DEMOS

OVER THE CHARACTER LIMIT

JOSH SMITH
AGNES CHAUVET
ELLIOT JONES
AVA BERRY

DECEMBER 2019

This project was supported by the Jubilee Centre for Character and Virtues at the University of Birmingham:

Published by Demos December 2019

76 Vincent Square, London, SW1P 2PD
T: 020 3878 3955
hello@demos.co.uk
www.demos.co.uk
Charity number 1042046

CONTENTS

ACKNOWLEDGEMENTS

This report would not have been possible without the generous support of the Jubilee Centre for Character and Virtues at the University of Birmingham. In particular, thanks go to Aidan Thompson, whose enthusiasm, patience and invaluable advice on virtue were critical to the project's development.

At Demos, this research owes a great debt to the multitude of researchers, assistants, proofreaders and designers who have helped bring it to fruition over the last year. Particular thanks go to Agnes Chauvet, Elliot Jones, Stanley Phillipson Brown and Izzy Little for their vital contributions. Thanks are also owed to the team at Sussex, who kept the servers running, and to Professor David Weir for his valuable input into, and advice around, training classifiers.

As ever, thanks are owedto the CASM team for their support– Carl, Alex and Ellen. To Pierre Ratinaud, whose excellent software produced some of the visualisations in this project, and to Oliver Marsh, who introduced me to the software and helped translate documentation. It goes without saying, but all mistakes, errors and omissions included in this paper are my own.

Josh Smith

November 2019

EXECUTIVE SUMMARY

"...there is a positive relationship between a person's use of virtue language on Twitter and their propensity to conduct virtuous actions on the platform."

Over the past few decades the internet has played an increasingly important role in shaping our societies and the individuals within them. We still know surprisingly little, however, about how these online spaces are used by people to explore and enact the moral virtues vital to the development of good character. This report aims to address this gap in our knowledge, providing an evidence base for the ways in which people discuss moral virtues and perform virtuous actions on Twitter.

In this study, conducted by Demos in partnership with the Jubilee Centre for Character and Virtues at the University of Birmingham, we analyse over 1 million tweets sent from the UK which use the terms 'courage', 'empathy', 'honesty' and 'humility', exploring the ways in which these terms are used and defined online. In doing so, we aim to lift the concept of virtue out of an academic, theoretical setting and situating it squarely in the real world, examining the themes, topics and institutions raised within everyday discussions of morality.

We also explore examples of virtuous action carried out online, examining the use of Twitter to express thanks, promote charity and help acquire and perfect skills. We show that this online interaction has positive effects which reach beyond the timeline, helping to raise money for fundraisers and encourage others to apply themselves to learning. Crucially, we also find that there is a positive relationship between a person's use of virtue language on Twitter and their propensity to conduct virtuous actions on the platform.

This research demonstrates how concepts associated with moral virtue are already being used by people in the UK to praise others and hold the powerful to account, and stresses the importance of a sound understanding of virtue-related language in encouraging virtuous action. As numerous past studies have shown, the complex ecosystems of social media have the ability to cause and enable real harm. However, by seeing these online spaces as a proving ground in which character can be positively developed, there is a vital opportunity for educators, policymakers and technology companies to encourage virtuous action and the development of good character.

Key findings

MORAL VIRTUE TERMS ARE WIDELY EMPLOYED BY TWITTER USERS IN THE UK.

During 203 days over late 2018 and early 2019, just over a million tweets were sent from the UK using one of the terms 'courage', 'empathy', 'honesty' and 'humility', with 71% of these using terms in a non-neutral sense – for example, praising or condemning the character of others. Topics discussed alongside virtues range from religion to football, with a focus on politics cutting across all four of the terms in our collection.

THERE IS A POSITIVE RELATIONSHIP BETWEEN THE USE OF VIRTUE TERMS ON TWITTER TO EXPRESS AN OPINION, AND THE PERFORMANCE OF VIRTUOUS ACTION ON THE PLATFORM.

Two statistical models were built which found that there is a statistically significant positive relationship between the use of virtue language and virtuous behaviour. The models predict that, for every 10 virtue language tweets sent by a user, that account will send 5 tweets expressing gratitude, and 1 link to fundraising campaigns. These models explained 10% of the variance in gratitude and 5% in fundraising.

DISCUSSION OF MORALITY ON TWITTER DIFFERS MARKEDLY FROM ITS USE IN OTHER PUBLIC SPHERES.

We compared language used across three spaces - the UK Twitter population, parliamentary speeches, and broadcasts on the BBC. We found that discussions of empathy take a larger role on Twitter, and courage a smaller role, than either of the other two. We also found that, overall, Parliament tends to use virtue language 10 times as often as Twitter or the BBC.

MORAL VIRTUE TERMS ARE MORE OFTEN USED ON TWITTER TO CRITICISE INSTITUTIONS THAN TO PRAISE THEM.

Around a quarter of tweets using a virtue term (24%) discussed virtue in institutions, including politicians and political parties, the media, public services and large corporations. 48% of these tweets used virtue terms in a negative sense, a figure substantially higher than the 36% of negativity which occurred in virtue tweets in general.

TWITTER IS USED TO ALLOW PEOPLE TO PUBLICLY STATE THEIR OWN, PERSONAL DEFINITON OF MORAL TERMS.

Below, we analyse a sample of 200 tweets which make a public, definitional statement about one of our four virtues. Notable amongst these is a belief that empathy is on the decline in public life and the workplace, which appeared alongside questions about whether empathy is an innate or developed capacity. Users also discussed the courage inherent in acts of writing and questioned whether humility was overrated.

SHARING A LINK TO AN ONLINE FUNDRAISING CAMPAIGN HAS A WEAK BUT POSITIVE EFFECT ON THE AMOUNT THAT CAMPAIGN RAISES.

In an analysis of 7,900 tweets containing a link to a campaign on fundraising site justgiving.com, we found that the number of tweets sent which link to a campaign is positively related to the overall amount of money raised by that campaign. A statistical model was trained which predicts that each link sent is associated with an increase of £56 in the total amount raised by a campaign; though the model explains only 2% of the variance observed in these totals. As an example of social media posts leading to good outcomes, this adds evidence to the theory that sharing links in this way should itself be considered as a form of virtuous action.

INTERACTION ON SOCIAL MEDIA HAS A POSITIVE EFFECT ON PERSEVERENCE IN DEVELOPING A SKILL.

To investigate perseverance and application on Twitter, 7,900 tweets were collected using a hashtag which pledged to practice a skill for a set period – '#100daysofanimation', for example. We found that there is a significant positive relationship between the extent to which peers on Twitter engaged with others committing publicly to these programmes and the length of time for which people continue using them. This early engagement is relatively decisive, counting for 28% of the variance in the dataset. This suggests that interaction on social media, even at the basic level of liking someone's tweet, can provide the support and encouragement they need to persevere. As above, this suggests that certain types of interaction on social media can themselves be virtuous acts.

Recommendations

FOR POLICYMAKERS:

Building character should continue to be a priority for the Department for Education.

This research shows that developing a sense of virtue, and comfort with discussing virtue online, is linked to virtuous actions. The Department for Education should restate its commitment to building character, as discussed in early 2019 by then education secretary Damian Hinds. Furthermore, the Department should reinstate funding cancelled in 2017 to promote character development in schools.

Policies designed to develop character should encourage the development of cyber phronesis.

Developing phronesis, defined as the moral wisdom which allows an individual to determine how to act well in real-life situations, is key to building a practical moral character. It is important that this sense is developed to guide good action online as well as offline. This will require promoting the value of thoughtful reflection on the effect of online actions, informed by a solid evidence base connecting behaviour online to tangible positive outcomes.

FOR EDUCATORS:

The role of online space in positively developing character should be recognised.

This research shows clearly that online spaces are utilised by people as areas which can enable personal flourishing, and where people can support others to flourish, through expressions of gratitude, charity and as an impetus to learn and grow. Where appropriate, existing initiatives to encourage the development of character in students, and to encourage them to discuss and define what virtue means to them, should take online social space into account as a meaningful arena in which virtues might be discussed and practised.

Programmes should encourage virtuous behaviour online, as well as building resilience to perceived or potential harms.

Existing initiatives which address online space, often presented through personal, social, health and economic (PSHE) programmes, tend to take a risk-based approach to social media, educating children on the harms they might encounter, and increasing resilience to exploitation, bullying and disinformation. This work is vital, and Demos has often made the case for its importance.[1] However, painting the online world as a place fraught solely with danger risks undermining the potential for students to develop their moral sensitivities in a positive sense. A framing should be found which enables discussion of good action online, without minimising the dangers faced on social media platforms. The current curriculum brings offline concepts such as personal safety and media literacy into an online space; it must now do the same for civics, citizenship and the development of good character.

FOR SOCIAL MEDIA PLATFORMS:

Virtuous action should be studied across the online ecosystem.

This report analyses virtuous action amongst the UK users of two particular platforms - Twitter and JustGiving. In both cases, this analysis was made possible through data available to researchers through each platform's API. These two spaces, however, afford only a partial picture of the overall online ecosystem. For many platforms, including Instagram, TikTok and other spaces likely to be particularly influential in developing the character of young people, data is currently simply not accessible. As part of a wider culture of transparency, and in order to support the promotion of positive action online, social media platforms should provide access to data for independent research to be conducted into virtuous speech and action online. This would help to fill a vital gap in our understanding of how character develops in the modern world, and how online senses of virtuous speech and action differ from our understanding of them offline.

