetaka-kurita-standards-manual/

Cancel Culture and its Effects on Human Growth

An extract from this short-listed Independent Learning Assignment (ILA)

Ashwin Vishwanath, Upper Sixth

PART I: INTRODUCTION

In recent times, there has been a rise in debate over the polarising use of 'cancel culture'. Some argue its usefulness in eliminating toxic, homophobic, or racist thoughts and ideas from picking up traction in the community and letting marginalised or minority demographics in society have a voice. There are arguments to be made that suggest cancel culture can help change communities for the better, creating a more open minded, accepting, and safe space for people to reside in. Others argue that this orthodox school of thought has silenced many contrarian views, muted conservative ideologies, and constricted academic freedom of thought. This movement in society, it has been argued, has all but eliminated the possibility of holistic debate and acceptance in the wider community.

But does the literature support this? Can we even use today's literature to answer these questions? If this is the case, then what are the impacts of cancel culture on the future of human growth? In this paper I aim to answer these questions and address the possible pitfalls society may fall into if continuing this path. Here, I confirm the idea that cancel culture does restrict freedom of thought, both in academia and in general discourse, although there are some nuances to consider.

PART 2: ORIGINS

The phrase 'Cancel Culture' has been used increasingly by the public on social media apps such as Facebook and Twitter. However, the general public rarely stop to understand what the phrase means and where it came from. In some of the literature, cancel culture is broadly regarded as: "the ostracism from media of people based on their past and present views, opinions and actions" (1). This process is remarkably similar to consumer boycotts of certain brands and firms that are perceived to be unethical or thought to have views or policies that go against social norms. In recent years, due to the rise and accessibility of social media, this battleground has moved from the storefront and headquarters of big corporations to platforms such as Twitter and Facebook. The use of the word 'ostracism' is indicative of a much older social construct, first introduced into Ancient Athenian society by Cleisthenes in his reform of the Athenian constitution, as documented by Aristotle in his Constitution of Athens. Ostracism, put into effect around 487 BC (2), ensured the stability of society by removing potentially dangerous people from the society before any damage was caused. Similarly, in recent years people have taken to different mediums, such as social media to do so, mainly because of social media's accessibility to millions of people around the world. This herd behaviour, at its most basic level, prevents radical ideas from gaining traction in society, possibly preventing riots or general unrest from disrupting people's way of life.

Like all social constructs, cancel culture has evolved over time. From exile in 500 BC, to the house arrest of Galileo in 1633 (3), to the cancelling of celebrities such as J.K Rowling, and David Chappelle via social media, only one thing remained the same. The public shunned them and stopped them from expressing their views. The question is, when does the use of cancel culture stop being a punishment and start being an infringement on fundamental human rights and a drag on the growth of society?

PART 3:

Public shaming may be a valid weapon in the fight for social justice by the people who are unable to use the legal system to get an apology or justice. Examples range from the #MeToo movement against extremely influential sexual predators who have committed crimes such as paedophilia and rape, to the #BLM movement, using public shaming and boycotting to bring attention to racist school syllabuses and textbooks, cases of fatal police violence, and what some would see as a suspicious lack of diversity in places of higher education, such as universities and research organizations. As aforementioned, society is constantly evolving, changing its views on topics such as race and gender at an incredible pace. Therefore, it is reasonable to argue that influencers and people in the public eye should adapt to the socio-political landscape in the media, in order to avoid unnecessary offence due to what they have or have not done. From this point of view, cancel culture is a valid tool as it enables the public to express their discomfort or disagreement with the use of racist, homophobic, or sexist language. It lets them see the abuse of power that may not have been visible before, and it enables criticisms of cultural appropriation in society.

In academia however, when opinionated pieces that are published by universities or scholarly sites are being retracted from the public in fear of being misconstrued as being racist, homophobic, or sexist, there is an argument to be made that says that this is an extremely dangerous trend. Removal of contrarian views and arguments, they argue, implies that opinions must be moderated in order to meet certain standards put by the publisher to fulfil other criteria. This may, in turn, filter out what can and cannot be said by academics, depending on what is shown through the Overton Window at the time, for example, Professor Selina Todd's invitation to a conference celebrating women was withdrawn due to her views on transgender rights (4). Although this area of discussion is heavily opinionated, if similar behaviour is followed in other fields of study, it may impact the scientific method: the process of objectively establishing facts through testing and experimentation. It depends on finding pieces of evidence to disprove the current prevailing theory (5), as now opposite views are stifled, meaning that the evidence or ideas that may prove the current theory wrong are never heard. Thus, prohibiting any progress from being made as everyone is thinking the same thing, and as Benjamin Franklin said, "If we all think alike, no one is thinking" (6).

The tolerance of a non-conforming idea and freedom of speech, especially when that expression is extremely controversial or contrarian, is one of the best ways to acknowledge our past values and beliefs as in the process of disagreeing and arguing, one's subconscious points of view and political leaning (Right or Left Wing) come to the forefront of one's minds. When these inner values and beliefs are challenged, we are given new pieces of information that we may not have known before, and with that knowledge, people could

change for the better. Conservatives argue that on worldwide platforms such as social media, there has been a rise in the silencing of contrarian views and ideas that go against the 'socially acceptable' at the time. Anyone who believes otherwise is outcast and looked down upon. This encourages an 'us-them' type segregation, intolerance, and censorship, both self-conducted and external. This inevitably leads to a kind of 'mob mentality' that constantly judges anything that is said or published. This segregation also leads to people who are 'cancelled' by mainstream social media to form their own communities, leading to echo chambers and arguably less regulation and safeguarding against radicalism. Lukianoff and Haidt, in their book The Coddling of The American Mind, described this as a "sanctimonious coddling of student minds," where cognitive demanding work, research and questioning is likened to real world harm or damage.

This imbalance in academia has impacted the general society today. There has been a stark rise in the stigma around 'antivax' people. Freedom of expression and conscience is a basic human right, and people, by law, are allowed to believe what they want, without fear of being discriminated against or treated differently. However, data suggests a growing stigma around antivax people. In a paper published in December of 2021 (13), one person by the name of Jay said:

"I lost contact with family members because of [it] .... I got harassed by a bunch of cousins .... and I found out that a whole bunch of them had started a group, a secret group, and were badmouthing me behind my back."

This sums up most of the experiences detailed in the report, ranging from fear of going to hospitals, to adapting behaviours depending on who they are with to stay safe. One can make an argument that these people's basic human rights are being infringed upon. The fear of going out and being themselves implies that they do not feel safe in society. Surely, after more than a century of democracy in some first world countries such as the USA, UK, and Australia (in which the study was taken), society must have improved to the extent that people feel safe and able to express their ideologies in a way that keeps them safe? This study suggests the opposite is true. In the case of 'antivaxxers', with information so abundant, accessible, and with the school curriculum telling children to trust science, the general public, using the 'mob mentality', have turned on people who believe anything contrary to main-stream academia and belief, effectively segregating them and discriminating against them.

On the other hand, there are some counter arguments to this thought. People who are part of the 'antivax' community opt their children and/or themselves out of state-run vaccine programs largely due to the perceived risk factor and growing complacency in society. Since vaccines have all but stopped most major diseases from spreading quickly in first world countries, citizens are increasingly relying on herd immunity to keep them safe from disease. However, since these people aren't getting vaccines, the people who are not immune grow larger in number, thus de-stabilising the immunity of the herd, putting everyone in the community at risk. This begs the question of where do we (either the governing body or society in general) draw the line between expressing beliefs and enforcing rules that silence one demographic for the safety of all others?

Freedom of expression and belief is a qualitative right. This means that the right can be taken away if it infringes on other basic human rights on others according to Mills harm principle (14). Governments have a duty to stop hate speech, protect national and territorial interests, prevent disorder or crime, protect health or morals and reputation or rights of others, prevent the disclosure of information received in confidence and maintain the authority and impartiality of the judiciary (15)(16). If radical and harmful expression is allowed, the stability of society may be threatened. This is the worst-case scenario, but everyday life is not typically worst case. In the case of the 'antivax' community, this is more of a grey area. On the one hand, as explained previously, they should be allowed to express what they want to, but, on the other hand, problems arise as their choices impact society in general. To add further complexity to the problem, this form of expression technically does not directly harm national interests, or encourage disorder or crime, hence, governments cannot directly intervene and change legislation to force people to take the vaccine just to stop this movement.

To conclude, I would like to address the question regarding the impact of cancel culture on the future of human growth posed at the start of this paper. So far data suggests that cancel culture does exist in society, and that a significant percentage of people who do believe it exists think it has had a negative impact on the community. Historical records suggest that a form of cancel culture has existed since the ancient Greek periods, with ostracism first invented to protect society from radical ideologies.

However, as with most things, too much cancel culture action, in my view, is extremely damaging to society. A streamlining of thoughts and a fear of standing out from the crowd, coupled with the swift and somewhat shallow judgement by the mob, brews a feeling on constriction and imprisonment in the global community, and as argued before, impedes scientific and educational progress.

The permanence of social media means that tweets and comments made in the past stay on the internet forever. This means that comments made by celebrities that may have been accepted in the at the time, may not necessarily be accepted today, leading to people being held accountable for things they said 10-15 years in the past when the context was much different (17). This point is especially true for comedians and entertainers, who, by the very nature of their job, may say potentially risky things with the intention to be funny. Now, there are arguments to be made to suggest that comments about people's abilities/disabilities, race, or gender have no place in humour as it is discriminatory, and even if it is for comedic effect, there is a fine line between comedy, and it being normalised by society. But surely these same arguments cannot be made in relation to scientific studies?

When these studies are conducted, their sole aim is to bring light to information that may have been previously misunderstood, or that is simply wrong. These studies, which have a profound impact on the direction that humanity is heading, must be as objective as possible so that only the facts are presented. Total objectivity can only be achieved if there is no fear of the repercussions, and because of the prevalence of cancel culture in academic society, the hesitation to speak one's mind is increased. This would mean that humans would reach a plateau in both economic, psychological, and sociological growth.

IS THIS A BAD THING?

A plateau in growth would result in the problems with society not being solved. People who are suffering from medical issues may continue to suffer with them due to the academic stagnation. Shuttering schools of thought may lead to a slowing down of economic growth in the form of lower GDP growth. Most importantly, in my opinion, the mistakes humans will inevitably make with regards to justice legislature, gender and race issues may not be solved; the effect of this is wide ranging. The very people who want to change society for the better by cancelling people may end up making society even worse.

Cancel culture impact firms as well. Historically there has been a lack of politicisation of advertisement in media campaigns from mainstream companies, as supporting one political party risks alienating large demographics of customers, negatively impacting revenues and profits on the consumer side, and further risks alienating workers, meaning that people are less willing to apply for

jobs in that company. This would restrict the growth of the company severely. Nowadays, however, neutrality is seen as implicit compliance (18). Consumers are pressuring companies to take a political stance on current affairs. According to a study by Edelman, 64% of consumers are willing to boycott a company based on their stance on a current socio-economic or political issue (19). Boycotting companies for their stance on certain issues could lead to a form of discriminatory behaviour against consumers of those products, linking back to the anti-vax experience, of being fearful of speaking their mind.

A FINAL THOUGHT EXPERIMENT.

In order to sum up the arguments made in this paper; I propose a thought experiment. An external speaker has been invited to a place of higher education, either a secondary school (High school), or a university. This speaker happens to be a devout Christian, who does not believe that gay marriage should be legal. Upon learning this, a section of the student body of this school decide to protest this decision in ways that make speaking there extremely difficult. As a result, the invitation to speak is withdrawn. Personally, I do not agree with what the school did to manage the situation, but I empathise with the protesting students. They would argue that to continue to debate same-sex marriage would de-humanise gay people in the school, implicitly classing them as second-rate citizens. In the act of prohibiting the speaker from expressing themselves, the gay students can go about their lives without the unnecessary anxiety induced by such debate. Furthermore, it helps the school prove that the environment they foster is an inclusive one.

The assumptions made in this thought experiment provide us a glimpse of who is in the right or wrong as well as illustrating one of the aforementioned nuances to the question of cancel culture: the harm caused to others. The first assumption is that there is no more debate to be had around the topic of same-sex marriage. A poll taken by Pew Research Centre in 2019 showed that over 1 in 5 Americans thought that same-sex marriage should not be accepted by society (20). More than a fifth of the American population in 2019, according to this survey, still need convincing on this basic right, so surely one of the best ways of doing so is by publicly scrutinising each sides arguments through the medium of public debate. The open and public nature of such discourse will force people to see the logic in the campaigners' arguments, thus contributing to the effort to build a more inclusive society. Moreover, in the process of justifying their opinions, irrational people become unhinged (21), further adding to what some would call 'ridiculousness' in their arguments. If people are afraid of debate, and live in fear of the arguments posed by the other side(s), how can the campaigners possibly convince other people? Closing off debate and 'no-platforming' contrarian views only serve to preserve them, as those views get circulated in closed off circles on the internet, growing increasingly radical as they are fed back and forth through the hidden echo chambers, thus prolonging an un-inclusive state rather than getting closer to inclusivity.