1 See e.g. Harrison P., Krasodomski, A. (2017) 'The Moral Web: Youth, Character, Ethics and Behaviour', conducted in partnership with the Jubilee Centre for Character and Virtues, and Reynolds L., Scott R. (2016) 'Digital Citizens: Countering Extremism Online'.

INTRODUCTION

The question of how to act virtuously is, perhaps, our single overarching moral challenge. In Aristotle's formulation, this good 'has rightly been declared to be that at which all things aim.'[2] As societies develop, they find ways to code this universal goal into guidelines for right action, from which behavioural norms, systems of justice, and religious doctrine emerge and are formalised. For individuals, a sense of virtue is developed through practice; through the repeated application of both these societal codes and that person's internal sense of what is right.

The advent of the internet has allowed new, globally dispersed forms of online society to develop, with their own distinct senses of what constitutes right action. One of the great challenges posed by social media stems from the fact that we are still developing an agreed set of social codes for how to behave well – or even politely – towards others occupying this digital space. This has contributed to the modern proliferation of abuse and the coarsening of public debate, often with the encouragement of those at the apex of our political structures. It has also allowed new kinds of moral flourishing, facilitating novel means of social connection, charity and reciprocity.

This lack of articulated social codes does not mean people have no sense for how to act well online, and there are actions on social media which 'feel' instinctively right - supporting a campaign, say, or congratulating a friend for an achievement. Developing and applying this instinct for what constitutes virtuous action on social media represents a kind of cyber 'phronesis', a term defined by Aristotle as 'a true and reasoned state of capacity to act with regard to the things which are good or bad for man.'[3] Under this definition, phronesis relies not only on a strong moral sense but also on practical experience. Developing the capacity to act well in different moral scenarios requires us to act, and reflect thoughtfully on our past decisions. This reflection, in turn, requires rationality, and an ability to take into account the result of our actions.

It is this aspect of reasoned thought, however, which makes cyber phronesis so elusive. Actions on social media do not have clear, predictable outcomes, subject as they are to the chaotic forces of large networks and inscrutable content-controlling algorithms. This ambiguity, and the difficulty of determining the motive behind short online posts, makes it tempting to reduce online virtuous action to mere 'virtue signalling', and argue that virtuous actions are conducted in public online space not because this is the right thing to do, but because it makes the actor look good. Below, we attempt to inform some of the reasoning required for true cyber phronesis, by developing an evidence base for the positive real-world impact of actions which 'feel right' on social media.

With continued advances in machine-led approaches to analysing natural language, we are increasingly able to study online societies at scale. The irreverent human hubbub present on social media offers a rare chance to examine virtue as it is discussed 'in the wild'; used as part of people's everyday speech, and in discussions where the deeper meaning of a moral concept may be, for an author who mainly wants to discuss, say, the actions of public figure, the least important part of a message. This social listening approach poses significant challenges. In moving away from the structure of survey questions and moderated discussions, you force yourself to deal with the exuberant, ambiguous nature of human speech, delivered in high volumes, and in messages often too short to allow a full assessment of motive and meaning. Enlisting machines to help interpret this text brings its own set of challenges. Studying human speech on public social platforms, not only avoids the Hawthorne effect of answers changing under analysis, but also allows for a crucial element

2 Aristotle, 'Nichomacahean Ethics' Book 1:1,2
3 Aristotle, Nicomachean Ethics – Ross translation, 1984, 1140b5

of surprise. By conducting our research in spaces where speech is already underway, we discover new and unexpected contexts for virtuous behaviour and use of moral terms.

In this paper, conducted with the University of Birmingham's Jubilee Centre for Character and Virtues, we study how the ancient concepts of moral virtue are presented on Twitter. Below, we conduct a detailed examination of how the moral virtues of honesty, empathy, courage and humility are used on the platform, as well as conducting a series of experiments to measure online moral action – tweets which might themselves constitute or indicate the virtuous behaviours of charity, gratitude and application to learning a skill. In doing so, we aim to provide a novel evidence base for the real-world impact of actions in online spaces, as well as the connection between discussion of moral concepts and virtuous actions online.

This report builds upon a previous body of research into the development on character online, including previously published collaborations between Demos and the Jubilee Centre concerning the internet's influence on both positive and negative character traits, and an in-depth examination of the role of character in education.[4] It also stands on the shoulders of recent developments in the philosophy of language – in particular Vasalou's 2012 paper on the importance of mastering terms related to moral virtue, and work from Wheeler et. al. which shows a decline in the relative frequencies of moral terms appearing in books since 1900.[5] In conducting this research we hope to fill some of the space between these investigations, aiming first to measure virtuous action online, but also to evidence any relationship between people's use of virtuous language and propensity for virtuous action on social media. While we cannot here match the century-long breadth of Wheeler et al.'s work, it is hoped that this paper will also lay the foundation for future investigations in to the changing frequency of virtuous language in online spaces.

This paper is split into two main sections, each based on machine-driven analysis of substantial quantities of publicly shared data on Twitter, connected with secondary datasets. The first concerns discussion within the UK of the moral concepts of courage, empathy, honesty and humility. In the second section, we focus on three virtuous online behaviours in turn – the sharing of links to charitable fundraisers, expressions of gratitude on Twitter, and the use of hashtags which encourage the user to apply themselves to learning a skill and publish their progress online. In each case we attempt to assess the real-life impact of these behaviours, tying them to actual money raised, or violin practise recorded. Each of these sections was written to be self-contained, and can be read alone without reference to the rest of the paper.

The document also contains two addenda – a methodological annex, which outlines the processes involved in training the natural language processing classifiers used in this research, and a full literature review which gives a theoretical introduction to previous research related to virtue and social media.

4 Harrison-Evans, P. and A. Krasodomski-Jones (2017). The Moral Web: Youth Character Ethics and Behaviour, Demos.
Birdwell, J. Scott, R. Reynolds, L. (2015) Character Nation, Demos.

5 Vasalou, S. (2012) Educating Virtue as a Mastery of Language Ethics 16:67
Wheeler MA, McGrath MJ, Haslam N (2019) Twentieth century morality: The rise and fall of moral concepts from 1900 to 2007. PLoS ONE 14(2): e0212267.

NOTES ON INTERPRETATION

In the course of this research, we make assessments as to the virtue inherent in actions taken online, based on the content of online speech. In doing so, we face the problem of assessing motive. Given the extreme brevity of tweets, it is difficult to ascertain the moral impetus behind any given message. We have attempted to mitigate this within our experiments into virtuous action by assessing the real-world results of online discussion, in terms of money raised, for example, but our results should be interpreted in the light of the inscrutability of moral motivation through observed behaviour.

While Twitter is regularly used by millions of people in the UK, the platform is not representative of the UK population.[6] Twitter users are disproportionately male, and predominantly young – though there are more elderly users on the platform than is often estimated.[7] As of mid-2018, around a fifth of UK adults had accessed Twitter within the last three months (21%).[8]

Below, we employ several machine-learning algorithms built specifically for this research. These are trained to decide, for example, whether a tweet using a virtue term is using that term in a negative or positive sense, or whether a tweet expresses genuine gratitude. Classification of this type is an inherently probabilistic process and especially challenging when interpreting intrinsically human concepts such as the nature of virtue. As a result, none of these algorithms are 100% accurate - the average accuracy for classifiers trained for this report is around 77%. Full details on how these algorithms were trained, the accuracy of each, and exemplar messages falling into each category, are laid out in full in the methodological annex to this report.

6 This report estimates 13.1 million monthly active users in the UK: avocadosocial.com/the-latest-uk-social-media-statistics- for-2018/

7 Sloan (2017): Who tweets in the United Kingdom? 12 Profiling the Twitter Population Using the British Social Attitudes Survey 2015. Retrieved July 2019 from journals.sagepub.com/doi/ full/10.1177/2056305117698981

8 See Ipsos MORI's Tech Tracker, available from ipsos.com/ipsos-mori/en-uk/2-3- adults-britain-use-social-media

SECTION 1
WHAT DO WE TWEET ABOUT WHEN WE TWEET ABOUT VIRTUE?

Examining the language of moral virtue online

Language is plastic. It changes and distorts with use, moulded through reinterpretation and misunderstanding. The evolutionary theory of linguistics argues that language changes through adaptation and propagation; over time, set conventions are broken, and new uses are picked up, passed on and amplified.[9] This change, of course, is a fundamentally social process.

We live in an age where conditions for linguistic propagation are riper than they have ever been. A near-ubiquitous internet, and the social platforms which have grown over it like so much glowing moss, has provided a global forum for linguistic exchange; spaces where words and concepts can be misused, borrowed, altered and evolved. As Goel et al. point out in a 2016 paper, numerous studies have found a link between social media and 'an increase in linguistic diversity and creativity'. This change can be global - the paper itself finds in particular that language change on Twitter diffuses through close ties formed online, which disregard geographic location.[10]

Encouraging the development of virtuous thought and action is one of society's highest goals. Character and resilience remains a significant focus of English educational policy; in February 2019, then Education Secretary Damian Hinds announced a new programme for building character, enabling pupils to appreciate "the importance of positive personal attributes – such as self-respect and self-worth, honesty, courage, kindness, generosity, trustworthiness and a sense of justice."[11] As we continue to struggle through an era of social and political uncertainty, it is vital that this focus is maintained, and these attributes are enabled and nurtured.

The language underpinning those virtuous attributes listed above, however, is not immune from social media's forces of linguistic change. If we are to succeed in encouraging the development of character, it is vital that we understand how people in the UK are already using and interpreting these terms online.

The following section aims to tackle this gap in our knowledge. Below, we analyse just over a million tweets sent from the UK between 2018 and 2019, which discuss the moral virtues of honesty, empathy, humility and courage. This data is used to develop an evidence-driven view of the ways in which these terms are used today.

It is clearly possible to behave virtuously without having a strong grasp of the related concepts. A toddler might display empathy towards a younger sibling, although she is highly unlikely to be able to

9 See e.g. Baxter, Gareth & Blythe, Richard & Croft, William & J. McKane, Alan. (2006). Utterance Selection Model of Language Change.
10 Goel, R., Soni, S.P., Goyal, N., Paparrizos, J., Wallach, H.M., Díaz, F., & Eisenstein, J. (2016). 'The Social Dynamics of Language Change in Online Networks.' ArXiv, abs/1609.02075
11 Department for Education (2019) 'Education Secretary sets out vision for character and resilience' - retrieved 13/08 from https://www.gov.uk/government/news/education-secretary-sets-out-vision-for-character-and-resilience

define 'empathy'. As characters develop, however, reading and writing about the concepts of virtue can help us navigate our own boundaries of good behaviour – to define what exactly about morality's abstract concepts is important to us, both as individuals and as a part of society. Social media offers a uniquely public platform on which this exploration can occur.

Methodology

In order to understand how the UK discusses moral virtues on Twitter, the platform's streaming API was used to collect tweets which contained one or more of the following terms, over a period between July 2018 and April 2019:

- courage
- empathy
- honesty
- humility

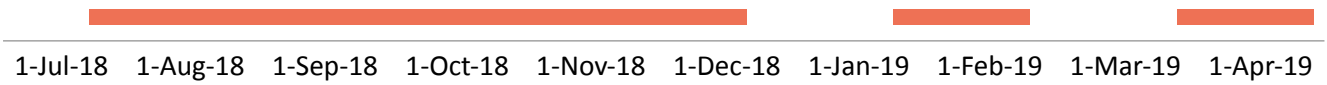
These terms were purposefully chosen in these forms to increase the chance of collecting discussions of virtues as concepts, rather than as adjectives or in everyday phrasal use. The shortlist above was arrived at after a process of iterative elimination after some initially considered terms – such as 'justice' – were found to produce too much irrelevant discussion (e.g. about the 'Justice System'). Tweets were collected using Method52, a suite of tools for analysing large free-text datasets developed by Demos in partnership with the University of Sussex.

Due to technical issues with accessing Twitter's API continuously, the collection was broken into three large periods, during which all tweets containing a relevant term were collected: 15th July to 4th December 2018; 15th January to 15th February 2019; and 16th March to 12th April 2019. In total, tweets were collected for 203 days. These periods are shown in Figure 1.

FIGURE 1.

PERIODS FOR WHICH TWEETS ARE PRESENT IN OUR COLLECTION

Represented in orange



Our aim here was to study, where possible, discussions around virtue sent by people within the United Kingdom. Accordingly, an algorithm was used to remove from the dataset tweets from accounts which were not likely to be based in the UK. While locational information is not always directly available from Twitter, the country from which a user is tweeting can often be inferred given information they provide on their public profile - the contents of a free text 'location' field, for example. Accordingly, all tweets collected during this research were passed through an existing classification algorithm within Method52, which removed all tweets not likely to have been sent from the UK. This algorithm assigned tweets to the country they were likely sent from using metadata fields, such as a free text 'location' field, included with the tweet. This algorithm estimated that around 20% of tweets collected globally were sent from the UK. The resulting collection is displayed in Table 2.

TABLE 2.

NUMBER OF TWEETS AND USERS USING EACH TERM IN OUR COLLECTION.

Tweets which contain more than one collected term have been counted in the rows for each term.

Term	UK Tweets	UK Users	Average Tweets per year
Total	1,014,546	385,302	2.6
Courage	451,695	206,352	2.2
Honesty	272,997	155,640	1.8
Empathy	230,650	122,660	1.9
Humility	76,347	52,288	1.5

These numbers give us a sense of the scale of our dataset, as allows us to compare the usage of the four terms collected. Interestingly, it shows us that courage is not only the most often discussed of our terms, but that it is used more frequently by those who use it. As we will see below, it is possible that this dominance could be related to the political events of late 2018 – many users in the dataset discuss the courage, or lack thereof, of political and public figures.

These volumes, however, do not reflect discussion on the platform as a whole, as they are taken from a sample of people already using virtue terms. This data alone cannot tell us how relatively often each term occurs on Twitter in general, or, more interestingly, how this usage compares to other public spaces.

To answer these questions, a large random sample of around 500 thousand tweets was collected, using an API provided by Twitter which returns a random 1% sample of all tweets sent on the platform. Table 3 displays the number of times each of our collection terms occur in this random sample of UK Twitter usage, alongside a measure of roughly how often these terms occur. These measures are then compared to two other sources of British language: speeches made by MPs and members of the House of Lords between October 2018 and July 2019, as recorded in Hansard, and SUBTLEX-UK, which contains word frequencies for 201 million words broadcast on various BBC channels.[12] Comparing three fields which have long influenced UK public discussion - parliamentary speeches, our broadcast media and discussion online - throws up some interesting findings.

Table 3 clearly shows that the moral virtues of honesty, courage, empathy and humility are mentioned relatively often in parliament - ten to fifteen times as often as the other spheres we measured. It is possible that this increased use of moral terms may be being thrown back at MPs by the tweeting public, who often use them to level criticism at politicians.

More surprising is the gap between use on Twitter and the BBC, with virtue terms appearing 1.5 times as often on the former than the latter. Even on a random sample of UK tweets, moral virtues seem to play an important part in discussion on the platform. Indeed, while the language of moral virtue is less prevalent on Twitter than in the Houses of Parliament, this comparison with the BBC shows that it is very much part of everyday discussion on the platform; more than we might expect.

Examining the relative occurrence of terms on each platform shows another difference between these spheres of discussion, as seen in Figure 4. This shows, firstly, that the BBC and Parliament are remarkably close in the balance of terms used; though 'honesty' appears more often in BBC transcripts than it does in Hansard.

TABLE 3.

USES AND FREQUENCY OF VIRTUE TERMS ACROSS VARIOUS SPHERES OF PUBLIC LIFE.

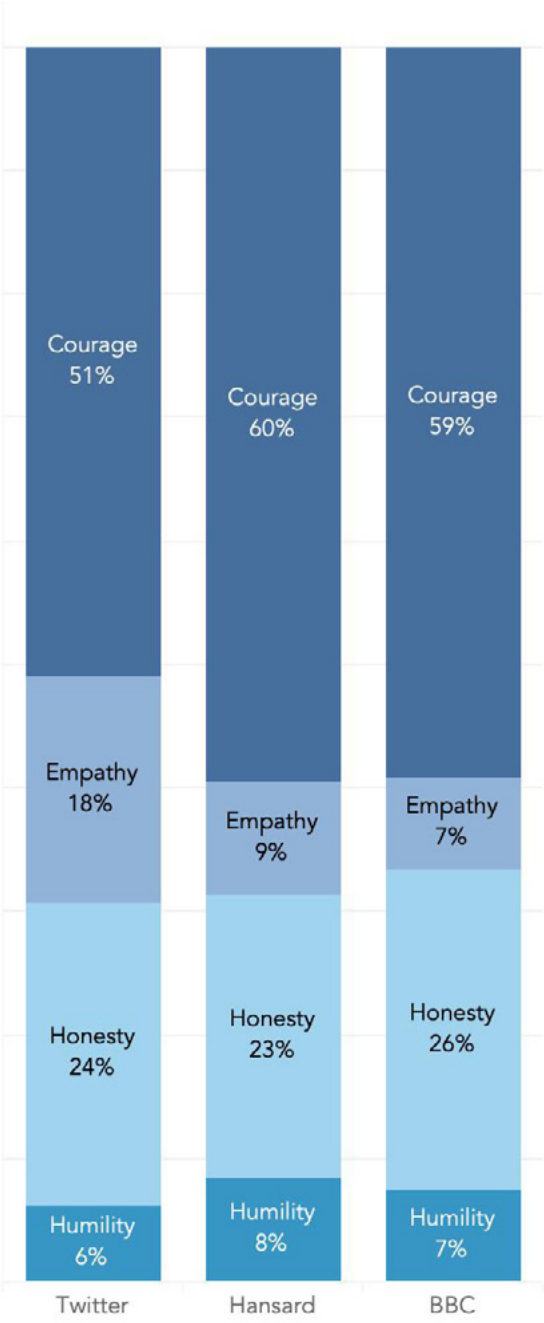
Source	Virtue terms used	Sample size in words	Approx. virtue term frequency
Twitter	148	3 Million	Once every 20 thousand words
Hansard	1285	2.5 Million	Once every 2 thousand words
BBC	5324	202 Million	Once every 30 thousand words

12 Both Hansard and SUBTLEX-UK are open source datasets, and each in its way is extremely rich. A basic API for querying parliamentary speeches since 1803 is available at https://api.parliament.uk/historic-hansard/api, and SUBTLEX-UK can be found at http://crr.ugent.be/archives/1423

The most notable aspect of this graph is in the use of 'empathy', which appears twice as often on Twitter as anywhere else. This does not, of course, mean that the platform's users are more empathetic - only that it is more often discussed, in relative terms, on Twitter.

FIGURE 4.