In the pursuit of complete tolerance, society will lose the ability to be tolerant. (22)

BIBLIOGRAPHY

1. Romano, A. (2019). Why we can't stop fighting about cancel culture. Retrieved 2 July 2022, from https://courses.bowdoin.edu/sociology-1101-spring-2020/wp-content/uploads/sites/319/2020/05/What-is-cancel-culture_-Why-we-keep-fighting-about-canceling-people.-Vox.pdf
2. Ostracism | Definition, Examples, & Facts. (2022). Retrieved 25 May 2022, from https://www.britannica.com/topic/ostracism
3. Wilde, M. The Galileo Project | Biography | Inquisition. Retrieved 2 July 2022, from http://galileo.rice.edu/bio/narrative_7.html
4. Oxford University professor condemns exclusion from event. (2020). Retrieved 1 July 2022, from https://www.bbc.co.uk/news/uk-england-oxfordshire-51737206
5. scientific method | Definition, Steps, & Application. (2022). Retrieved 18 June 2022, from https://www.britannica.com/science/scientific-method
6. Franklin, B. (2022). A quote by Benjamin Franklin. Retrieved 2 July 2022, from https://www.goodreads.com/quotes/4294692-if-we-all-think-alike-no-one-is-thinking
7. Human attention span 2000-2013 | Statista. (2022). Retrieved 17 June 2022, from https://www.statista.com/statistics/689457/human-attention-span-worldwide/
8. Maybin, S. (2017). Busting the attention span myth. Retrieved 17 June 2022, from https://www.bbc.co.uk/news/health-38896790
9. Yar, S., & Bromwich, J. (2019). Tales from the teenage cancel culture. Retrieved 16 June 2022, from https://oglethorpe.edu/wp-content/uploads/2020/01/teenage-cancel-culture.pdf
10. Keane, D. (2022). New rules to ban teachers from politicising Black Lives Matter and British Empire. Retrieved 1 July 2022, from https://www.standard.co.uk/news/uk/schools-impartiality-guidance-teaching-racism-empire-nadhim-zahawi-b983005.html
11. Examples of British Values. (2022). Retrieved 1 July 2022, from https://www.bucks.ac.uk/study/apprenticeships/safeguarding-student-welfare/examples-british-values
12. 10.9 What are my human rights? (2022). Retrieved 1 July 2022, from https://www.mygov.scot/human-rights
13. Wiley, K., Leask, J., Attwell, K., Helps, K., Barclay, L., Ward, P., & Carter, S. (2021). Stigmatized for standing up for my child: A qualitative study of non-vaccinating parents in Australia [PDF] (1st ed., pp. 1-8). Elsevier. Retrieved from https://reader.elsevier.com/reader/sd/pii/S2352827321002019?token=12AD51BD410CC94920AB6DB0967E14D2FD6FB52412D3A97659B343D7E407D83557ADB7C1108900043CD9E1EF3FE5A0BF&originRegion=eu-west-1&originCreation=20220702223626
14. The harm principle. (2022). Retrieved 1 July 2022, from https://www.futurelearn.com/info/courses/introducing-humanism/0/steps/37114#:~:text=John%20Stuart%20Mill%2C%20On%20Liberty,prevent%20that%20harm%20from%20occurring.
15. Freedom of Expression. Retrieved 17 June 2022, from https://www.amnesty.org/en/what-we-do/freedom-of-expression/#:~:text=Governments%20have%20a%20duty%20to,laws%20criminalising%20freedom%20of%20expression
16. Bychawska-Siniarska, D. (2017). Protecting the Right to Freedom of Expression Under the European Convention on Human Rights. Retrieved 16 June 2022, from https://rm.coe.int/handbook-freedom-of-expression-eng/1680732814#:~:text=Paragraph%201%20thus%20provides%20for,authorities8%20and%20regardless%20of%20frontiers
17. Guardians of the Galaxy director fired over offensive tweets. (2018). Retrieved 3 July 2022, from https://www.sbs.com.au/news/article/guardians-of-the-galaxy-director-fired-over-offensive-tweets/mnv9lbvfb

18. Bakhtiari, K. (2020). Why Brands Need to Pay Attention to Cancel Culture. retrieved 30 June 2022, from https://www.forbes.com/sites/kianbakhtiari/2020/09/29/why-brands-need-to-pay-attention-to-cancel-culture/?sh=1b13fa63645e

19. Two-Thirds of Consumers Worldwide Now Buy on Beliefs. (2018). Retrieved 18 June 2022, from https://www.edelman.com/news-awards/two-thirds-consumers-worldwide-now-buy-beliefs#:~:text=Nearly%20two%2Dthirds%20(64%20percent,13%20points%20from%20last%20year.

20. Pew Research Centre. (2019). The Global Divide on Homosexuality Persists. Pew Research Centre. Retrieved from https://www.pewresearch.org/global/2020/06/25/global-divide-on-homosexuality-persists/

21. How to Deal With Irrational People - The Overwhelmed Brain. Retrieved 3 July 2022, from https://theoverwhelmedbrain.com/irrational-people/

22. Elsby, C. (2019). Asses, Arrows, & Undead Cats: An Introduction to Philosophy through Paradox [PDF] (1st ed., pp. 3-4). 11th Dimension Press Toronto, Canada

Are the economic sanctions placed on Russia justified?

A short-listed Independent Learning Assignment (ILA)

Louis Wilby, Upper Sixth

INTRODUCTION

As a response to Putin's decision to invade Ukraine, the UK, the EU, the USA and multiple other nations have decided to impose economic sanctions onto Russia in an attempt to stop the conflict. Putin believes the Western world to be against him and they "have publicly designated Russia as their enemy." (Kremlin, 2022). He is attempting to reunite Russia and its neighbours into a more Soviet like state, aiming to "demilitarise and de-Nazify Ukraine"(Kremlin, 2022) and to bring Ukraine back under Russian control. The economic sanctions have been put in place in an attempt to keep Ukraine's independence from Russian rule. I accept that there must be action to prevent this conflict, however the purpose of this essay is to question whether the current response to the conflict, targeted economic sanctioning, is a useful and justified route to go down.

In order for the economic sanctions on Russia to be justified, I believe two overarching criteria must be met. The first criteria is whether the impact on Russia is significant enough to cause a stop in fighting. This can occur if sustained damage to the Russian economy occurs, leaving them in an almost unrecoverable recession if they continue fighting. This will either cause public unrest and attempts to overthrow the Putin regime, or lead to Putin pulling out of Ukraine in order to protect Russia from further damage and to prevent his population from suffering. The sanctions placed on Russia are in alignment with Smart Sanctioning principles developed by the UN as a response to the humanitarian disaster of the Iraq sanctions. To assess these sanctions, we must first look at Smart Sanctions and see whether they have had any positive effect in the past and whether they seem to be working currently. We must also review Putin's dependence on public approval ratings and the extent to which his power is dependent on the public's perception of him. This will dictate whether Putin will likely pull out of Ukraine if the Russian population suffer, or whether he will continue regardless of the cost to the Russian population. The smart sanctions on Russia look unlikely to have any real effect in stopping the Putin invasion of Russia, however more comprehensive sanctions, including total trade embargoes, may be more likely to stop the invasion. If the sanctions meet this first criteria, we must then review the sanctions to assess whether they are morally acceptable. To do this we can look at the concept of sanctioning through different moral lenses. For this essay, I believe Utilitarianism and Aquinas' Just War Theory to be the most useful lenses to look through. While society is not wholly utilitarian, the utilitarian concept of pursuing pleasures is popular amongst many and is central to the majority of economic theory. Therefore a utilitarian lens would be fairly representative of common moral perception on the economic sanctions. Similarly the Just War Theory is a popular moral approach to conflicts, which many accept to be a very useful way of judging whether there is a moral justification for the conflict. As such, it is seemingly another useful moral lens to look through to decide whether economic sanctions on Russia are justified. Both Utilitarianism and the Just War Theory seem to reach similar conclusions, both suggesting economic sanctioning to be morally unjustified. However, utilitarianism seems to be a far more useful moral lens to use when assessing the morality of the sanctions, as I believe the Just War Theory to be unfit for assessing the morality of sanctioning as they are a replacement for war.

THE IMPACT OF THE SANCTIONS

When exploring the impact of economic sanctions on Russia, it is a three part process. We must look at what the current sanctions on Russia are. From there, we must assess their effectiveness. This is done in two parts, the first is assessing past effectiveness of sanctioning to give a good indication of the likelihood of success. The second is to review the current consequences on Russia and whether they will work in forcing Putin to end the invasion of Ukraine.

For this essay, it must be noted that an economic sanction is a tool used by governments to alter a nation's behaviour by limiting a nation's trade, putting economic pressure on the nation, and limiting access to financial assets. Pape puts it simply, "Economic sanctions seek to lower the aggregate economic welfare of a target state by reducing international trade in order to coerce the target government to change its political behaviour." (Pape, 1997).

Over the years, the nature of sanctioning has changed. The UN ordered the development of smart sanctioning as a response to the humanitarian crisis that followed from comprehensive sanctioning placed on Iraq. The sanctions currently placed on Russia follow suit, they are targeted, smarter sanctions which are " designed to hurt elite supporters of the targeted regime, while imposing minimal hardship on the mass public" (Drezner, 2011).

CURRENT SANCTIONS

The UK has put various targeted sanctions on Russia. There are roughly three elements to the economic sanctioning on Russia, economic, financial and sanctions on Russian oligarchs. The sanctions are as follows:

Economic- The UK aims to phase out all Russian oil imports by 2022. Both the US and EU are attempting to implement similar sanctions, however the EU imports 40% of its oil from Russia, making it hard to reduce the oil imports as the EU depends heavily on Russian oil. I believe this to be a big issue; there is a risk of the whole sanction system being undermined if Russia is able to continue exporting oil to the EU. This ensures Russia still receive vast sums of money for their main export and as such their economy will not suffer anywhere near enough to force a change in policy.

Financial- £470 billion worth of assets belonging to the Russian central bank have been frozen, giving them no access to any foreign reserves thus making it far harder for the central bank to implement any monetary policy to combat the current high inflation levels.

Oligarchs- The UK has also targeted many wealthy Russians, such as Roman Abramovich, with asset freezes and removal of visas. These sanctions, targeting the wealthy, are an attempt to hurt those with the closest links to Putin, encouraging the oligarchs to persuade Putin into realising he is doing the wrong thing.

As a whole, these sanctions are all targeted, they have been placed on Russia's largest export- oil; they have been placed on the Russian Central Bank, reducing their ability to combat recession and influential Russian oligarchs. The targeted nature attempts to avoid severe harm to the general Russian population, but enough harm to cause Putin to change his mind.

PAST EFFECTIVENESS OF SANCTIONING

Gary Hufbauer, Jeffrey Schott, and Kimberly Ann Elliot were the first scholars to do extensive research into the effectiveness of economic sanctions. They investigated 115 cases between 1914 and 1990, and of these 115 cases, they showed 40 successes (Doxey, Hufbauer and Schott, 1985). Initially this seems promising, a 34% success rate is enough to be optimistic about ending conflict in Russia. An example of successful sanctioning is the US sanctioning the Dominican Republic in the early 1960s in an attempt to overthrow Rafael Trujillo and his repressive regime. The US targeted Dominican sugar and oil exports with tariffs and embargoes and by 1964, a new pro-US government was in place in the Dominican Republic. Using the 34% success rate, there is justification for sanctions to be used on Russia as their likelihood of success seems quite high for a tool trying to prevent conflict. The initial perception of sanctioning is a seemingly optimistic one, suggesting sanctions are in fact a very useful tool for a government to use.

However, I believe that Hufbauer, Schott and Elliot's findings are flawed and paint too positive of a picture. The Trujillo regime mentioned above was brought to an end by his US state sponsored assassination, on the back of US naval forces ready to invade if the Trujillo's did not leave. Seemingly military force resolved that conflict, not the sanctioning. The flawed review Hufbauer, Schott and Elliot produced on sanctioning was heavily criticised by Robert Pape, describing it as "overly optimistic" (Pape, 1997). Pape believed only five of the forty successes were actually as a direct result of economic sanctioning. The thirty-five failed successes fell foul to one of these issues:

- Eighteen were settled by force
- Eight were simply failures
- Six were trade disputes, not economic sanctions
- Three were indeterminate

Of the five successes, they only resolved small issues, such as Canada moving its embassy in Israel. I believe that sanctioning is not quite as useful of a tool as Hufbauer, Schott and Elliot present it to be, and often when sanctions have been put in place, it is actions elsewhere which resolve the conflict. I believe sanctions cause the problem of creating a heightened sense of nationalism rather than causing disruption, people stand together when under the attack of sanctions rather than breaking apart. This suggests to me that in Russia sanctions could cause people to stand by the Putin regime rather than turn on it and call for the end of the war. I agree with Pape that if we want to avoid this issue, it must be done "by achieving dramatically higher levels of economic punishment than sanctions have achieved in the past." (Pape, 1997).

Although I do believe Hufbauer, Schott and Elliot to be too positive about sanctioning, I believe at times Pape fails to understand the true nature of economic sanctioning. I believe that sanctions are not simply a stand-alone policy, they are a tool in a wider tool box of foreign policy, resulting in an overly harsh outlook on sanctioning. I believe Pape's binary 'successful' or 'unsuccessful' ranking to be flawed when reviewing sanctions, as Baldwin says, "If the sanctions were modestly successful with respect to several different goals and targets but highly successful with respect to none, Pape would classify them as failures." (Baldwin and Pape, 1998). Sanctioning often has positive impacts in helping achieve an aim, but these impacts are not always large enough for them to be justified as a stand-alone policy. When viewed as part in a large set of foreign policy tools, sanctioning can be viewed as useful. Using the US example in the Dominican Republic, whilst sanctioning was not the final cause of the end of the Trujillo regime, it added to the pressure on the regime and helped ensure the regime was overthrown. In the case of Russia, sanctions are being used alongside Ukrainian military resistance as well as the UK providing weapons to the Ukrainian army, meaning sanctioning can work alongside these other two tools the help stop the Russian invasion.

Past experience suggests that sanctioning could be of use in the case of the Russia-Ukraine conflict. Whilst I believe Hufbauer, Schott and Elliot's findings are "overly optimistic" (Pape, 1997), I do not think they are as useless as Pape presents them to be. When used as one tool in a wider toolbox of foreign policy, they can help create the required pressure to resolve conflict. As such, according to past sanctioning, the current sanctions on Russia could be useful.

EFFECTIVENESS OF THE CURRENT SANCTIONS

The current targeted sanctions have seemingly had very little effect on the Russian economy.