USE OF EACH COLLECTION TERM AS A PERCENTAGE OF USES OF THE FOUR VIRTUE TERMS COLLECTED, BY PLATFORM



Use of virtue terms to praise and criticise online

This report seeks to answer a series of questions concerning the use of moral virtue terms on Twitter's modern milieu. To this end, it is to some extent philosophical in nature. The data it examines, however, is very much grounded in the messy, human reality of social media – a format highly reactive to events, and where the simplest most, emotionally charged messages can be seen and shared by thousands. While, as we will see later, Twitter does offer its users space to help define their understanding of what it means to act morally, many uses of virtue terms we collected involved people looking outwards – using the virtues we examined to praise or criticise the actions and character of others, particularly public figures.

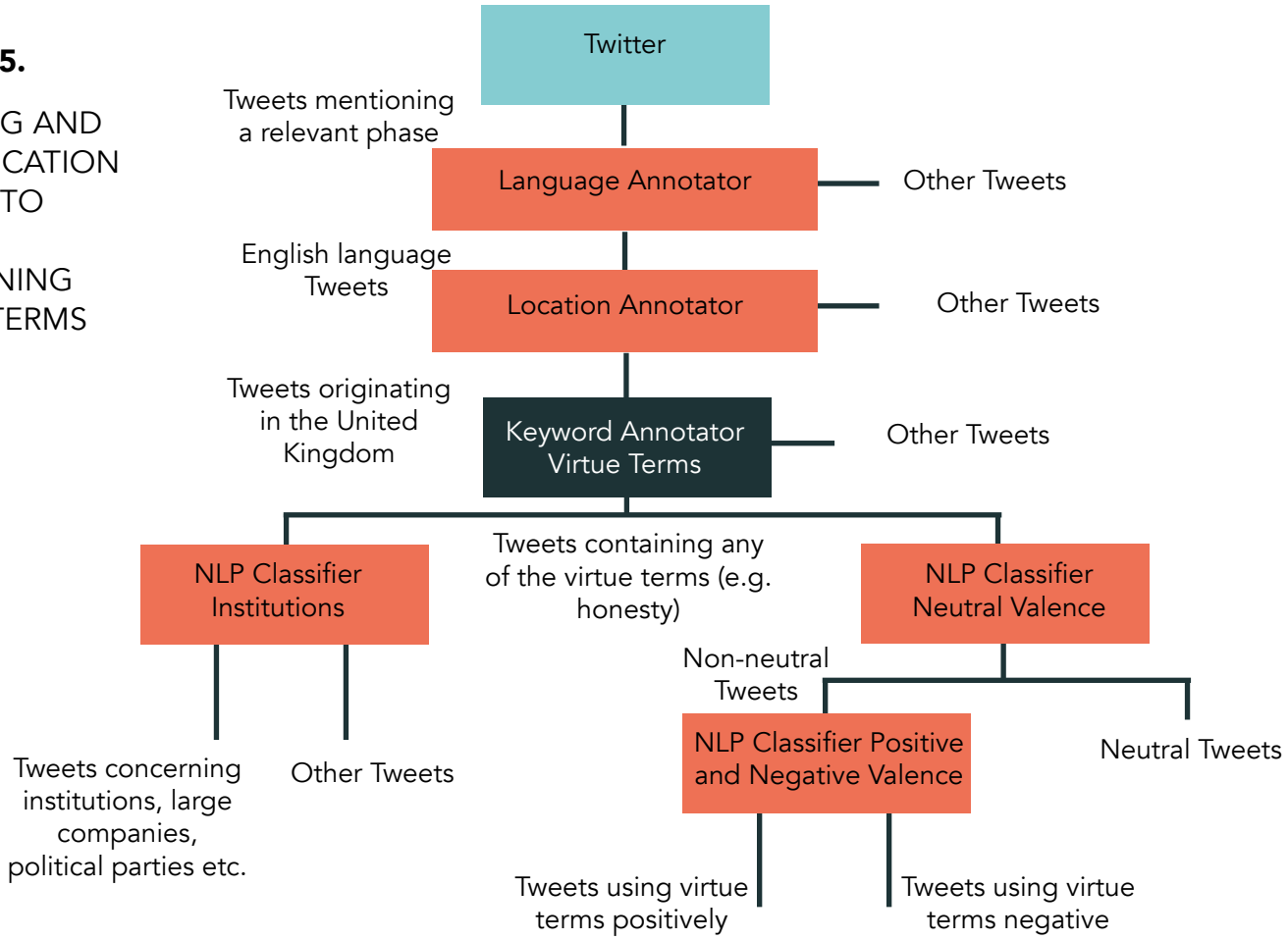
To study this at scale, we used Method52 to train a series of 'Natural Language Processing' (NLP) classifiers. The process for this is explained in detail in the report's technical annex, but these can roughly be seen as algorithms which can be trained to recognise patterns within sets of human language and use these to make distinctions which you would traditionally need a human to make; for example, given a tweet which contains the word 'honesty', is this being used in a phrasal sense (e.g. to say 'in all honesty…') or to discuss the honesty as a form of moral behaviour?

In this case, we used these classifiers to answer two separate questions of each tweet in our dataset. Firstly, a classifier was trained to work out the sentiment within use of virtue terms – whether they were being used to praise, for example, the humility of an individual, or bemoan a lack of courage. To do this we first built a classifier to identify and remove neutral uses of terms, including the common phrasal uses mentioned above (discussing 'Dutch courage' etc.) as well as promotion for events designed to help with courage, published papers studying empathy in animals etc. Remaining tweets were then classified to establish whether they were positive or negative.

A second classifier was trained to establish the subject of each tweet – in particular, whether it concerned the virtue (or lack thereof) of a national institution, including politics and politicians, the media and the health service. A diagram of this classifier pipeline is included in Figure 5.

FIGURE 5.

FILTERING AND CLASSIFICATION APPLIES TO TWEETS CONTAINING VIRTUE TERMS



Using algorithms to label tweets in this way allows us to examine trends within the entirety of our million tweet dataset, without having to rely on small manual samples hand-coded by researchers.

As Figure 6 shows, virtue terms are overwhelmingly likely to be used in order to express judgement, with non-neutral uses accounting for 71% of the dataset. Within these tweets, however, usage is remarkably balanced – 35% use virtue terms positively, against 36% negatively.

This balance, however, shifts with the subject under discussion. Around 248,000 (24%) of the tweets in our dataset discussed the perceived virtue of British institutions, those who represent them, and the actions taken on their behalf. Institutions here are broadly defined and include political parties and politicians, the media, public services and large corporations and brands.

We found that tweets concerning these 'institutions' were far more likely to comment negatively than tweets in general, with criticism taking up 48% of the conversation around 'institutions' – 60% of tweets which expressed an opinion on institutions expressed a negative one. As we will see later, this is very likely to be influenced by the national sense of uneasy turbulence produced by the alarming political events of late 2018, and very much still present at the time of writing.

This dominance of negativity aligns with trends noted elsewhere in the declining public trust in many facets of public life, especially politics. In a study conducted at the end of our collection period in 2019, the Hansard Society found that 72% of the UK feel that the system of governing needs 'quite a lot' or 'a great deal of' improvement – representing the lowest levels of faith in the political system for 15 years.[13]

FIGURE 6.

SENTIMENT WITHIN USES OF VIRTUE TERMS

13 The Hansard Society (2019) 'Audit of Political Engagement 16 -The 2019 Report.' Retrieved September 2019 from https://www.hansardsociety.org.uk/publications/reports/audit-of-political-engagement-16

Crucially, what this data adds to this is the finding that this criticism goes deeper than governmental competence, or the effectiveness of the UK's political system – rather, many of these criticisms are being made on moral grounds, questioning the character as well as the capability of politicians and other public figures.

DISCUSSION OVER TIME

In order to understand how conversations using virtue terms changed during our collection period, we examined two three-week samples from our collection in detail. The first of these periods, from 16th July to 6th August 2018, contains tweets sent right at the outset of our collection – the second, from 16th March to 8th April 2019, falls in the final weeks.

It is important to situate this discussion within the context of the political and social events the last half of 2018, and the fractious public conversation they engendered. Many of these, of course, relate to Brexit. In the first month of our collection, Boris Johnson resigned as foreign secretary, Vote Leave was fined for breaking electoral law and ministers revealed that, in response to a no-deal scenario, they would resort to the Army to deliver food, medicine and fuel.

This set the tone for the months to follow. Theresa May's withdrawal agreement prompted a broadside of resignations. When put to the House of Commons for a 'meaningful vote', her government in quick succession claimed the titles of 1st, 4th and 8th biggest government defeats in modern history. As hope of a resolution engendered by the December 2017 agreement with the EU was overwritten by political uncertainty and constitutional novelties, the pound slid, resignations mounted, and hundreds of thousands of people marched for a second referendum. The period closed with an extension until October 31st being agreed, avoiding no-deal but infuriating huge swathes of the population.