- The Ruble is at 61 Ruble to the US Dollar (Russian Ruble - 2022 Data - 1996-2021 Historical - 2023 Forecast - Quote - Chart, 2022), better than pre-war rates of around 70 Ruble to the US Dollar.
- Russian unemployment rates are sitting at 4% (Russia Unemployment Rate - May 2022 Data - 1992-2021 Historical - June Forecast, 2022)
- A severe rise in inflation, from 9.2% in February to 17.8% in April (Russia Inflation Rate - May 2022 Data - 1991-2021 Historical - June Forecast, 2022), the only notable economic damage.
- Supply side issues such as car companies unable to import microchips. As such car production in Russia is beginning to slow.

With seemingly little impact on the Russian economy, the sanctions clearly provide only a small challenge to the Russian economy and one which the Russian people will have little problem overcoming. To add to this, Putin's public approval rating currently sits at 83% (Putin approval rating Russia 2022 | Statista, 2022) and has only dipped below 60% once in the past 20 years, suggesting Russian nationalism is already very high and the people will easily stand together to fight these rather limited sanctions.

The limited effect the current sanctions are having on Russia makes them unlikely to stop Putin's invasion of Ukraine. Even when used as one policy in a group of many, these current sanctions are unable to provide enough pressure on the Russian economy and people to have any useful impact in stopping the invasion of Ukraine. As such, the current sanctions on Russia are unjustified as they are having very little effect in stopping the invasion of Ukraine.

A SOLUTION

In order for the sanctions to be justified, they must put enough pressure on Russia to complement the military pressure Ukraine is putting on Russia, and result in the end of the invasion of Ukraine. I believe more comprehensive sanctioning can do that. More comprehensive sanctioning in the form of total trade embargoes will drive inflation rates up further and increase the unemployment rate. This will help increase the pressure on Russia as Putin is left with a decision as to whether to pull out of Ukraine to help the suffering people of Russia, or to continue on and leave his people to suffer and potentially turn against him. Stronger economic sanctions are going to be far more effective than the targeted

economic sanctions currently in place. However, although they are more likely to be successful than targeted sanctions, there is still no guarantee of success with comprehensive sanctioning.

THE MORALITY OF ECONOMIC SANCTIONS

The reason for more comprehensive sanctions being avoided today stems from the humanitarian disaster that came from economic sanctions placed on Iraq in the 90's. There has been little moral justification for the approximately 250,000 deaths (Gordon, 2020) caused by these sanctions.

I agree that there is little justification for everything that occurred in Iraq, but I disagree with the logic that all following sets of comprehensive sanctions will follow suit. As such, I believe it to be important to review the morality of economic sanctioning separate to the Iraq case as I believe we have learnt from the mistakes made in Iraq.

LOOKING AT ECONOMIC SANCTIONS USING THE JUST WAR THEORY

The Just War Theory, first proposed by St. Thomas Aquinas, is an ethical theory used to determine whether a war is justified or not. As economic sanctioning is often viewed as an alternative to war, I believe it to be a useful lens to look through to judge whether sanctioning Russia is justified or not.

There are three main components to the Just War Theory: 'jus ad bellum', 'jus in bello' and 'jus post bello', each concerned with the build up to the war, the war itself and the aftermath of the war respectively. When linking economic sanctioning with the Just War Theory: 'jus ad bellum' and 'jus in bello' are usually the two components most heavily focussed on.

The sanctions on Russia currently seem to align with the 'jus ad bellum' principles of the theory. Within the 'jus ad bellum' section of the theory, there are six criteria which need to be met. The sanctions must have a 'Just Cause'- which in this case they do, as a response to the invasion of Ukraine; they must be of 'Proportional' force to the harm being done currently, I do not think that they exceed the force currently being used in the Russian invasion. They must also have the 'Right Intention' which they clearly do, the intention being to stop Russia from invading Ukraine; the sanctions have also been commissioned by a 'Legitimate Authority'. The concept of 'Last Resort' is irrelevant to sanctioning as sanctioning comes before the deployment of the military. The only criteria the current sanctions may not fulfil is the 'Reasonable Chance of Success' criteria. The current targeted sanctions, using the reasoning above, are seemingly unlikely to have the desired effects of stopping the Russian invasion. If the sanctions were raised to more comprehensive ones, which I believe to have a higher (but not complete) chance of success, I believe all 'jus ad bellum' criteria to be met by the sanctions placed on Russia.

The principles of 'jus in bello' are where economic sanctioning seemingly does not align with the Just War Theory. Whilst it is accepted the criteria of 'proportionality' and 'necessity' are often met within war, it is usually the criteria of 'discrimination' which sanction fails to fulfil. 'Discrimination' requires non-combatants and combatants to be distinguished between and only combatants are allowed to be attacked, combatants being soldiers and those in government making policy to do with the war, and non-combatants being civilians not fighting in the war. All economic sanctions, whether that be the targeted ones currently on Russia (to a lesser extent), or totally comprehensive sanctions, focus on non-combatants as a way of stopping war. The current sanctions target oligarchs who have had no part to play in the Russian invasion of Ukraine, and comprehensive sanctioning on Russia would target the whole population. Joy Gordon likens the focus on non-combatants to siege warfare, saying, "the harm is done to those who are least able to defend themselves, who present the least military threat, who have the least input into policy or military decisions, and who are the most vulnerable" (Gordon, 1999). The concept of sanctioning is centred around causing harm and distress to non-combatants in an attempt to get them to help force a change in behaviour from the sanctioned government. This completely undermines the 'discrimination' criteria in the Just War Theory and as such, the Just War Theory finds all economic sanctioning morally unacceptable.

Despite this, I believe that the Just War Theory cannot be applied to economic sanctions at all. I believe war and economic sanctioning to be too dissimilar to warfare for them to both be judged by the Just War Theory. Despite sanctioning being seen as an alternative to warfare, I do not believe that makes them like warfare and as such, the two cannot be treated the same. I agree with Christiansen and Powers' belief that too many scholars "often treat economic sanctions as analogous with acts of war... along with blockades and siege warfare" (Christiansen and Powers, 1995) but in reality, sanctioning should have a different moral status to warfare for three reasons. The first reason, and what I believe to be the most compelling reason, is that sanctions are not imposed as warfare themselves but rather as an alternative; the second reason is that some harms to non-combatants are accepted as they have bought in to the government's ideas and plans; and the final criticism is that if appropriate humanitarian measures are built in, the harms of sanctioning are far less than the harms of war. Christiansen and Powers give an analogy in support of my argument when comparing sanctions with warfare, 'warfare is akin to the death penalty, whereas sanctions are more like attaching someone's assets in a civil proceeding' (Christiansen and Powers, 1995). I agree that the idea the two scholars try to convey is a useful one, there are differences between the two, however I believe that their presentation of economic sanctions is not fully accurate. They present economic sanctions as mere inconveniences which a nation has to tackle for a short period of time, however in reality, sanctions pose major problems to civilians, with the aim of causing problems so big that people protest against the government. Sanctions are slightly more severe than Christiansen and Powers make them out to be and as such their criticisms must be adapted slightly to show the Just War Theory to be useless in judging the morality of economic sanctioning. Despite this, I still believe that the argument that war and sanctioning are too different to be compared does hold true, and despite Christiansen and Power's slight misrepresentation of economic sanctions, their main message is still vital, that warfare and sanctioning are too different to both be judged by the Just War Theory.

On the whole however, there are too many differences between sanctioning and warfare. Sanctioning is often preventative, in the case of these sanctions, despite sanctioning being there to help Ukraine, I argue it is not unreasonable to suggest that the UK and other nations are also implementing sanctions as preventative measures, stopping Russia from amassing further power and military strength and shifting their focusses elsewhere. Preventative measures are not justified by the Just War Theory as they have no 'Just Cause', yet a preventative measure in the form of freezing assets is far less detrimental than a preventative measure in the form of a bomb.

I believe there are too many differences between sanctions and war for the two to be judged by the same moral criteria and whilst the view of Christiansen and Powers presents a slight oversimplification as to what economic sanctions truly are, their main message is vital in showing that economic sanctions cannot be judged by the Just War Theory.

A UTILITARIAN VIEW ON ECONOMIC SANCTIONS

If the Just War Theory offers little help in the moral justification of economic sanctioning, another ethical theory must be used to work out whether the sanctions on Russia are morally acceptable. Utilitarianism, despite its flawed nature, is an ethical theory at the centre of economic and political theory as well as often being used as a tool by the individual to make moral judgement. As Robert McGee puts it, "To completely ignore utilitarian ethics would be to have an incomplete analysis" (McGee, 2014). As such, utilitarianism is a helpful moral lens to judge economic sanctions by, as it is a widely used ethical theory and is representative of many people's beliefs and moral outlook on the world.

Basic utilitarian theory, according to Jeremy Bentham, is the idea that an action is the right thing to do if it maximises pleasure and minimises pain (Bentham, 2012). Thus, in the case of sanctioning, sanctions are justifiable if the pleasure they create outweighs the harm and suffering they cause.

A key component to judging whether sanctioning is justified or not from a utilitarian perspective is assessing the sanctions' likelihood of success. With little likelihood of success, sanctions could never be justified as the pain and suffering felt by the civilians being sanctioned comes with no benefits of preventing warfare or achieving the political aim. As an inherently consequentialist ethical theory, if the consequences of the action are unlikely to be the desired ones, it can never be seen as morally acceptable as the so called 'greater good' is not achieved and only suffering occurs.

When applied to the current sanctions on Russia, it was concluded above that the current smart sanctions are unlikely to be enough to end the war. Any suffering they have caused to Russian civilians, which, although limited when compared with a set of more comprehensive sanctions, is unjustified as there is no pleasure (stopping the war) to outweigh the suffering.

When looking at more comprehensive sanctions, which have been suggested to be more useful, the suffering of the Russian population will be far greater. However, with an increased likelihood of success, does the pleasure and happiness given to the Ukrainians and the Western world from stopping the Russian invasion outweigh this suffering? It is beneficial to the majority outside of Russia if the war is stopped and as such, it seems that the happiness derived from the stopping of Russia's invasion could outweigh the suffering of the Russian people.

However, this highlights major flaws in the utilitarian approach to ethics. Utilitarianism requires us to predict the future. In the case of sanctioning, there is a chance of success, but it is by no means a certainty. Therefore, it is hard to judge whether an action is the right thing to do if you are unsure of the consequences. As a purely consequentialist theory, where an action is right or wrong based on its outcome, when there is uncertainty over the success of an action, it is very hard to work out the best decision to make. In the case of sanctioning Russia, with uncertainty over the success of these sanctions, it is very hard to judge whether comprehensive sanctioning is the morally correct thing to do. In addition, utilitarianism has a very impersonal nature for a moral criteria. It groups people as a whole, rather than taking into account the individual and this causes issues, especially in regards to human rights. Pain and pleasure gauges do not take into account human rights principles and as such, regardless of any human rights abuses sanctions may cause, such as a loss of rights to safety from violence, to healthcare, education and adequate living standards, as well as rights to humane treatment. This disregard for the individual is an issue for many, as human rights are central to the moral outlook of many. As such utilitarianism has it flaws in being a useful tool to review the morality of economic sanctioning, however is still a good representative of basic public mindset.

I believe that it is clear, if the comprehensive sanctions were to work quickly and effectively, utilitarian theory would find sanctioning to be morally acceptable. However, it is easy for us to also imagine a scenario where sanctions are unsuccessful and result in only suffering for the Russian population as well as the continued invasion of Ukraine. This uncertainty, for me, is enough to suggest that from a utilitarian point of view, comprehensive economic sanctioning is unjust. I believe that due to a lack of clear evidence to suggest that sanctioning is a guaranteed success, a utilitarian would err on the side of caution and suggest comprehensive sanctioning to be the wrong action to do.

It is clear, that the current sanctions placed on Russia are unjust according to utilitarian theory as they are unable to achieve their goal of stopping Russian invasion, and have caused some suffering to Russian civilians. With comprehensive sanctions, my suggested amendment, it is unclear as to whether these are right or wrong due to the problems of utilitarianism listed above and the uncertainty of success. However I believe this lack of clarity to suggest that a utilitarian approach would find comprehensive sanctioning unjust as there is too high of a chance of the sanctions not succeeding and as such only the large scale suffering of the Russian population occurring.

CONCLUSION

I believe that the answer to the question, 'Are the economic sanctions placed on Russia justified?', simply put: no. The current targeted sanctions have been shown by history to be highly ineffective and very unlikely to resolve a major conflict like the Russian invasion of Ukraine. In addition, they have little moral justification from a utilitarian point of view.

However, I believe the answer to the question when using a more comprehensive set of economic sanctions to be more nuanced. While comprehensive sanctions are far more likely to be successful than targeted sanctions, this likelihood of success is still by no means a guarantee. Although being used in combination with Ukrainian military action and Western aid to Ukraine, it is still uncertain as to whether these will bring an end to the war. These uncertainties, whilst okay from an economic point of view as they can bring success, are not okay from a moral point of view. Utilitarianism , despite being slightly unclear on the matter, would likely judge comprehensive sanctioning to be morally unacceptable.

If a comprehensive set of sanctions could be drawn up, so useful and effective that the chances of success are almost guaranteed, utilitarian theory could justify sanctions use. However, with limited likelihood of success, even with a more comprehensive set of sanctions, from a moral standpoint, there is little justification of the economic sanctions on Russia.

REFERENCES

Baldwin, D. and Pape, R., 1998. Evaluating Economic Sanctions. *International Security*, [online] 23(2), pp.189-198. Available at: <https://www.jstor.org/stable/pdf/2539384pdf?refreqid=excelsior%3A6c322cceff79015b5ec6ff3e89c2a1bc&ab_segments=0%2FSYC-6398%2Fcontrol&origin=> [Accessed 2 June 2022].