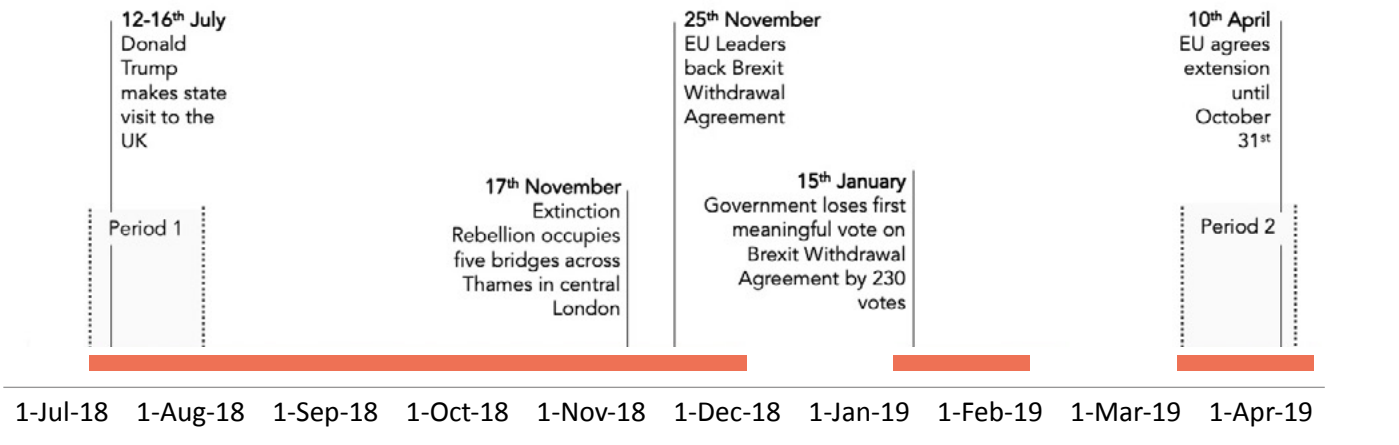
Alongside the Brexit drama, the climate crisis forced its way into the headlines thanks to activist groups Extinction Rebellion and the Youth Strikes for Climate. In November, Extinction Rebellion blockaded bridges across the Thames, and in April occupied much of central London, including the surroundings of Parliament Square.

During both periods picked out for closer analysis, we see a consistently high volume of tweets using moral terms – an average of around 5,500 tweets sent from the UK per day, most of which express a positive or negative opinion. This constant activity is punctuated by the surges in volume characteristic of social media use; often with regards to a particular discussion, or in response to a single event. Some of these are labelled, and discussed, below.

FIGURE 7.

POLITICAL CONTEXT FOR THIS RESEARCH.

Orange bars show periods for which tweets were collected.



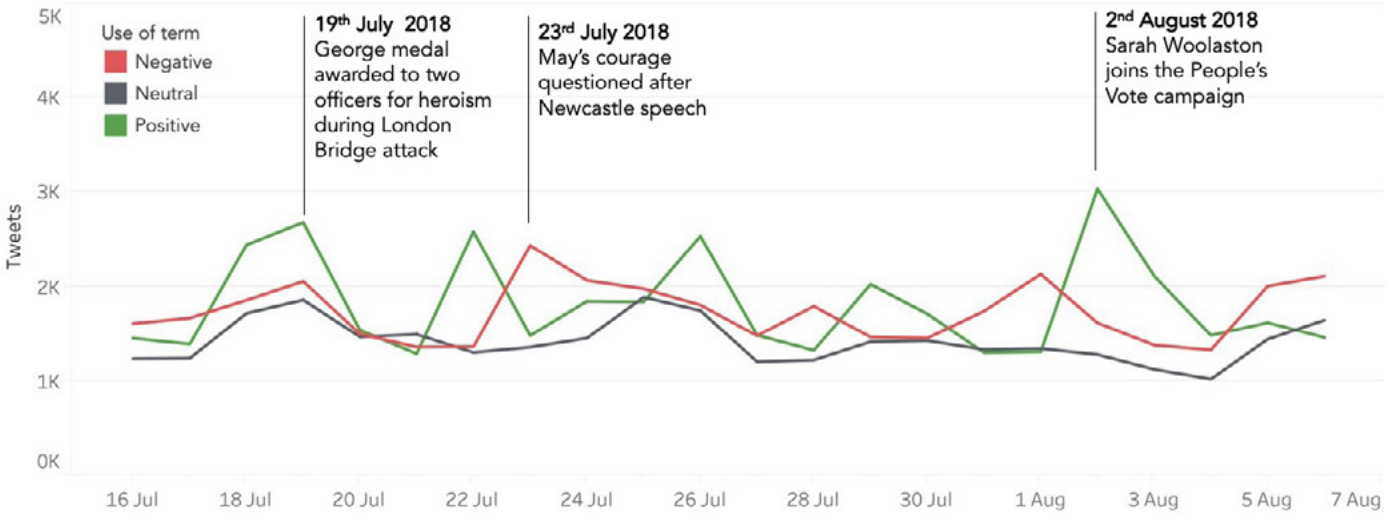
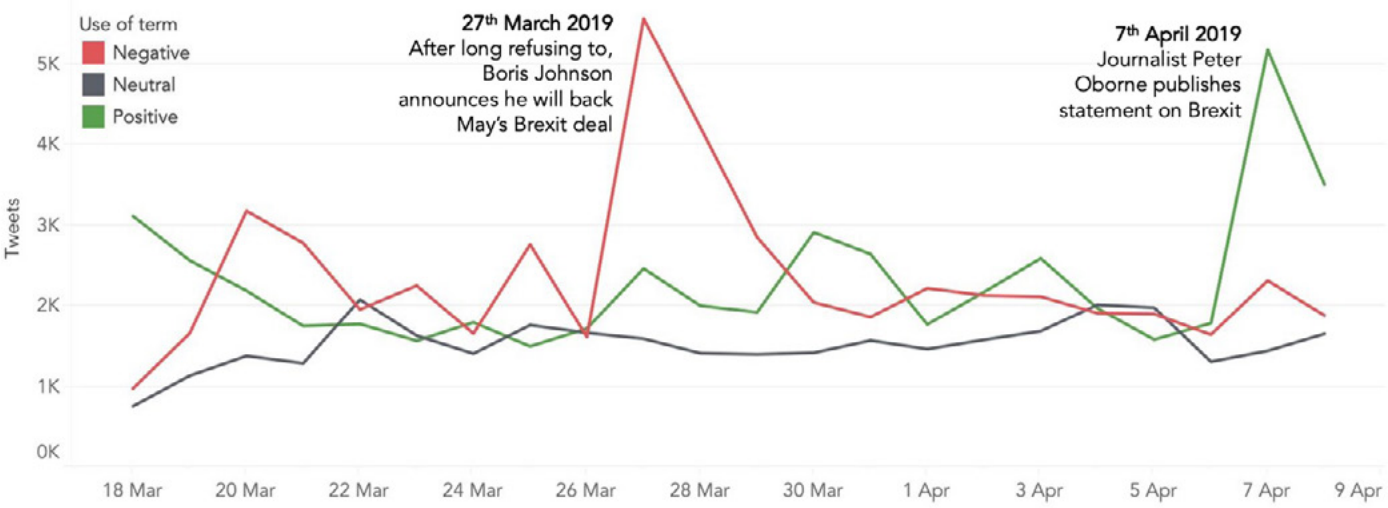


FIGURE 8.1. (Above)

PERIOD 1: SENTIMENT WITHIN TWEETS SENT BETWEEN 16TH JULY AND 6TH AUGUST 2018

FIGURE 8.2. (Below)

PERIOD 2: SENTIMENT WITHIN TWEETS SENT BETWEEN 27TH MARCH AND 8TH APRIL 2019



It is perhaps not surprising given the UK's recent history that much of the most eagerly shared content above concerns the actions of politicians. On 23rd July 2018, for example, a spike can be seen around a series of videos criticising May's performance in a recent speech in Newcastle, claiming she lacked 'vision and courage'. On 27th March, a widely retweeted message, sent by journalist Piers Morgan, sarcastically applauded Boris Johnson after he came to the 'sad conclusion' that he would support Theresa May's proposed withdrawal agreement on its third, and eventually unsuccessful, attempt at being passed by the House of Commons:[14]

> *BREAKING: Boris Johnson says he'll now back Theresa May's deal, which he's repeatedly trashed as a terrible deal he could never support. This is because if it passes, she quits & he may become Prime Minister. Such courage! Such principle! Such a shameless little ****.*
>
> *@piersmorgan, 27th March 2019*

This message alone was retweeted by 4.6 thousand people in the UK, making it the most commonly shared tweet in the entire dataset.

14 Mairs, N., (March 2019): 'Boris Johnson reaches 'sad conclusion' that he must back Theresa May's Brexit deal', Politics Home. Retrieved September 2019 from https://www.politicshome.com/news/uk/political-parties/conservative-party/news/102877/boris-johnson-reaches-sad-conclusion-he

Not all of the prominent political discussion is negative. On 2nd August, Sarah Woolaston's courage was praised for her announcement that she was joining the 'People's Vote' campaign for a confirmatory referendum on Britain exiting the EU. Nine months later, on 7th April 2019, journalist Peter Oborne, a forthright pro-Brexit voice during the referendum campaign, published a piece in which he said those who thought Britain should leave the EU should 'swallow our pride, and think again. Maybe it means rethinking the Brexit decision altogether.'[15] Oborne's willingness to openly discuss this change of view was applauded widely within our dataset – five of the ten most retweeted tweets sent that day praised the piece for its honesty. Praise, this time from across political boundaries, was also given to PCs Wayne Marques and Charles Guenigault who received, in July 2018, the George medal for their heroic courage during the Westminster terrorist attacks of June the previous year.[16]

Underpinning this daily discussion is a sobering trend only glimpsable in the busy graphs above. Tweets sent during the second period were on average more polarised, and more negative, than during the first - UK users sent, on average, 89 more negative tweets per day, and 164 fewer neutral tweets per day.