Bastiat, F.,1850. Ce Qu'on Voit et Ce Qu'on ne Voit Pas [What Is Seen and What Is Not Seen], first published as a pamphlet; reprinted in Oeuvres Complètes de Frédéric Bastiat (1862) Vol. V: 336-392. Also reprinted in English in Bastiat 1964: 1-50; 2007, I: 1-48.

Bentham, J., 2012. *Introduction to the Principles of Morals and Legislation. Dover Publications.*

BBC News. 2022. *What sanctions are being imposed on Russia over Ukraine invasion?*. [online] Available at: <https://www.bbc.co.uk/news/world-europe-60125659> [Accessed 6 June 2022].

Christiansen, D. and Powers, G., 1995. *Economic Sanctions and Just-War Doctrine. In: Economic Sanctions: Panacea or Peacebuilidng in a Post-Cold War world?*, 1st ed. New York: Taylor And Francis.

Doxey, M., Hufbauer, G. and Schott, J., 1985. Economic Sanctions Reconsidered: History and Current Policy. *International Journal*, 41(1), p.264.

Drezner, D., 2011. Sanctions Sometimes Smart: Targeted Sanctions in Theory and Practice. *International Studies Review*, [online] 13(1), pp.96-108. Available at: <https://www.jstor.org/stable/pdf/23016144.pdf?refreqid=excelsior%3A72153dbfa49c30eb48c1e1f1dd9642a0&ab_segments=0%2FSYC-6398%2Fcontrol&origin=> [Accessed 2 June 2022].

Ellis, E., 2020. *The Ethics of Economic Sanctions*. PhD. University Of Edinburgh.

Ellis, E., 2020. The Ethics of Economic Sanctions: Why Just War Theory is Not the Answer. *Res Publica*, 27(3), pp.409-426.

Giumelli, F., 2013. *How EU sanctions work*. [online] European Union Institute for Security Studies. Available at: <https://www.jstor.org/stable/pdf/resrep06969.6.pdf?refreqid=excelsior%3Adf0651e0d70027eb4b1183b29a32e0f6&ab_segments=0%2FSYC-6398%2Fcontrol&origin=> [Accessed 6 June 2022].

Gordon, J., 1999. A Peaceful, Silent, Deadly Remedy: The Ethics of Economic Sanctions. Ethics & International Affairs, 13, pp.123-142.

Gordon, J., 1999. Economic Sanctions, Just War Doctrine, and the "Fearful Spectacle of the Civilian Dead." 49, pp.387-400.

Gordon, J., 2020. The Enduring Lessons of the Iraq Sanctions - MERIP. [online] MERIP. Available at: <https://merip.org/2020/06/the-enduring-lessons-of-the-iraq-sanctions/> [Accessed 2 June 2022].

GOV.UK. 2022. Russia sanctions: guidance. [online] Available at: <https://www.gov.uk/government/publications/russia-sanctions-guidance/russia-sanctions-guidance> [Accessed 6 June 2022].

Greene, S. and Robertson, G., 2022. Putin's power depends on his popularity. That makes him vulnerable.. [online] The Washington Post. Available at: <https://www.washingtonpost.com/outlook/putins-power-depends-on-his-popularity-that-makes-him-vulnerable/2019/08/27/c5e0cf1a-b4a2-11e9-8e94-71a35969e4d8_story.html> [Accessed 2 June 2022].

Jones, L., 2022. Ukraine sanctions: What pain lies ahead for Russia's economy?. [online] BBC News. Available at: <https://www.bbc.co.uk/news/business-61381241> [Accessed 2 June 2022].

Justifiable?, I., 2022. In-Depth: Are Economic Sanctions Justifiable? - McGill Journal of Political Studies. [online] McGill Journal of Political Studies. Available at: <https://mjps.ssmu.ca/2020/08/04/in-depth-are-economic-sanctions-justifiable/> [Accessed 3 June 2022].

Kremlin, T., 2022. Address by the President of the Russian Federation. [online] President of Russia. Available at: <http://en.kremlin.ru/events/president/news/67843> [Accessed 22 May 2022].

McGee, R., 2014. Ethical Aspects of Economic Sanctions: A Third Theory. SSRN Electronic Journal, [online] Available at: <https://deliverypdf.ssrn.com/delivery.php?ID=499084127102074118069071107109029028036054068006069016124066127112003097024081078073000123007101038043029024096003100073104029114042012092028104127088022006038087094117085112082108031103065107103006016085125006090073108028076112079069105007005&EXT=pdf&INDEX=TRUE> [Accessed 3 June 2006].

Muchakazi, P., 2018. The Ethical Dilemma of the Imposition of Economic Sanctions as a Deterrent Tool against a Sovereign State: A Critical Analysis with reference to Cuba, Iraq and Zimbabwe. PhD. UNIVERSITY OF KWAZULU–NATAL.

Pape, R., 1997. Why Economic Sanctions Do Not Work. International Security, [online] 22(2), pp.90-136. Available at: <https://www.jstor.org/stable/pdf/2539368.pdf?refreqid=excelsior%3A2e0230638655de7e229b81b1ee87a6b5&ab_segments=0%2FSYC-6398%2Fcontrol&origin=> [Accessed 2 June 2022].

Statista. 2022. Putin approval rating Russia 2022 | Statista. [online] Available at: <https://www.statista.com/statistics/896181/putin-approval-rating-russia/> [Accessed 2 June 2022].

Tradingeconomics.com. 2022. Russia Inflation Rate - May 2022 Data - 1991-2021 Historical - June Forecast. [online] Available at: <https://tradingeconomics.com/russia/inflation-cpi> [Accessed 2 June 2022].

Tradingeconomics.com. 2022. Russia Unemployment Rate - May 2022 Data - 1992-2021 Historical - June Forecast. [online] Available at: <https://tradingeconomics.com/russia/unemployment-rate> [Accessed 2 June 2022].

Tradingeconomics.com. 2022. Russian Ruble - 2022 Data - 1996-2021 Historical - 2023 Forecast - Quote - Chart. [online] Available at: <https://tradingeconomics.com/russia/currency> [Accessed 2 June 2022].

Winkler, A., 1999. Just Sanctions. Human Rights Quarterly, [online] 21(1), pp.133-155. Available at: <https://www.jstor.org/stable/pdf/762739.pdf?refreqid=excelsior%3A46172fccdef321415974106a327af9f2&ab_segments=0%2Fbasic_phrase_search%2Fcontrol&origin=>.

Data Visualisation: An Art or Science? Exploring the Effects of Aesthetic & Dataset Complexity on Graphical Effectiveness

A research report written following an Original Research in Science (ORIS) placement

Michael Wu, Upper Sixth

ABSTRACT

Many business decisions need to be supported and affirmed with data, and data visualisations are often used to present the data in such a way that it is made easier to interpret and gain insights from. Although data can be visualised with various forms (e.g. bar charts, pie charts, and tables), little is known on their effectiveness and accuracy in assisting decision making. Furthermore, in order to enhance the appeal of visualisations, frequently beautification attempts are made with little to no regard as to whether there are any effects on efficiency of interpretation. Therefore, in this study, we conducted a comprehensive analysis of volunteer-participant surveys to determine the effectiveness of three of the most commonly used data visualisation types – Table, Bar Chart, and Pie Chart. We used four real datasets (of which only one, Dataset 2, is assessed in this report for the sake of brevity), for which we evaluated five common data analysis and interpretation tasks. We examined the effectiveness of bar and pie charts, both in both black & white as well as colour, and measured the perceived visual appeal of all visualisation types. We find that tables are the most accurate and often fastest for categorical and numerical tasks, followed by the bar charts. Also, the bar charts displayed the strongest pattern identification ability and the coloured bar charts were the most visually appealing. We also find that the colour of the visualisation has little effect on its accuracy or response times.

1. INTRODUCTION

It is reported that humans generate 2.5 quintillion bytes of data daily (Statista, 2021). It goes without saying that we have more information than we know what to do with. Hence it becomes imperative that we find ways to simplify data into more approachable and accessible forms. This is generally realised with data visualisations, the aim of which is to increase the efficiency of interpretation and enhance understanding of the data. Humans have evolved an incredible ability to detect and recognise patterns through our visual processing system, which is largely effortless and intuitive (Pinker, 1988). Thus, data visualisation seeks to leverage the visual processing to improve the communication of, and insight into datasets. This should enable analysts to communicate their insights more effectively, and for decision-makers to make better informed decisions and thus improve the operation of business (Yigitbasioglu & Velcu, 2012).

Nowadays, computers are used to store, obtain and process large volumes of data. This data can be easily processed and converted into various visual forms, of which the most common in business are still the simple bar chart, pie chart and line graph. With the progress in data processing, improvements in data presentation are also made to the point where the data analyst often has a specific choice of data visualisation type, colour, font type and size, line width, category order, and so on. However, these raise many questions on the use of these visualisation types. For instance, does colour affect ability to compare bars? Does it affect ability to spot patterns, to calculate totals, to identify the greatest or lowest datapoint? Considerable efforts have hence been devoted to examine the impact of data visualisation on decision making (Bishop et al., 2013; Chertov et al., 2002; Chinnaswamy et al., 2019; Quattrone, 2017).

The efficiency of graph types in data visualisation was first investigated by Cleveland and McGill (1982, 1984, 1985, 1993) who explored how human-subjects interpret different visualisation types, resulting in a categorisation of different graphs into an implicit hierarchy for different data comprehension and analysis tasks. The hierarchy suggests that positions against scales and lengths were more accurate graphical encodings than angles and areas, thus implying that bar charts are more accurate than pie charts. Since then, cognitive models of graphical perception were proposed (Carpenter & Shah, 1998; Friel et al., 2001; Hegarty, 2011; Lohse, 1993) and further explored using such advanced technologies as eye tracking (Okan et al., 2015; Strobel et al., 2016) and brain sensing (Peck et al., 2013).

The influence of colour and aesthetic appearance on graphical perception was also investigated. Most research on the influence of colour was attributed to Benbasat et al. (1986), who aimed to *"assess the influence of graphical and colour-enhanced information presentation on information use and decision quality in a simulation setting"*, and Cawthon & Moere (2007) who investigated the "*under-represented*" effect of aesthetics on efficiency and effectiveness of information retrieval in various complex visualisation types.

Some findings have ultimately found their way into common usage in business analytics and data visualisation in general. For example, many advocates of data visualisation now vehemently disown the pie chart (*"Pie charts are evil"* - Knaflic, 2015). However, these commandments – thou shalt not use pie charts, thou shalt not use 3D – lack subtlety, and the research that supports them is often based on small studies of, at best, tens of university students. The reality is likely far more nuanced, with context being vital in which graph type is appropriate. For example, pie charts are primarily designed to show the part-whole relationship, and the evidence appears that they are very good for this task. Also, there is now substantial evidence that people are not comparing area when using a pie chart, but rather comparing arc length, or some other factor (Skau & Kosara, 2016a). New technology (e.g. Virtual Reality and Augmented Reality) now open up the possibility that 3D graphs – properly represented in 3D space – as viable visualisation techniques.

In summary, there is still a lot of research to be done in data visualisation. There is lots of dogma based on sketchy data, and we believe there is still a much better balance to be struck between (useful) accuracy, design theory and simple aesthetics.

Therefore, the aim of this study is to perform a comprehensive evaluation on the efficiency and accuracy of the most commonly used chart types in presenting real datasets. For this purpose, we conducted a survey to assess the graphical efficiency of two different 2-dimensional graph types (Bar Chart, Pie Chart) and using a Table as a control on five different graphical interpretation and analysis tasks. We also assessed the effect of colour on the responses to the bar and pie charts and analysed the data using the Monte Carlo random sampling approach.

2. HYPOTHESES

A priori, we assume that relevant metrics to assess the efficiency of a graph type are accuracy, response time, and visual appeal. Based on prior research and personal expectation and experience, we decided to test the following hypotheses:

- When participants are asked to compare the difference or sum of two categories, tables should outperform both bar and pie charts in terms of accuracy and response times, but bar charts should be more accurate and faster than pie charts.
- When participants are asked whether there is a discernible pattern in the dataset, bar charts will outperform pie charts, which will outperform tables, in accuracy and speed of response.
- The coloured bar and pie charts are the most visually appealing, followed by their black & white counterparts. The table is the least aesthetic visualisation type.
- The difference in accuracy between bar and pie charts is smaller than the natural variance typically observed in real-world datasets.

3. METHODOLOGY

3.1. SURVEY & STUDY DESIGN

An online survey was designed based on 4 datasets which were taken from real life data rounded to varying degrees of accuracy (to resemble real imperfect data), and were chosen because they contained not only a variety of contexts but also, more significantly, a relatively large number of categories: the lowest being 7 and the highest 12, which are much more than that used in previous studies where often no more than five categories were used (Simkin & Hastie, 1987; Skau & Kosara, 2016a, 2016b; Spence, 1990; Zacks & Tversky, 1999). This made the questions more realistic and could be used to evaluate participants' graphical perception to data related to daily life. The data for each dataset was converted to percentage compositions, and the participants were informed in the question description that the total percentage would always be 100%.

With these datasets, we devised an online survey using Qualtrics software (the survey can be accessed via this link: https://surreyfbel.qualtrics.com/jfe/form/SV_9ZuURBpZuc5JN42), with the same types for questions for each dataset and graph type. Prior to the questions for each dataset, a preface was included to give context to the data about to be shown.

Each dataset was plotted with five different graph types: i) Bar chart with colour, ii) Pie chart with colour, iii) Bar chart in black & white, iv) Pie chart in black & white, and v) Table. The table set was included as a control variable against the bar and pie chart variations. For each participant and dataset, only one of these five graph types was randomly selected with uniform probability and presented to a participant. Regardless of the graph/table being selected, the survey contained the same questions in the same order for all datasets.