The surges in volume examined above give us a good sense of the most popular messages on UK Twitter at any time. Given that the platform, in pursuit of the universal social media goal of engagement, is designed to show its users tweets which people like them have found engaging in the past, examining these spikes of virality offers a useful window into the messages which large numbers of people were likely to be reading at the time. Focusing too hard on these spikes in interest, often pushed by the loudest voices online, risks obscuring the vital wider discussions underway here. Underpinning the events discussed above, this dataset contains a constant stream of political commentary across party lines, from passionate support for Jeremy Corbyn's administration to hardline anti-immigration messaging; but also thousands of tweets which talk about people's everyday lives. These tweets don't tend to attract much attention, and it is to this wider conversation we turn our focus next.

VARYING USE OF MORAL TERMS

Sentiment within uses of moral terms was also affected by the terms themselves. Interestingly, as shown in Figure 9, UK Twitter users were more likely to use the terms 'courage', 'humility' and 'honesty' in a positive sense, either to praise that virtue in others or to stress its importance in the modern world.

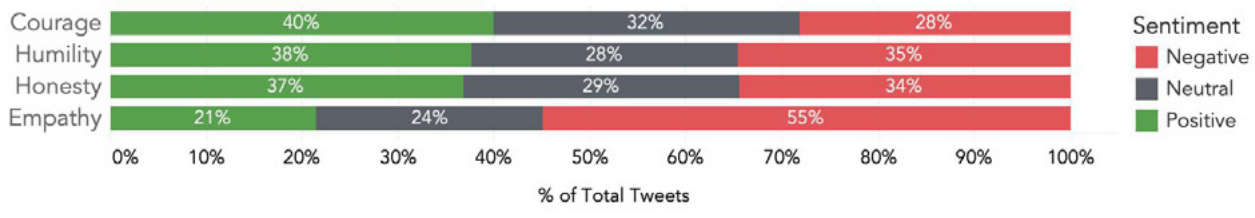
Strikingly, this trend is reversed for 'empathy', with over half (55%) of the uses of that term appearing in a negative sense. This is likely explained by the high number of discussions around this term which involve abuse and adversity – the topic analysis conducted below indicate that these themes could take up around 46% of the dataset.

TOPICS WITHIN VIRTUE DISCUSSION

Twitter, as said above, is a reactive medium. People tend to post in response to events which have happened to them in their personal life – things they've experienced, read about in the news, seen flash past on a screen. This isn't to say people don't tweet about concepts in the abstract; they do, and we explore this below. Most concepts, however, are typically used in connection with something else. This section aims to uncover those connections – to explore the links between moral virtues and the noisy, human chatter of everyday social media, and place these terms in their true social context.

FIGURE 9

SENTIMENT WITHIN USES OF VIRTUE TERMS



15 Oborne, Peter (April 2019) – "I was a strong Brexiteer. Now we must swallow our pride and think again" Open Democracy – Retrieved September 2019 from https://www.opendemocracy.net/en/opendemocracyuk/i-was-strong-brexiteer-now-we-must-swallow-our-pride-and-think-again/

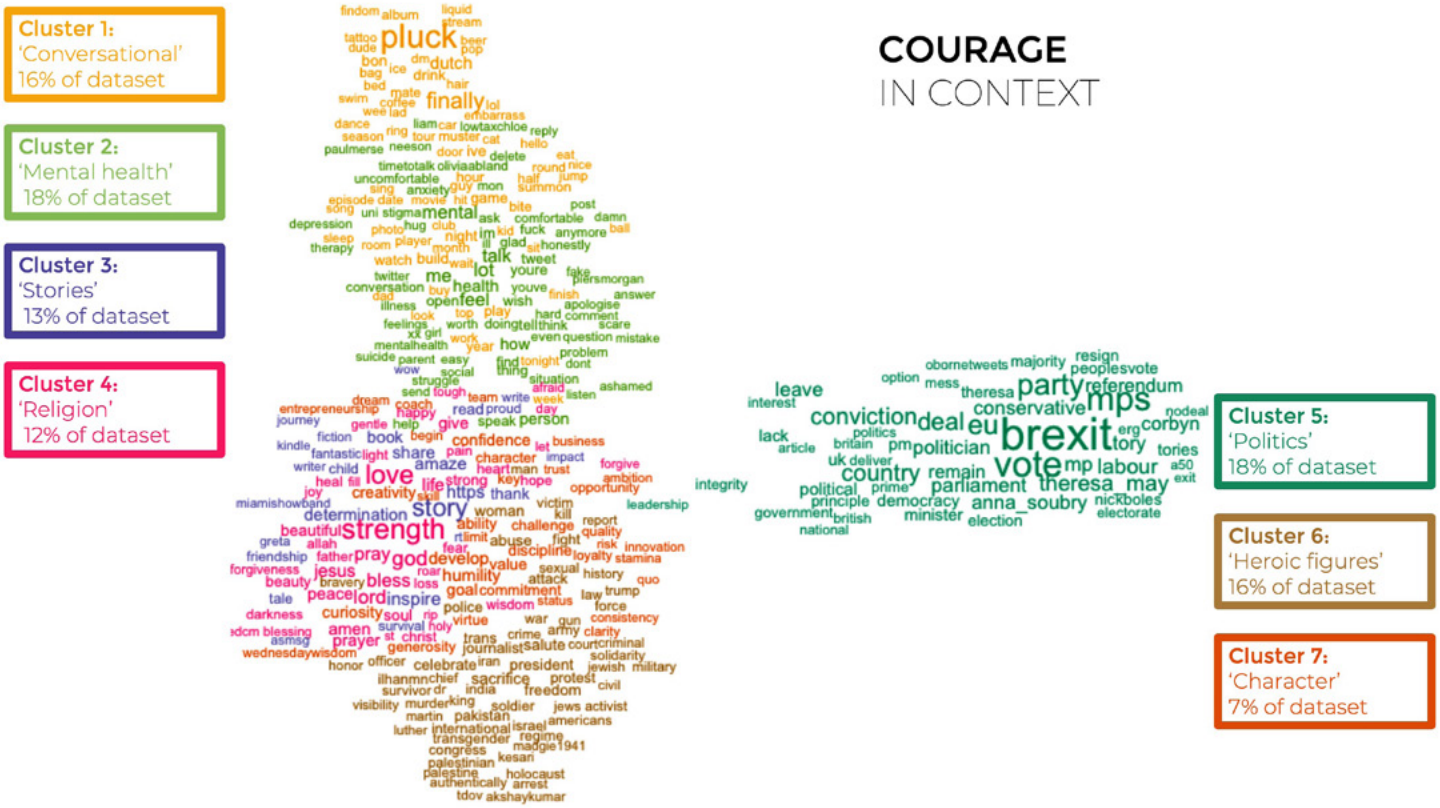
16 https://news.sky.com/story/london-bridge-terror-attack-heroes-among-those-receiving-gallantry-awards-11441311

In order to do so in a way which took the size of this vast dataset into account, we present a series of 'correspondence factor analysis' graphs.[17] These show clusters of terms which often appear together, which can be useful in unearthing themes within large, noisy datasets without relying on manual reading of a small sample of tweets.

Some notes on interpretation:

- Colours below indicate different word classes, which are collections of words that frequently occur closely together but rarely with words from other classes. These classes develop purely from these interactions within the data - their contents are not defined through human intervention.
- The position of the word classes on the graph shows how similar the classes are to one another; two coloured classes positioned next to one another contain words which are relatively likely to appear close together, though not likely enough to be placed in the same class.
- The size of the word indicates how 'characteristic' it is of that class; large words are very likely to occur alongside other words from that class and very unlikely to occur alongside words from other classes.
- After production, clusters were labelled by an analyst according to the theme they were judged to discuss. (e.g. 'Politics'.) These labels were not machine-generated, and while contentious themes have been verified through a manual reading of tweets present in the clusters, they rely on fallible human interpretation of the terms present.
- Clustering in this way is a chaotic process, and groups of terms occasionally arise due to peculiarities present during a given run of the algorithm. To guard against this, each graph was produced multiple times to ensure that emergent clusters were likely to be a genuine product of connections present within the dataset, rather than programmatic quirks.
- Importantly, while they surface what is most distinctive about each dataset, clusters shown do not represent the totality of themes present within a discussion.