3.2. DESIGN OF GRAPHS & TABLES

An example of the coloured bar and pie charts from Dataset 1 are given in Figures 1 and 2 below.

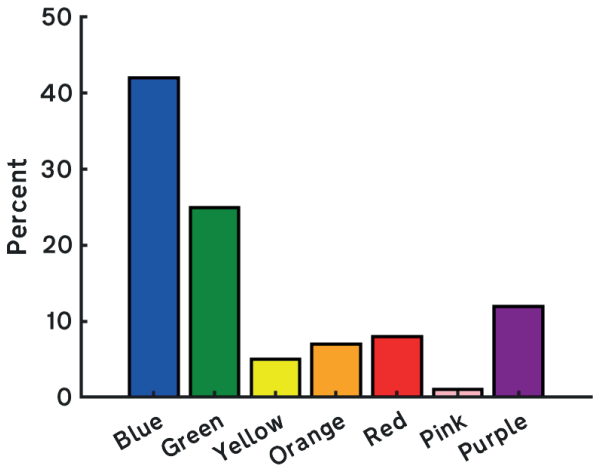


Figure 1: Example of a coloured bar chart presented in survey. Here, the coloured bar chart for Dataset 1.

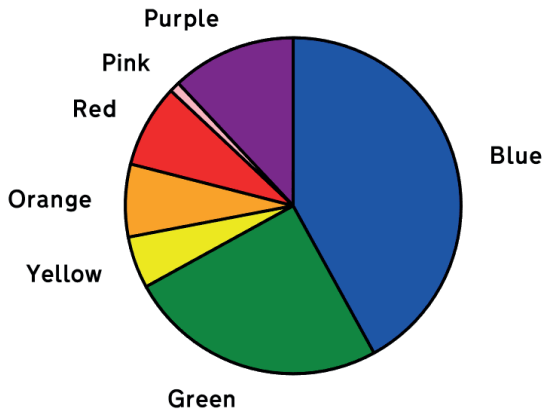


Figure 2: Example of a coloured pie chart presented in survey. Here, the coloured pie chart for Dataset 1.

The graphs and tables were all labelled using transport font (which was specifically designed for ease of reading on road signs) to reduce the level of mental load that participants would expend on reading the category labels.

The black & white versions of the bar and pie charts had white bars or slices with black outlines, axes and labels. The colour schemes used in the colour charts are shown in Figures 3 & 4. The tables were all black & white.

Figure 3: Colour scheme used for Dataset 1.

Figure 4: Colour scheme used for Datasets 2-4.

3.3. SURVEY QUESTIONS

A diverse range of question types was carefully considered for each dataset to maximise the use of the data visualisation type. An attempt was also made to increase the difficulty of the questions as the respondent progressed in the survey to challenge them and augment the difference in response accuracy if any

discrepancies did exist. The question types used are explained in detail in Table 1. These questions were always presented in the order as listed in Table 1 and were always placed under the corresponding graph or table.

Table 1: Question types used in the survey.

Question Type		Datasets Used In	Explanation	Example
Categorical	Placement	1, 2, 3	This question involved asking the respondent to identify a certain category from the data visualisation type shown. These were all multiple choice and a list of all categories in the given dataset were shown from which the participant could select only one option. This question type was not used in dataset 4 as it was deemed too easy.	*"From the graph, which store has the smallest market share?"*
Numerical	Difference	All	These questions required the participant to estimate the absolute difference in size of two different categories. The phrase *"percentage point difference"* was used in an effort to obtain the absolute difference, not relative difference in size.	*"From the table, what is the difference in percentage points of the numbers 2 and 8?"*
	Valuation	All	Valuation questions asked the responder to estimate the value of a stated category.	*"What is the percentage of people who had a favourite colour of Red?"*
	Summation	All	Summation questions asked the partaker to approximate the total size of two different specified categories.	*"What is the total market share of Communications and Energy?"*
Pattern		All	The pattern question asked the respondent's opinion on whether the data shown contained a pattern, with three optional answers *"Yes"*, *"No"* and *"I don't know"*, from which only one could be chosen.	*"Is there a pattern in this data?"*
Visual Appeal		All	Visual Appeal questions sought to find out how visually appealing the given visualisation type was. This was done via a Likert scale with options *"Very unappealing"*, *"Unappealing"*, *"Neutral"*, *"Appealing"* and *"Very Appealing"*.	*"How visually appealing is this graph?"*

3.4. SURVEY CAMPAIGN

The survey was created using the Qualtrics experience management software. We promoted the survey using online forums to attract as many participants as possible. The survey was distributed and posted on numerous social media platforms and internet forums over a 3-week period. In doing this, participants with as wide a range of graphical interpretation experience as possible could be engaged. A total of 2,900 unique participants voluntarily did the survey.

3.5. DATA PROCESSING

Upon receiving all responses, any incomplete questionnaires were removed as well as those from anyone who did not accept the terms of agreement. Responses from participants who took longer than 1800 seconds (i.e. 30 minutes) to complete the survey were then removed as well as responses from anyone who took longer than 180 seconds (i.e. 3 minutes) on any question. For the numerical response questions, all numbers were formatted identically and invalid responses were replaced with "NA". Participants who had more than 50% of their responses as "NA" were then removed from the dataset. The original sample (2,900 participants) and final cleaned sample (1,142 participants) are shown in Figure 5. Then accuracies of the remaining responses in the final sample were then calculated. For categorical questions this was a simple "TRUE" or "FALSE" depending on whether the participant selected the correct answer or not and for numerical questions this was the absolute difference between the participant's response and the actual answer.

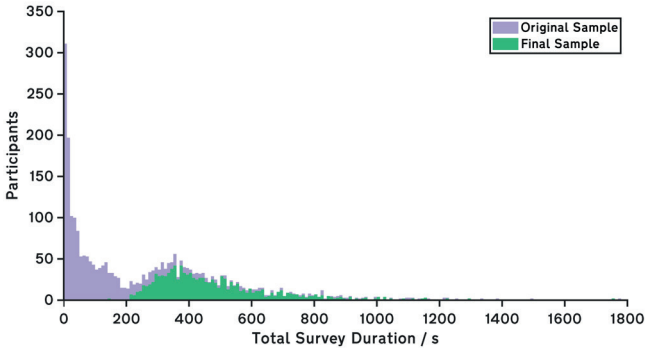


Figure 5: Histogram of original sample (2,900 participants) survey duration against final cleaned sample (1,142 participants) survey duration.

To determine whether there was an effect on the time taken to answer between the coloured and black & white charts and also between different devices, two-sample Wilcoxon signed-rank tests were performed. No significance was obtained so the times taken were combined into one set for each of the three visualisation types (Bar, Pie, and Table). These same tests were also undertaken to check whether there was an effect of device used and colour or black & white on the accuracy of response. Again, they came with no significance in most questions so the data was again combined.

The combined survey results were then further analysed to assess the accuracy of response; variation in response times of all questions; and response times between data visualisation types. Percentages of response options for the categorical, pattern and visual appeal questions were determined and compared. The data for the visual appeal questions was split into five graph types – Bar Chart with Colour, Pie Chart with Colour, Bar Chart in Black & White, Pie Chart in Black & White, and Table – for comparison, rather than just three – Bar Chart, Pie Chart, and Table.

4. RESULTS

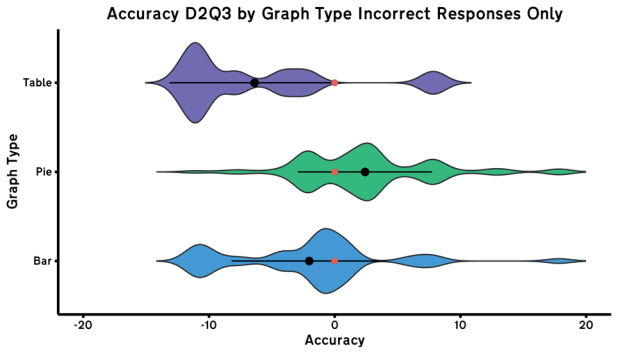
4.1 EXAMPLE OF RESULTS

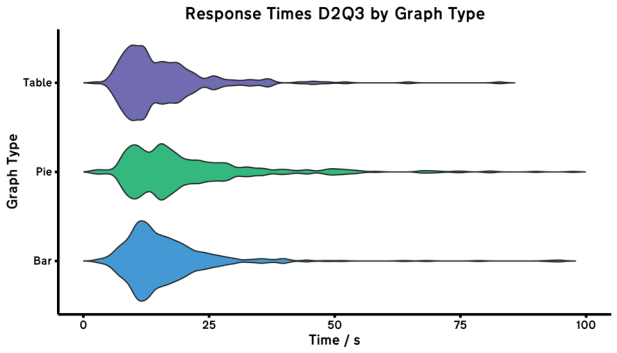
For the 1,142 respondents with valid responses the average time to complete the survey was 486 seconds (8 minutes 6 seconds), of which 799 used a smartphone and 343 used a computer. An example of the results for Question 3 of Dataset 2 questions can be found in Table 2.

Table 2: Overview of the results of Question 3 from Dataset 2, containing elaborations of statistical significances. Included is a violin plot of response times in seconds; a table containing contains the percentage of correct answers, which we refer to as accuracy, for each graph type where "correct" for numerical questions is answers within ±0.5 of the true value; and a violin plot of the distribution of accuracies of the incorrect answers, where 0 (the true value) is shown as a red dot and the mean and one standard deviation is displayed by the black dot and line respectively.

D2Q3 – VALUATION

Graph Types:	Percent Correct / %:
Bar	90.89
Pie	18.25
Table	96.51





Accuracy: (T>B>P)

The Table is most ($p < 0.01$) accurate, followed by the Bar Chart, which is significantly ($p < 0.001$) more accurate than the Pie Chart.

Time: (T~B>P)

Results show that the Table and Bar Chart are both significantly ($p < 0.001$) faster than the Pie Chart.

4.2. ACCURACY

From the analysis based on Monte Carlo random sampling of the responses and the Kruskal-Wallis tests to the categorical questions, it was found that presenting the data with a table was always significantly ($p < 0.05$) more accurate than the bar and pie charts. This is to be expected as the actual values of each category can be determined from the table whereas from the bar and pie the sizes of the categories have to be compared and the values to be estimated, which is significantly more difficult if the categories are similar in size. No significance was found between the accuracy of the bar and pie charts.

For numerical questions, the table was also always significantly ($p < 0.005$) more accurate than the bar and pie charts. This is interesting to notice as the participants could easily locate the exact values required and sometimes perform a short calculation to obtain the exact answer. Interestingly, the bar charts always performed significantly ($p < 0.001$) better than the pie charts at these questions. There are a number of factors which could have led to this outcome, some being the difficulty of the questions, the fact that the bar charts contained a labelled y-axis which would give the participants an approximate range for which their answer could reasonably lie etc.

The data for the accuracies of the pattern recognition questions were much more scattered than any other question types. After all, whether or not there is a pattern in the data is very subjective and despite the selected datasets being very definitive as to whether or not they represented a pattern or not, there could still be underlying social and economic principles which mean that there is actually a pattern in some of the first three datasets even with the rearranged orders of some of the categories. We have split the results of this section into four parts in order to discuss the results of the pattern question for each dataset separately.

In dataset 2 there was no apparent pattern associated with the data and the Monte Carlo sampling returned no significance between any graph type. The highest percentage correct was from the coloured bar responses, 55.2%, and the lowest was from the table, 44.2%.

4.3. VISUAL APPEAL

Ten comparisons were required for the Kruskal-Wallis tests and Monte Carlo samples of the Bar Colour (BC), Pie Colour (PC), Bar B&W (BBW), Pie B&W (PBW), and Table (T), of which the Coloured Bar was always significantly ($p < 0.05$) the most popular. It is expected that the Coloured Bars and Pies are the most popular and the least visual. The Tables are expected to be the most unappealing, but surprisingly, the Coloured Pie Chart and Black & White Bar were equally appealing, with the Table coming next. Despite being visual, the PBW was the least popular, being significantly ($p < 0.05$) disliked compared to all other graph types in all datasets apart from the T in dataset 2 where $p = 0.071$.

4.4. RESPONSE TIME

The analysis based on the Kruskal-Wallis tests and the Monte Carlo random sampling on the response times to all questions showed that significance was present in all question types apart from the visual appeal questions. Across all question types however, we find that the table was overall the quickest, followed by the bar chart, with the pie chart generally the slowest visualisation type.

The response times to the categorical questions were significantly different, although the quickest varied between questions, and were generally conducted fastest by participants presented a bar chart and slowest by those presented a pie chart. Our results also suggest that the order in which the data is presented has an effect on the speed of responses from the tables.

The difference questions were performed fastest by table users, and slowest by pie chart users. We expect the tables to be the most efficient at this task as it is possible to calculate the exact value of the absolute difference whereas in the bar and pie chart a complicated comparison often had to be performed.

The order of response times for valuation questions was the same for all datasets, where they were expectedly completed fastest on the tables. Second fastest was the bar chart, likely due to the labelled y-axis which they could quickly obtain an approximate answer for. The pie chart was slowest again, presumably due to the large slice count making it difficult to estimate values and the ease at which this task could be performed on the other visualisation types.

Interestingly, the times taken for the summation questions were significantly different, but not consistent: with the pie charts surprisingly quickest the most times, followed by the tables then the bar charts. Interestingly, Cleveland & McGill (1984) also observed the similar phenomenon. We find it relevant to note that in all summation questions, the answer was always less than the value of the greatest bar, meaning that the answer lied on the labelled y-axis which could therefore be used to help approximate an answer. However, the pie chart was overall the fastest.

5. DISCUSSION

5.1. MEANINGFUL MAGNITUDE OF DIFFERENCES

Besides being able to show significant differences between the three visualisation types, it is worthy discussing whether these differences would actually have a meaningful effect in the real world. We therefore present the following tables showing accuracies as percentages of correct responses and response times in seconds.