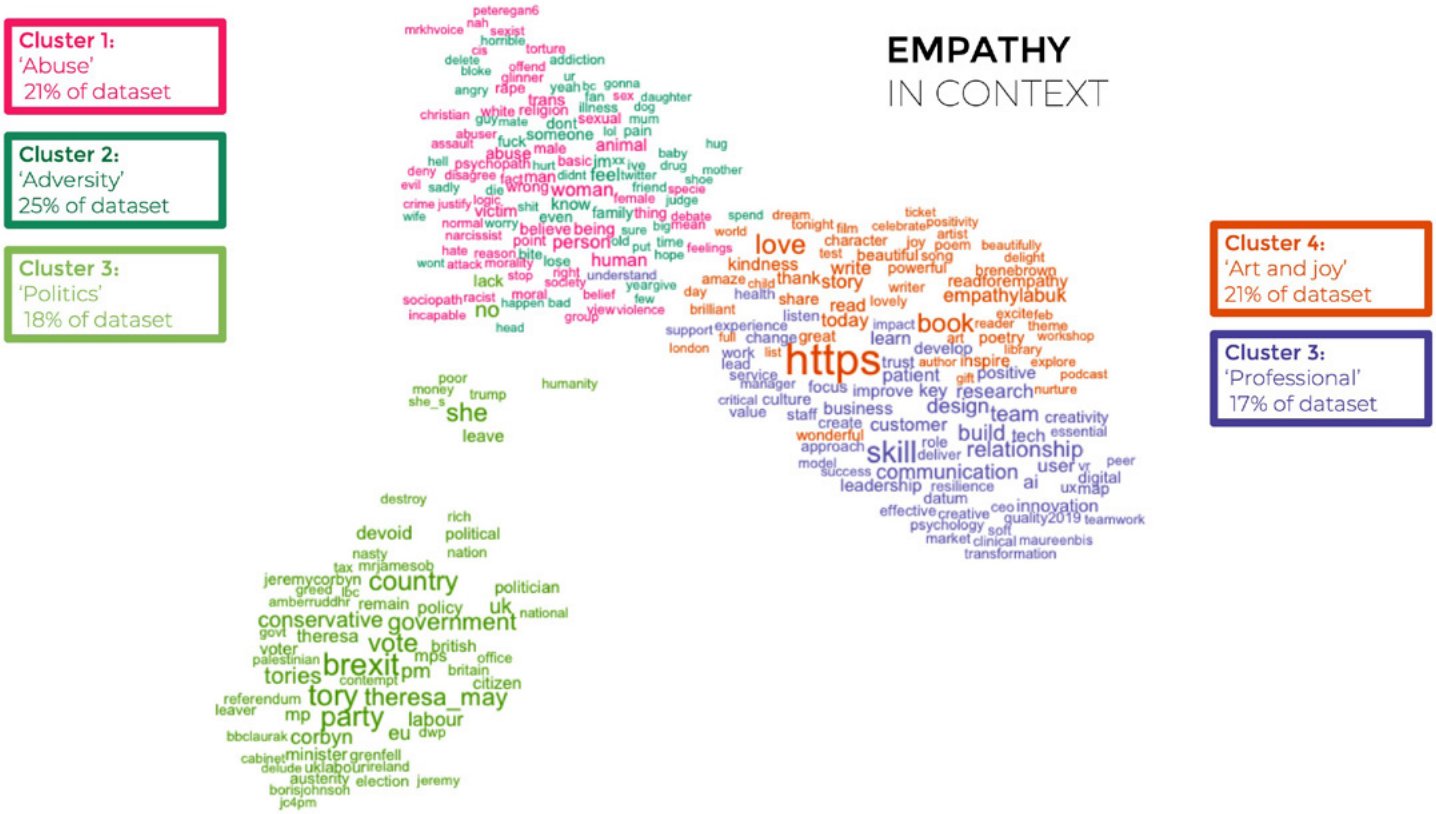
17 These graphs were produced using a piece of software called 'Iramuteq', developed by Pierre Ratinaud; another excellent open source resource. It can be downloaded, for free, from iramuteq.org



As with each of the moral terms collected, political discussion formed a distinct cluster around the concept of 'courage', with users urging leaders to resolve issues around Brexit. This cluster is also well separated from others, meaning people are likely to be discussing politics in a vacuum, without referring to other meanings of courage. As Figure 3 shows, this term is also well used by politicians, and we may be seeing here political discourse mirrored back by the public.

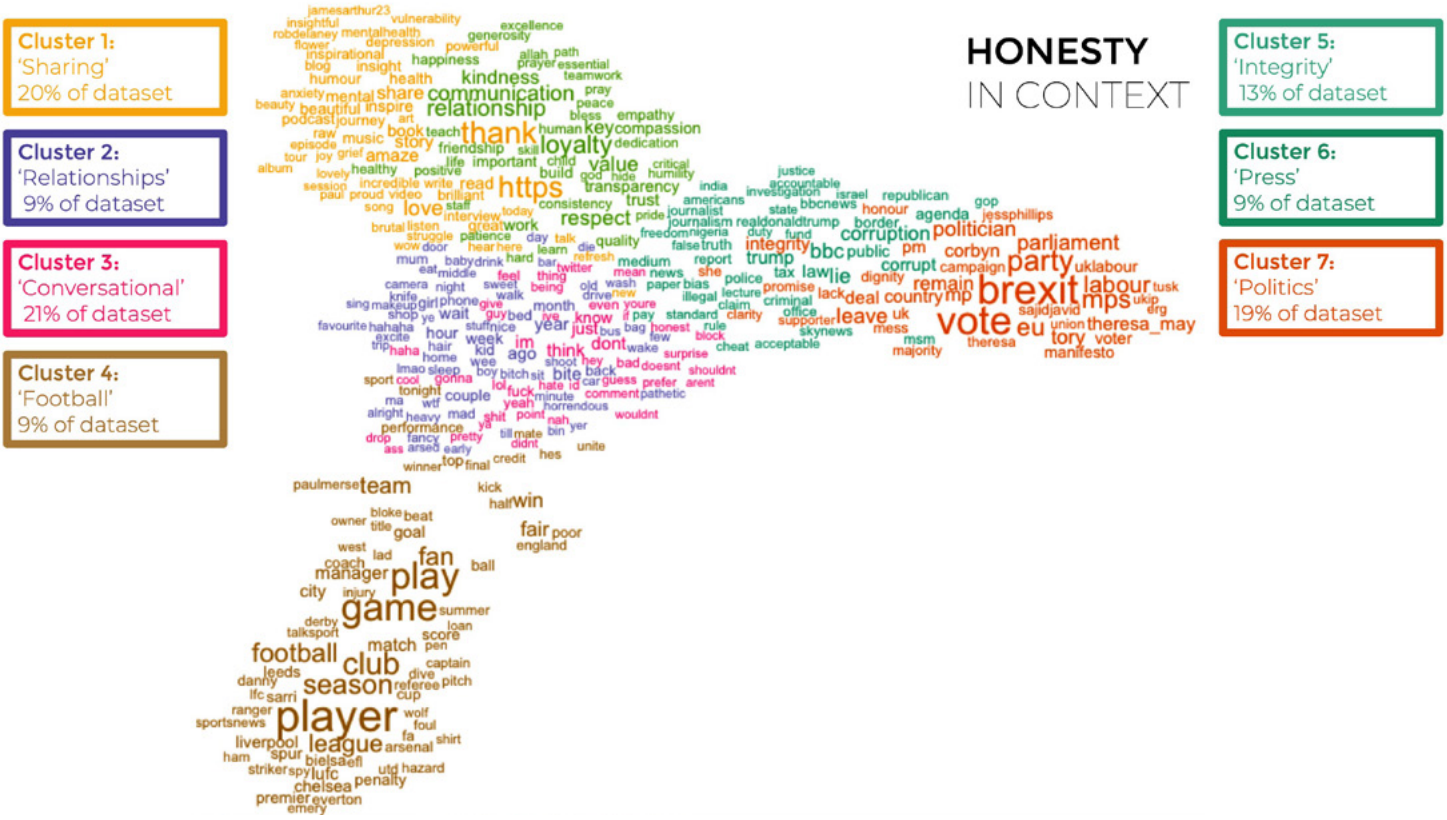
The concept also saw people sharing life experiences, with encouragements to open up around mental health, and discussion of heroic figures such as Martin Luther King and transsexual Pakistani newsreader Marvia Malik. Notably, well connected to discussions around these figures, religion and personal stories, we see a small cluster centred around character, comprising 7% of the dataset, and containing terms such as 'confidence', 'creativity', 'wisdom' and 'generosity'.

Finally, 'courage' was notable as the only term with a distinct 'phrasal' cluster, containing terms like 'pluck', 'dutch' and 'muster - though this only accounts for 16% of the dataset.



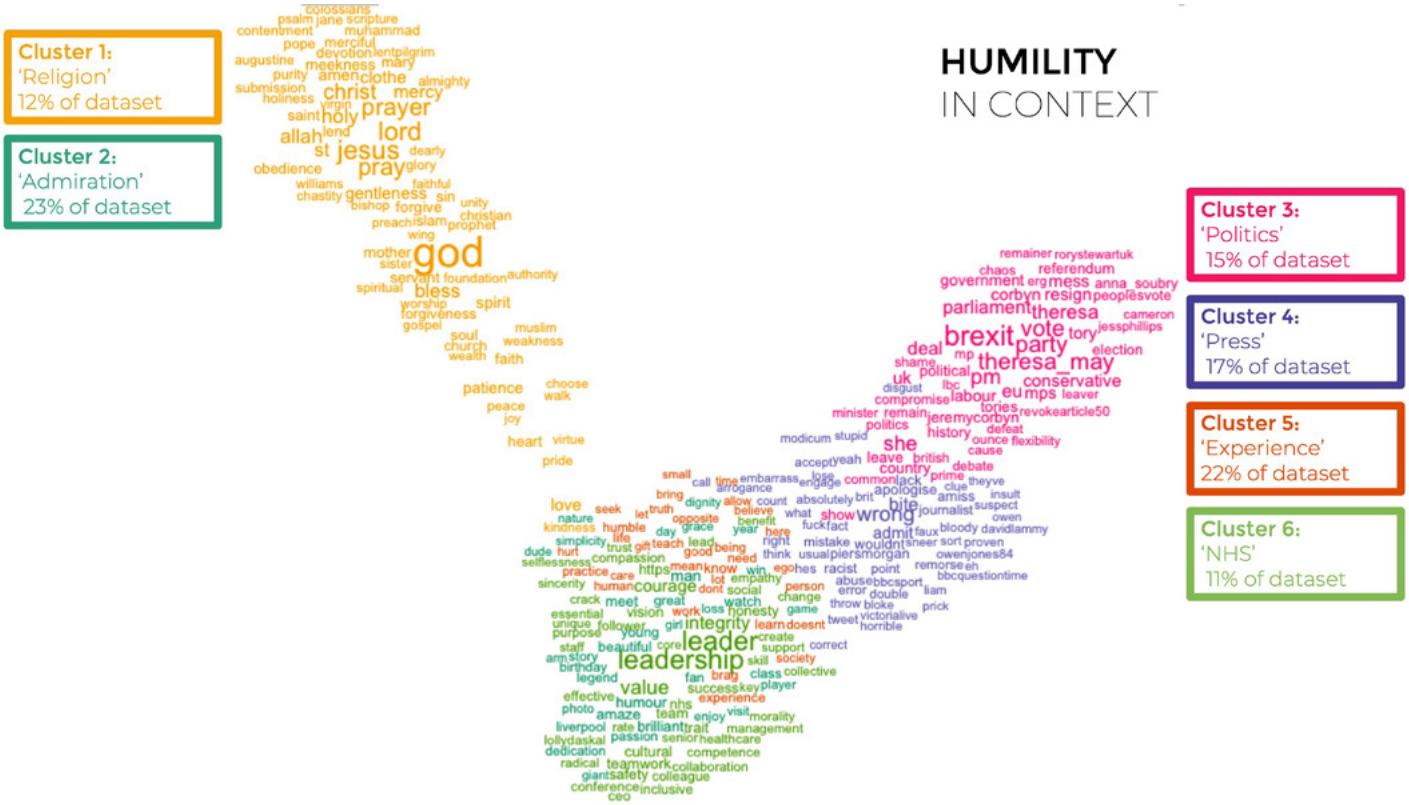
Again, we see here a separated political cluster alongside the concept of empathy, primarily decrying a perceived lack of the virtue in politicians, with a number of highly characteristic terms related to the Conservative party. Notably, the word 'Grenfell' also appears in this cluster.