We also note that some of the means, especially for response times, seem to contradict significance findings based on Kruskal-Wallis and Monte Carlo testing. This is likely due to a long tail of the response times, as shown in Figure 6.

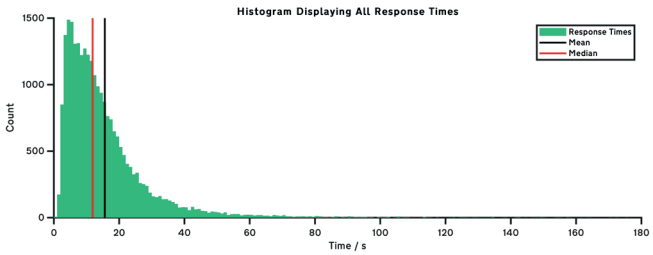


Figure 6: Histogram of response times to all questions. The x-axis is presented up to 180 seconds as participants taking longer than this length of time in any question were removed.

Table 3: Accuracies of each graph type for each question.

INACCURATE-ACCURATE

Question	Bar	Pie	Table
D2Q1	74.89	76.17	82.13
D2Q2	60.41	8.93	92.00
D2Q3	90.89	18.25	96.51
D2Q4	45.41	4.44	90.99
D2Q5	50.24	49.64	44.19

Table 4: Mean response times in seconds of each graph type for each question.

SLOW-FAST

Question	Bar	Pie	Table
D2Q1	19.51	23.75	20.79
D2Q2	29.65	35.85	27.95
D2Q3	16.65	21.16	16.12
D2Q4	20.54	19.66	23.22
D2Q5	10.55	9.41	8.61

Table 5: Percentage of responses to visual appeal questions by graph type.

LOW-HIGH

Dataset	Graph Type	Very Unappealing	Unappealing	Neutral	Appealing	Very Appealing
2	Bar Coloured	7.02	21.49	36.84	28.51	6.14
	Bar Black & White	14.35	40.43	24.78	16.52	3.91
	Pie Coloured	18.48	36.49	25.59	13.27	6.16
	Pie Black & White	36.97	34.03	19.33	8.40	1.26
	Table	17.87	33.19	34.89	9.36	4.68

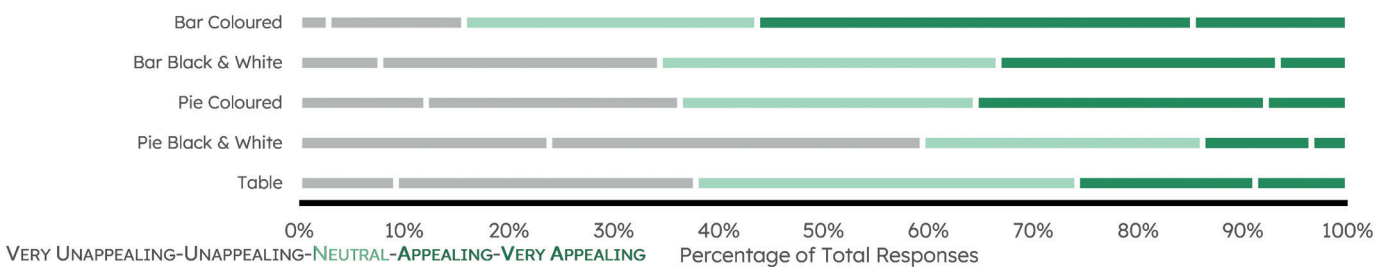


Figure 7: Mean percentage of responses to visual appeal questions by graph type over all datasets.

Combining this data with the significances discussed above reveals whether a significant difference is really meaningful in the real world. For this purpose, we regard that a difference is meaningful and impactful if: for accuracies and percentage responses the values more than 10% apart; and for response times if the values are more than 5 seconds apart. We believe that these differences are enough to be noticeable in real world situations, and for accuracies, larger than the natural variance typically observed within real datasets. Consequently, our study shows that:

1. There is not a meaningful difference in accuracies or response times of placement questions between any graph types.
2. There is a meaningful difference in accuracies of all numerical question types between all graph types, with the table being more accurate than the bar, and the pie being least accurate.
3. There is no meaningful difference in pattern question response times.
4. The pie charts were meaningfully the slowest at difference tasks and meaningfully slower than only the table in valuation tasks.
5. No meaningful difference is found in summation question response times.
6. The coloured bar chart was the most appealing, the black & white pie chart the least appealing, and the remaining three visualisations having the same perceived aesthetic

5.2. WHICH VISUALISATION TYPE SHOULD I USE?

Based on our findings, we suggest the following visualisation usage dependent on task:

- **Use tables for exact numerical calculations and finding exact values.** We found that tables produced the most accurate results and quickest response times for these tasks.
- **Tables are often more efficient on smaller datasets (for example up to 10-20 categories)**. Graphs are likely only useful when specifically asking questions around patterns or part-whole relationships.
- **Use bar charts for identifying patterns.** Our results show that the bar charts outperformed pie charts and tables in this area.
- **All else being equal, and if visual aesthetic is an important factor in selecting your visualisation, use a bar chart with colour over a pie chart of any colour.** Results of our study show that the coloured bar charts significantly outperformed the other visualisations tested in the visual appeal sector and that the black & white pie chart was least preferred visually.
- **If 7 or more datapoints are required do not use a pie chart.** We find that the bar chart and table were always significantly more accurate than the pie chart in all our datasets (containing between 7-12 datapoints). This finding is contrary to previous research where little or no significance is found (also where often no more than 5 different datapoints are presented).

5.3. ASSESSMENT OF HYPOTHESES

Based upon our systematic analysis of the survey responses, we are able to evaluate the hypotheses that we made in Section 2 as follows:

- When participants were asked to compare the difference of two categories, tables did outperform both the bar and pie in terms of accuracy and response times, and bar charts are also faster and more accurate than pie charts.
- When asked to sum two categories, we actually found that the pie chart is faster than the table and bar chart, with the table second fastest. However, we were correct in our hypothesis that the table would be the most accurate and pie chart the least accurate.
- When asked to determine whether a discernible pattern was present in the dataset, the bar charts are the fastest and most accurate, but we were wrong in our hypothesis that the pie charts would be significantly more efficient than the tables, instead finding that the pie chart and table are equally accurate.
- We did find that the coloured bar chart is the most appealing, but the pie chart is not on par. We are correct that the coloured versions of the bar and pie charts are more visually appealing than their black & white counterparts. However, we found that the table is not the least aesthetic but instead the black & white pie chart is.
- Contrasting with our hypothesis, we find that the difference in accuracy between bar and pie charts is larger than the natural variance typically observed in real-world datasets for numerical questions.

6. CONCLUSIONS

Since the visual appeal of graphical visualisations is not widely investigated as it is often grouped as a largely insignificant factor of "graphical preference" and other such variables, we conducted a systematic analysis of online voluntary participated survey to explore its impact on data visualisation efficiency and found that the table had the greatest performance in terms of accuracy and speed of response for both categorical and numerical tasks, followed by bar charts. The bar charts resulted in the strongest pattern identification ability and the bar charts with colour were overwhelmingly the most visually appealing. we also found that there was no effect of colour on accuracy or time of response apart from where it is a relevant identification tool.

7. REFERENCES

Benbasat, I., Dexter, A. S., & Todd, P. (1986). The Influence of Color and Graphical Information Presentation in a Managerial Decision Simulation. Human–Computer Interaction, 2(1), 65–92. https://doi.org/10.1207/s15327051hci0201_3.

Bishop, I. D., Pettit, C. J., Sheth, F., & Sharma, S. (2013). Evaluation of Data Visualisation Options for Land-Use Policy and Decision Making in Response to Climate Change. Environment and Planning B: Planning and Design, 40(2), 213–233. https://doi.org/10.1068/b38159.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. Journal of Experimental Psychology: Applied, 4(2), 75–100. https://doi.org/10.1037/1076-898x.4.2.75.

Cawthon, N., & Moere, A. V. (2007). The Effect of Aesthetic on the Usability of Data Visualization. 2007 11th International Conference Information Visualization (IV '07). https://doi.org/10.1109/iv.2007.147.

Chen, Z., Wang, Y., Wang, Q., Wang, Y., & Qu, H. (2020). Towards Automated Infographic Design: Deep Learning-based Auto-Extraction of Extensible Timeline. IEEE Transactions on Visualization and Computer Graphics, 26(1), 917–926. https://doi.org/10.1109/TVCG.2019.2934810.

Chertov, O., Komarov, A., Andrienko, G., Andrienko, N., & Gatalsky, P. (2002). Integrating forest simulation models and spatial–temporal interactive visualisation for decision making at landscape level. Ecological Modelling, 148(1), 47–65. https://doi.org/10.1016/s0304-3800(01)00437-9.

Chinnaswamy, A., Papa, A., Dezi, L., & Mattiacci, A. (2019). Big data visualisation, geographic information systems and decision making in healthcare management. Management Decision, 57(8), 1937–1959. https://doi.org/10.1108/md-07-2018-0835.

Cohen, P. N. (2012). Children's Gender and Parents' Color Preferences. Archives of Sexual Behavior, 42(3), 393–397. https://doi.org/10.1007/s10508-012-9951-5.

Cleveland, W. S.. Visualizing Data. Hobart Press, 1993.

Cleveland, W. S., Diaconis, P., & McGill, R. (1982). Variables on Scatterplots Look More Highly Correlated When the Scales Are Increased. Science, 216(4550), 1138–1141. https://doi.org/10.1126/science.216.4550.1138.

Cleveland, W. S., & McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. Journal of the American Statistical Association, 79(387), 531–554. https://doi.org/10.1080/01621459.1984.10478080.

W. S. Cleveland. J. Am. Statistical Assoc., 79:531–554, 1984.(1985), "Graphical Perception and Graphical Methods for Analyzing and Presenting Scientific Data," Science, 229, 828-833.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making Sense of Graphs: Critical Factors Influencing Comprehension and Instructional Implications. Journal for Research in Mathematics Education, 32(2), 124–158. https://doi.org/10.2307/749671.

Great Britain: Grocery market share 2015-2020. (n.d.). Statista. https://www.statista.com/statistics/280208/grocery-market-share-in-the-united-kingdom-uk/#statisticContainer.

Golbeck, J. (2015). Benford's Law Applies to Online Social Networks. PLOS ONE, 10(8), e0135169. https://doi.org/10.1371/journal.pone.0135169.

Hegarty, M. (2011). The Cognitive Science of Visual-Spatial Displays: Implications for Design. Topics in Cognitive Science, 3(3), 446–474. https://doi.org/10.1111/j.1756-8765.2011.01150.x.

Knaflic, C. N., (2015). Storytelling with Data A Data Visualization Guide for Business Professionals. Hoboken, Nj, Usa John Wiley & Sons, Inc.

Lohse, G. L. (1993). A Cognitive Model for Understanding Graphical Perception. Human–Computer Interaction, 8(4), 353–388. https://doi.org/10.1207/s15327051hci0804_3.

Okan, Y., Galesic, M., & Garcia-Retamero, R. (2015). How People with Low and High Graph Literacy Process Health Graphs: Evidence from Eye-tracking. Journal of Behavioral Decision Making, 29(2-3), 271–294. https://doi.org/10.1002/bdm.1891.

Peck, E. M. M., Yuksel, B. F., Ottley, A., Jacob, R. J. K., & Chang, R. (2013). Using fNIRS brain sensing to evaluate information visualization interfaces. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/2470654.2470723.

Pinker, S. (1988). Visual cognition. Mit Press.

Quattrone, P. (2017). Embracing ambiguity in management controls and decision-making processes: On how to design data visualisations to prompt wise judgement. Accounting and Business Research, 47(5), 588–612. https://doi.org/10.1080/00014788.2017.1320842.

Savva, M., Kong, N., Chhajta, A., Fei-Fei, L., Agrawala, M., & Heer, J. (2011). ReVision: automated classification, analysis and redesign of chart images. Proceedings of the 24th annual ACM symposium on User interface software and technology.

Schunk, D. (2008). A Markov chain Monte Carlo algorithm for multiple imputation in large surveys. AStA Advances in Statistical Analysis, 92(1), 101–114. https://doi.org/10.1007/s10182-008-0053-6.

Siblis Research. (2020, October 6). FTSE 100 Index Sector Weightings. Siblis Research. https://siblisresearch.com/data/ftse-100-sector-weights/.

Simkin, D., & Hastie, R. (1987). An Information-Processing Analysis of Graph Perception. Journal of the American Statistical Association, 82(398), 454–465. https://doi.org/10.1080/01621459.1987.10478448

Skau, D., & Kosara, R. (2016a). Arcs, Angles, or Areas: Individual Data Encodings in Pie and Donut Charts. Computer Graphics Forum, 35(3), 121–130. https://doi.org/10.1111/cgf.12888.

Skau, D. & Kosara, R. (2016b). Judgment Error in Pie Chart Variations. Eurographics Conference on Visualization – EuroVis 2016.

Spence, I. (1990). Visual psychophysics of simple graphical elements. Journal of Experimental Psychology: Human Perception and Performance, 16(4), 683–692. https://doi.org/10.1037/0096-1523.16.4.683.

Statista. (2021, June 7). Data Created Worldwide 2010-2025 | Statista. Statista; Statista. https://www.statista.com/statistics/871513/worldwide-data-created/

Strobel, B., Saß, S., Lindner, M. A., & Köller, O. (2016). Do Graph Readers Prefer the Graph Type Most Suited to a Given Task? Insights from Eye Tracking. Journal of Eye Movement Research, 9(4). https://doi.org/10.16910/jemr.9.4.4.

Yigitbasioglu, O. M., & Velcu, O. (2012). A review of dashboards in performance management: Implications for design and research. International Journal of Accounting Information Systems, 13(1), 41–59. https://doi.org/10.1016/j.accinf.2011.08.002.

Zacks, J., & Tversky, B. (1999). Bars and lines: A study of graphic communication. Memory & Cognition, 27(6), 1073–1079. https://doi.org/10.3758/bf03201236.

A long time ago in a galaxy far, far away: The Hunt for the Supermassive Black Hole

An extract of this short-listed Independent Learning Assignment (ILA)

Andrew Zhang, Upper Sixth

"Black holes are where God divided by zero" - Albert Einstein

1 INTRODUCTION

Black holes are a fundamental prediction of Einstein's theory of General Relativity [1]. A defining feature of black holes is the event horizon, a one-way causal boundary in space-time out of which not even light can escape; at first studied as a mathematical consequence of General Relativity rather than physically relevant objects [2], they are now known as generic and often inevitable results of gravitational collapse [3][4]. Over a century after the discoveries of Einstein and Schwarzschild, they still remain at the heart of fundamental questions unifying General Relativity with quantum physics [5][6].

Black holes are fairly common in astrophysics, and found over a range of masses (from stellar mass black holes to supermassive black holes). Evidence for stellar mass black holes come from X-ray [7] and gravitational wave measurements [8]. Supermassive black holes, with masses from millions to tens of billions of solar masses, are thought to exist in the centre of nearly all galaxies [9], including in the centre of the milky way [10] and in the nucleus of the nearby elliptical galaxy M87 [11][12].

In 2019, the first ever image of a black hole was released by the Event Horizon Telescope Collaboration [12], and only a few months ago, in March of this year, the first image of Sagittarius A*, the black hole at the centre of our own galaxy, was released [13].

The earliest mention of the concept Black Hole comes from John Michell, the English natural philosopher and clergyman, writing in the 1770s, and producing a paper on the topic in 1783. While arguing that the brightness of a star was related to its mass and area (and hence density), he calculated that the escape velocity of the sun was roughly 497 times smaller than the velocity of light (which was known fairly accurately [14]), and hence argued that a star with 497 times the area and the same density of the sun would trap all the light that it emitted. He went so far as to point out that it would be impossible to observe these so-called "dark stars" except by their gravitational effects on nearby bodies.

This idea relied on the fact that light would be slowed down, just like any other macroscopic body, by the gravitational field of the star from which it was emitted; however, when William Herschel later attempted to measure a decrease in the speed of detected light, he was unsuccessful. [15]

2 RELATIVITY AND THE BLACK HOLE

2.1 EINSTEIN'S FIELD EQUATIONS

2.1.1 EINSTEIN'S EQUATION AND SCHWARZSCHILD'S SOLUTION

It wasn't until over 100 years after they had first been dreamt up by John Michell, in 1915, that black holes would once again become a topic of serious scientific discussion. Having developed his theory of General Relativity, Einstein showed that gravity, rather than being a simple force as in Newtonian physics, was a distortion of spacetime and as such affected the path of light.

Einstein's field equations described the local curvature of spacetime of a region, relating it to the local energy, momentum and stress within that spacetime [16]. Only a few months after their publication, (and while serving as a soldier in the First World War!) German scientist Karl Schwarzschild found an "exact solution" to these equations, describing the curvature of space-time around stationary, spherical objects (for example, a star). However, Schwarzschild's solution seemed to show that a mass gathered in a small enough space would cause the equation to "blow up to infinity" [17]; terms in the field equations would become infinite. This is known as the Schwarzschild radius:

$$R_s = \frac{2GM}{c^2} \qquad (1)$$

where G is the universal gravitational constant, M is the mass of the object, and c is the speed of light. This describes the radius of the event horizon of a non-rotating black hole; any object with a radius lower than its Schwarzschild radius is a black hole [18].

At the time, Einstein and Schwarzschild did not really understand what they had found, and Schwarzschild died just months later. Despite the mathematical foundations for what would become known as the black hole being set, the concept would be overlooked for two decades.

3 IMAGING

Obtaining a close-up view of a black hole has been a long-term goal that has long evaded astronomers. In 2019, the first ever image of a black hole was published by the Event Horizon Telescope collaboration - an image of the supermassive black hole at the center of the galaxy M87. [19]

The plan was to image the event horizon of a supermassive black hole. Two sources were initially selected as targets: Sgr A*, the supermassive black hole at the centre of our own galaxy, and the huge, jet-producing black hole of one of our neighbour galaxies, M87.

3.1 THE EVENT HORIZON TELESCOPE

3.1.1 A PLANET-SCALE TELESCOPE

The Event Horizon Telescope is an extensive virtual telescope which is created by combining simultaneous observations from various radio arrays and dishes. In the case of M87, the observations were made by eight ground-based telescopes in Arizona, Hawaii, Mexico, Chile, Spain and the South Pole.

The EHT works by performing very-long-baseline interferometry: combining different telescopes around the world allows the EHT to function like a telescope with an effective size the same size as its longest baseline (the distance between

component telescopes). These are shown in Figure 1, where the solid lines are baselines used to observe-M87 and the dashed lines are baselines used the calibration source. The resolving power of the telescope can be calculated by the Rayleigh criterion, $\theta \approx \frac{\lambda}{b}$ where θ is the minimum angle that can be resolved, and b is the aperture width. This can be adapted to the EHT by using the longest baseline (distance between two telescopes) as b. Thus, it can reach an incredible resolution of 25μas at its observing wavelength of 1.3mm. [20] [12]

The EHT's imaging plans begun over a decade ago, but facilities had to be upgraded and new facilities built.

3.1.2 THE PHOTON SPHERE

Although the Schwarzschild radius of black hole defines its event horizon (if it is non-spinning), the closest to the black hole that the EHT can actually observe is the photon sphere (i.e. where photons orbit in the black hole in unstable circular orbits) [20]. This can be found in terms of its gravitational radius: Gravitational lensing means that the radius of the sphere appears between $(2\sqrt{3+2\sqrt{2}})r_g$ and $(3\sqrt{(3)})^{r_g}$ (for a Swarzschild BH), which depends on the spin and inclination of the BH (where rg is the gravitational radius, $r_g = \frac{GM}{c^2}$ or $\frac{r_s}{2}$. The photon sphere appears as the "shadow" of the black hole.

Consider a non-spinning black hole with a mass of ~ $4.1\times10^6 M_\odot$ and a distance of 8.34 kpc from earth (modeling Sgr A* as a Schwarzschild BH): by balancing the relativistic centripetal (where $v' = v\sqrt{1-\frac{2r_g}{r}}$, where $r = r_{ph}$ and $v = c$) and Newtonian gravitational forces we can find the radius of the photon sphere as seen from earth:

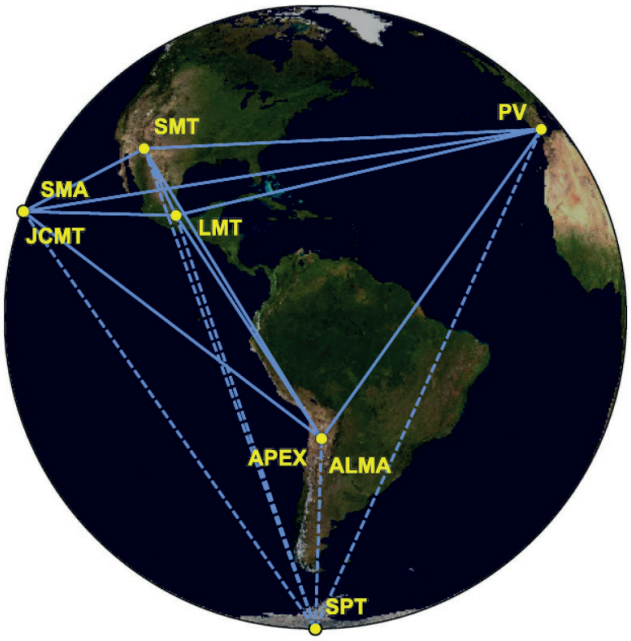


Figure 1: The eight EHT 2017 campaign stations over six geographic locations, as viewed from the equatorial plane. The solid planes represent mutual visibility on M87*. The dashed lines were used for the calibration source, 3C 279. Source: EHT Collaboration et al 2019 [12]

$$\frac{GMm}{r^2} = \frac{mv'^2}{r}$$

$$v'^2 = v^2\left(1 - \frac{2r_g}{r}\right) = \frac{GM}{r}$$

Now, substituting v for c (the speed of light, which the photons in the photon sphere must move at), r_g for $\frac{GM}{c^2}$ and r for r_{ph}:

$$\frac{GM}{r_{ph}} = c^2\left(1 - \frac{2r_g}{r_{ph}}\right)$$

$$\frac{GM}{c^2} = r_{ph}\left(1 - \frac{2r_g}{r_{ph}}\right)$$

$$r_g = r_{ph}\left(1 - \frac{2r_g}{r_{ph}}\right)$$

rearranging for r_{ph}:

$$r_{ph} = 3\ r_g \qquad (2)$$

Thus the photon distance is three times the size of the gravitational radius for a model of Sgr A* as a Schwarzschild black hole (non-spinning), and we can use trigonometry to find the angular diameter of this model (about 25 μas). Therefore it can be resolved by the EHT. [20] [21]

3.2 M87

3.2.1 WHY M87?

Although M87 is further away (about 2,000 times further), its black hole is around 1,500 times larger than that of Sgr A* (as well as being far more luminous), meaning that it has longer and more manageable variability timescales. Therefore M87 is comparatively straightforward to image compared to Sgr A*.

M87* is known to be active, having a "jet" of particles that shoots out and stretches 5000 light years from the galaxy's centre, while Sgr A* is far calmer (M87* feeds on massive reservoirs of gas, while Sgr A* sips from the stellar winds of a few dozen stars in its neighbourhood).

3.2.2 ESTIMATING MASS

An upper and lower limit for the mass of M87* can be estimated by using its angular diameter (determined from the images gained by the EHT as 42μas) and distance from us (16.8 Mpc, or ~ 5.18×10^{23}m to find its lensed physical radius, and then using the minimum and maximum of gravitational lensing that occurs to find the minimum and maximum gravitational radius and thus its mass [20]. The following was adapted from the 2020 BAAO paper:

We know that the angular radius is half of the angular diameter (θ_r = 21 μas or 1.02×10^{-10} radians. Using the small angle approximation and some trigonometry we can say that:

$$\theta_r = \frac{r'_g}{d}$$

$$r'_g = d\theta_r$$

$$r'_g = 1.02 \times 10^{-10} \times 5.18 \times 10^{23} = 5.29 \times 10^{13}\text{m}$$

Where r'_g is the lensed gravitational radius and d is the distance to M87*. From this the maximum and minimum bounds of M87*'s mass can be found by applying the two extremes of gravitational lensing to find the possible extremes of r_g, the actual radius and hence its mass:

$$r'_g = (3\sqrt{3})\, r_{g,min}$$

$$r_{g,min} = \frac{r'_g}{3\sqrt{3}} = \frac{5.29 \times 10^{13}}{3\sqrt{3}} = 1.02 \times 10^{13}\text{m}$$

but since $r_g = \frac{GM}{c^2}$:

$$r_{g,min} = 1.02 \times 10^{13} = \frac{GM_{min}}{c^2}$$

$$M_{min} = \frac{r_{g,min}\, c^2}{G} = \frac{1.02 \times 10^{13} c^2}{G} = 1.37 \times 10^{40}\text{kg}$$

Which is ~ $6.90 \times 10^9 M_\odot$. The same method can be used, but with $r_{g,max} = (2\sqrt{3}+2\sqrt{2})r_g$ to get $M_{max} = 1.48 \times 10^{40}$ kg or $7.41 \times 10^9 M_\odot$ and therefore the mass M87* can be estimated from its images. [20] [21]

3.2.3 IMAGING M87

In April 2017, conditions were finally right to attempt to image M87's event horizon. It was observed for four days (across the planet) and the data from all eight of the telescopes were combined in order to reconstruct images of the black hole.

The image matched predictions incredibly well, with a ring of light spanning 38 – 44μas, and the "southern" part of the ring appearing brighter than the rest.

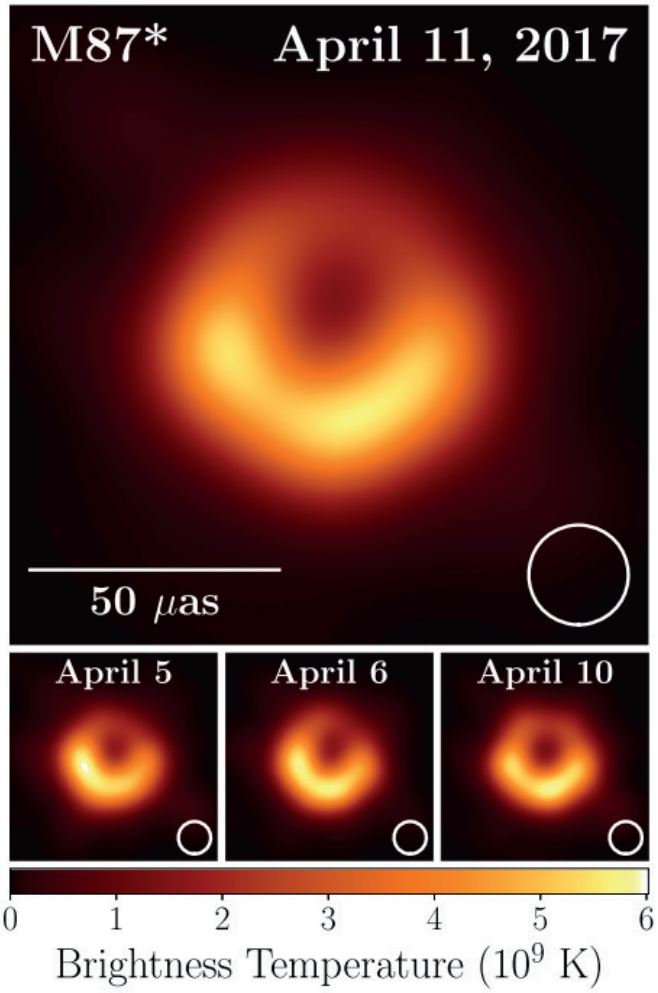


Figure 2: Top: EHT image of M87* from observations on the 11th April as a representative sample of the images collected. The image is the average of three different imaging methods, after convolving each with a circular Gaussian kernel to give matched resolutions. The image is shown in units of brightness temperature: $\mathrm{Tb} = \frac{S\lambda^2}{2k_B\Omega}$ where S is the flux density, λ is the observing wavelength, k_B is the Boltzmann constant and Ω is the solid angle of the resolution element. Bottom: similar images which were taken over different days, showing the stability of the basic image structure and equivalence on different days. North is up and East is left. Source: [12]

The bright orange ring of emission is caused by the dust and gas that forms the accretion disk of the black hole, as well as funnels at the base of its powerful jets. It was predicted that images of such an encased black hole would reveal a dark region - the black hole's "shadow", surrounded by the ring of emission which is produced by distorted paths of light from the surrounding material.

The southern part of the ring appears brighter because the matter moves towards us on this side; a relativistic effect beams light in our direction and causes the region to appear more bright. Combined with observations from previous measurements of M87's jet, which show that it is inclined at an angle of 17° relative to our line of sight, so that M87 likely spins clockwise from our point our view, with its spin axis pointed at an angle away from us. [12]

3.2.4 NON-SPINNING AND SPINNING ACCRETION DISK PREDICTIONS

The theoretical inner radius of the ring, the Innermost Stable Circular Orbit (ISCO), can be found by the value of r for which the total conserved energy of a circular orbit close to a Schwarzschild BH which minimises this energy, E:

$$E = mc^2 \left(\frac{1 - \frac{2r_g}{r}}{\sqrt{1 - \frac{3r_g}{r}}}\right) \qquad (3)$$

This can be done is by differentiating E with respect to r and setting the derivative equal to zero, using product rule:

$$\frac{dE}{dr} = mc^2 \left[\frac{2r_g}{r^2\sqrt{1 - \frac{3r_g}{r}}} - \frac{3r_g\left(1 - \frac{2r_g}{r}\right)}{2r^2\left(1 - \frac{3r_g}{r}\right)^{\frac{3}{2}}}\right] = 0 \qquad (4)$$

thus:

$$\frac{2r_g}{r^2\sqrt{1 - \frac{3r_g}{r}}} - \frac{3r_g\left(1 - \frac{2r_g}{r}\right)}{2r^2\left(1 - \frac{3r_g}{r}\right)^{\frac{3}{2}}} = 0$$

$$2\left(1 - \frac{3r_g}{r}\right) = \frac{3}{2}\left(1 - \frac{2r_g}{r}\right)$$

$$r = 6r_g$$

Therefore $r_{ISCO} = 6r_g$. This is analogous to the photon sphere, but for particles with mass. [20] [21]

However, this is true only for a non-rotating BHs: in reality many BHs rotate and have angular momentum. The Kerr metric describes black holes with spin angular momentum: the photon capture radius changes with the ray's orientation relative to the vector of angular momentum and therefore the black hole's cross section is not necessarily circular [22].

The spin can be quantified with a dimensionless spin parameter, $a \equiv J/J_{max}$ (where J is angular momentum of the BH and J_{max} is it maximum possible angular momentum, defined by $J_{max} = \frac{GM^2}{c}$. The spin parameter varies from −1 ≤ a ≤ 1; negative spins indicate black hole rotation in the opposite direction of accretion disk rotation and vice versa. When $a = 1$, $r_{ISCO} = r_g$ and when $a = -1$, $r_{ISCO} = 9r_g$. An equation for the angular velocity of a particle in the ISCO is:

$$\omega^2 = \frac{GM}{(r_{ISCO}^{3/2} + ar_g^{3/2})^2} \qquad (5)$$

Thus, if we know the mass of M87* as $6.5 \times 10^9 M_\odot$ we can find the time period for a particle in the ISCO of M87* for each of the cases of $a = 1$, $a = -1$ and $a = 0$ by first using $r_g = \frac{GM}{c^2}$ (to get $r_g = 9.59 \times 10^{12}$m) and then the relation between angular velocity and time period, T: $\omega = \frac{2\pi}{T}$. Thus for $a = 1$

$$\omega^2 = \frac{GM}{c^2}$$

$$T^2 = \frac{4\pi^2 c^2}{GM}$$

$$T = 4.02 \times 10^5 \text{s}$$

Or an orbital time period of 4.65 days. Repeating the same method for a = 0 and a = −1 gives time periods of 34.2 days and 60.4 days respectively.

During the imaging of M87*, over a period of 5 days, a bright patch of gas was observed to move a quarter of the way ar ound the ring. Therefore, the orbital period of this gas would be 20 days (which falls within the range of ISCO gas) and corresponds to a positive spin. [20] [21]

3.2.5 COMPARING M87* WITH MODELS

Comparing the images of M87* to an extensive library of simulated observations, which are dependent on the underlying models and conditions, showed incredible consistency with actual observations (confirming a picture of a spinning supermassive black hole as the source of the emission). They also suggest the most likely way that M87's jets are launched: via energy extraction directly from the black hole's spin [19].

Previous estimates of the mass of M87* ranged from ~ 3–7×$10^9 M_\odot$, depending on measurement methods. From these images, the mass has been estimated to be (6.5 ± 0.7) × $10^9 M_\odot$.

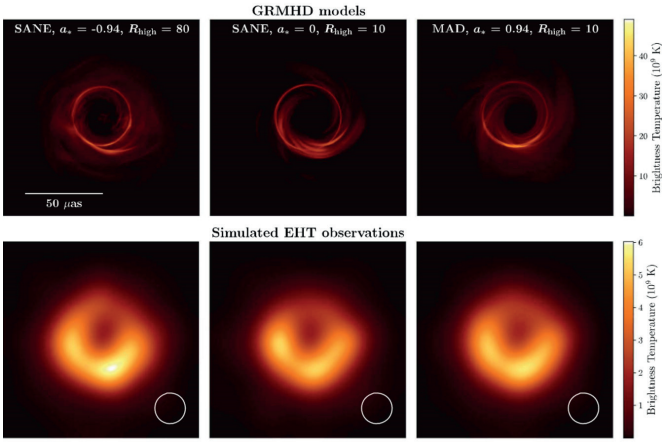


Figure 3: Top: three example models of some best-fitting snapshots from the image library of simulations, each corresponding to different spin parameters and accretion flows. Bottom: the same theoretical models, processed through a simulation pipeline with the same schedule, telescope characteristics and weather parameters as on 11th April. Note that although they all fit the image equally well, they each refer to radically different physical scenarios, highlighting the fact that a single good fit doesn't imply that a model is preferred over others. Source: [12]

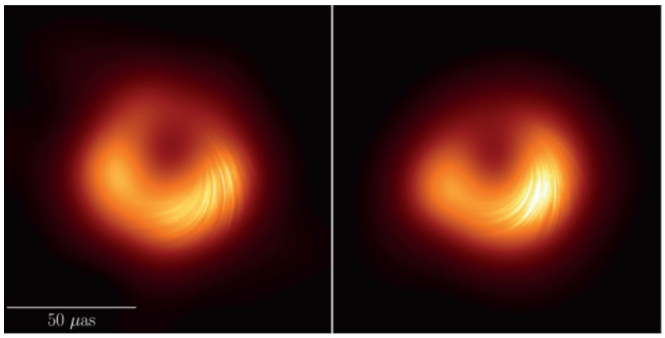


Figure 4: The polarized image of the M87 black hole shadow as observed on 11th April 2017 by the EHT (left panel). Right: an image from the EHT model Library with a MAD magnetic configuration (magnetially arrested disks; where the magnetic fields are the dominant force regulating the accretion onto the black hole). Source: source: [12]

Further analysis has been carried out: polarimetric analysis of these images, which probe the magnetic field and help determine the rate of accretion into M87's black hole.

3.2.6 SPACE-TIME NEAR A BLACK HOLE

General Relativity tells us about the curved spacetime near a black hole, resulting in the distance between two points being very different from the same situation in flat spacetime. Taking M87* as an example, a particle has to travel much further to fall into the black hole (going from rISCO to rph) than expected in flat spacetime [20]. The equation for the Relativistic length is:

$$\Delta l = \int_{r_2}^{r_1} (1 - \frac{2r_g}{r})^{-\frac{1}{2}} dr \qquad (6)$$

And since we know that it must travel from r_{ISCO} to r_{pb}, and that these are $6r_g$ and $3r_g$ respectively we integrate with these limits to get:

$$\Delta l = 4.14 r_g \qquad (7)$$

Comparing this to the distance through flat space-time ($r_{ISCO} - r_{pb} = 3r_g$), a photon must travel 1.1 × 10^{13}m more in comparison to flat spacetime! [20] [21]

3.2.7 LOOKING TO THE FUTURE

Future analysis of EHT observations will delve into the magnetic field near Sgr A* and in M87*, as well as investigating structural changes that may have accompanied a flare that was seen in the X-ray emission on the final day of EHT observations. Although the existing images of both black holes promise many future discoveries, even more exciting projects are down the pipeline. More detectors and recording bandwith had been added to the EHT since 2017, greatly enhancing the capabilities of the planet-scale telescope. The first video of hot gas circling a black hole may lie in our not-too-distant future! [13]

4 CONCLUDING REMARKS

The characteristics and formation processes of black holes are integral to our understanding of the universe and, from dark stars to general relativity and quantum physics, the journey to understand them has a long way to go.

The images of M87* released in 2019 helped us to determine a black hole's mass, spin and orientation, providing a valuable test of how massive objects bend spacetime. The EHT results for SgrA* and M87*, when paired with gravitational wave measurements from colliding stellar-mass black holes,

validate Einstein's general theory of relativity over an impressive range of masses and distances. [23], and, though the EHT's results give definitive answers to many questions about the black hole at the centre of our galaxy, many remain.

How do supermassive black holes create their particle jets? Do they hinder star formation in the galaxies they inhabit, or spur it onward? Exactly how, and when, did they form in the first place?

One thing that is clear, however, is that there has never been a better time in human history to observe the Universe - in all its beauty.

LIST OF FIGURES

REFERENCES

[1] A. Einstein, "Die Feldgleichungen der Gravitation", Sitzungsberichte der K¨oniglich Preußischen Akademie der Wissenschaften (Berlin, 844–847 (1915).

[2] K. Schwarzschild, "¨Uber das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie", Sitzungsberichte der K¨oniglich Preußischen Akademie der Wissenschaften (Berlin, 189–196 (1916).

[3] R. Oppenheimer and H. Snyder, "On continued gravitational contraction", Physical Review **56**, 455–459 (1939).

[4] R. Penrose, "Gravitational Collapse and Space-Time Singularities", **14**, 57–59 (1965).

[5] S. W. Hawking, "Breakdown of predictability in gravitational collapse", Physical Review D **14**, 2460–2473 (1976).

[6] S. B. Giddings, "Astronomical tests for quantum black hole structure", 10.48550/ ARXIV.1703.03387 (2017).

[7] R. A. Remillard and J. E. McClintock, "X-Ray Properties of Black-Hole Binaries", **44**, 49–92 (2006).

[8] The LIGO Scientific Collaboration and The Virgo Collaboration, "Observation of gravitational waves from a binary black hole merger", 10.48550/ ARXIV.1602. 03837 (2016).

[9] D. Lynden-Bell, "Galactic Nuclei as Collapsed Old Quasars", 223, 690–694 (1969).

[10] A. Eckart and R. Genzel, "Stellar proper motions in the central 0.1 PC of the Galaxy", **284**, 576–598 (1997).

[11] K. Gebhardt, J. Adams, D. Richstone, T. R. Lauer, S. M. Faber, K. G¨ultekin, J. Murphy, and S. Tremaine, "The Black Hole Mass in M87 from Gemini/NIFS Adaptive Optics Observations", ***729***, 119, 119 (2011).

[12] Event Horizon Telescope Collaboration et al., "First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole", **875**, L1, L1 (2019).

[13] E. C. et al., "First Sagittarius A* Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole in the Center of the Milky Way", **930**, L12, L12 (2022).

[14] D. Raynaud, "Determining the speed of light (1676-1983): an internalist study in the sociology of science", L'Ann´ee sociologique **63** (2013).

[15] S. Schaffer, "John michell and black holes", Journal for the History of Astronomy **10**, 42–43 (1979).

[16] A. Einstein, "The foundation of the general theory of relativity", Annalen der Physik (1916).

[17] A. Levy, "How black holes morphed from theory to reality", Knowable Magazine (2021).

[18] T. E. o. E. Britannica, "Schwarzschild radius", Encyclopedia Britannica (2021).

[19] S. Kohler, *First Images of a Black Hole from the Event Horizon Telescope*, AAS Nova Highlight, 10 Apr 2019, id.5013, Apr. 2019.

[20] A. Calverley, A. Sun, and N. Neog, *British astronomy and astrophysics olympiad 2019-2020*, Jan. 2020.

[21] A. Calverley, A. Sun, and N. Neog, *British astronomy and astrophysics olympiad 2019-2020 solutions and marking guidelines*, Jan. 2020.

[22] R. P. Kerr, "Gravitational field of a spinning mass as an example of algebraically special metrics", Physical Review Letters 11, 237–238 (1963).

[23] K. Hensley, *First Image of the Milky Way's Supermassive Black Hole*, AAS Nova Highlight, 12 May 2022, id.9419, May 2022.

www.rgs-guildford.co.uk

@RGSGuildford

RGS Guildford

Registered Charity No. 312028