The rest of the dataset is split in two. The first two clusters, making up 46% of the dataset, both concern empathy for those who have suffered, often the authors themselves - victims of sexual abuse, and those struggling with addiction and discrimination. Clusters 3 and 4, in contrast, focus on the positives of empathy, centring respectively around love, gratitude, poetry and literature, and empathy in the professional world.



A discussion around honesty in politics comprises close to a fifth of the dataset and is closely linked to cluster 6, concerning the press. Again, much of this discussion seems to be negative, with the terms 'corruption', 'lie' and 'bias' appearing. The graph also contains a distinct conversation around sport, with terms about football matches intermixed with 'dive', 'referee' and 'foul'.

On a more positive note, Cluster 1 references the honesty of people sharing their stories. As seen in topics around 'courage' this includes discussions of mental health and grief. Also, like courage, this positive discussion is linked to a cluster discussing character and integrity, which discusses religion alongside loyalty, trust and respect.



Humility is notable for a clear section on religion, comprising 12% of the dataset. While religious terminology has appeared alongside the other virtues examined above (though is notably absent from discussions around 'empathy'), worship has tended to be positioned close to other topics – it's one theme in a wider conversation. In this case, it is being discussed alone, possibly as users share a godly imperative to remain humble.

As seen before, public figures are closely linked here, with politics and the press appearing together. In this case, we also see the emergence of discussions around the NHS, alongside terms such as 'compassion', 'selflessness' and 'integrity'. A review of a sample of tweets from this cluster showed users praising the individuals in the health service, and discussing the need for humility in its management and leadership.

CONNECTIONS BETWEEN VIRTUE TERMS IN ONLINE USE

One finding from the topic analysis above is that virtue terms are often used together, with each of the graphs above displaying at least one group of words which relate to virtuous behaviour or good character. This analysis emphasises the connections between virtues in everyday speech.

To explore these connections further, we analysed every UK tweet in our dataset to build a map of connections between specific terms, based simply on whether they were used together in a message. Terms searched for were taken from the Jubilee Centre's Framework for Character Education in Schools, which outlines a typology of virtues fundamental to the development of flourishing individuals.[18]

This was then used to produce a linguistic network map shown in Figure 11. In this network, terms are linked according to whether they have appeared together in a tweet – in an effort to make the visualisation readable, we show these links (or 'edges') faintly in black below. The stronger a link gets, the closer two terms appear together, meaning that, as in the graphs above, terms will end up closer to those they often share a tweet with. Words have also been coloured according to the type of virtue they represent, according to the Framework:

- White: Collection terms (these are also moral virtues)
- Red: Moral virtues
- Blue: Performance virtues
- Olive: Civic virtues
- Pink: Intellectual virtues

Finally, the single overarching term 'wisdom' was added, in green.

In interpreting this linguistic map, it is worth reiterating that the only thing which determines the position of each term is the language they appear alongside. It is, therefore, striking that virtues of each, coloured type tend to appear together. Unsurprisingly, we see our collection terms at the core of the graph – after all, every tweet in the dataset will contain one of these terms – but they are also closely surrounded by the moral virtues of justice, integrity and respect. Intellectual virtues also appear together, with determination closely linked to perseverance and resilience, but also to the moral virtue of courage.

Some of the patterns seen above are through second-hand connections; they exist because of the network effect. For example, while intellectual virtues appear in the same section of the map, they are not in fact well linked to each other. Rather, they are often discussed in similar contexts; of gratitude and generosity, of optimism. We also see some terms separated from others of their type. Gratitude, for example, appears commonly alongside forgiveness but is more closely linked to confidence and generosity than to other moral virtues.

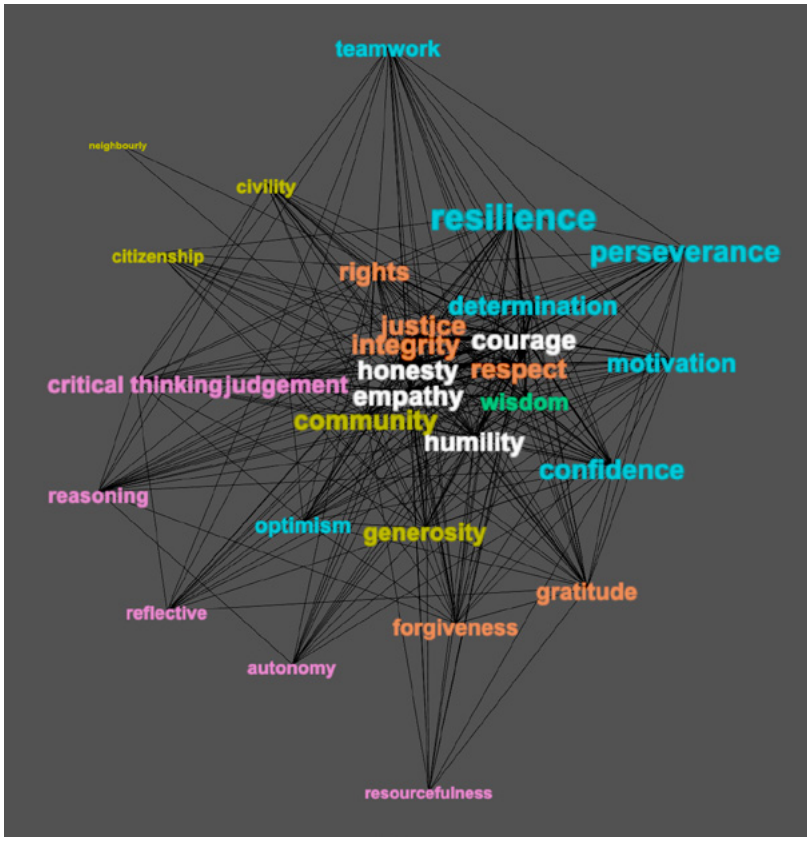


FIGURE 11

LINGUISTIC NETWORK OF VIRTUE TERMS APPEARING TOGETHER IN TWEETS

18 The Jubilee Centre for Character and Virtues (2017) 'A Framework for Character Education in Schools

Moral by Definition - Use of social media to define and explore virtues.

During the analysis conducted above, we found that we were often presented with tweets which, without necessarily addressing any specific person, laid out personal definitions of virtue and virtuous behaviour. These declarative tweets are a particularly interesting use of virtue terms, not only because they present a set of modern definitions of ancient concepts, but also because they represent a particular and interesting response to Twitter's promise of a universal audience. Given the potential of a global soapbox, some in the UK have chosen to define what they mean by morality.

Below we present an in-depth analysis of tweets which make a clear statement of definition, exploring and stating the meaning of virtue, or drawing the boundaries of virtuous behaviour. In order to submit these to scrutiny, we first searched within our dataset for tweets which contain one of a specific set of prescriptive terms, such as 'I define humility' or 'empathy is'.[19] A random sample of 50 tweets using each virtue term was then coded by hand, recording the broad sense of the definition, as well as the tweet's probable intended audience. It should be noted that these samples represent a tiny proportion of the overall discussion and that while this analysis presents some interesting examples of actual uses of virtue terms on Twitter, they cannot be taken to be representative of conversation on the platform as a whole – let alone conversation in the UK.

In analysing these tweets, a number of prominent themes arose, which are discussed below with examples. In all cases, to protect the privacy of their authors, tweets have been 'bowdlerised' – the wording of each has been altered in such a way as to preserve its meaning, but prevent the form of words from leading, through an online search, to the account of the user who posted them.

SELF-PRESENTATION

As Erving Goffman identified in The Presentation of Self in Everyday Life, people are continually constructing and reconstructing their identities through individual acts.[20] As our online identities have come to carry significant meaning, we might expect people to express their positive personal values on Twitter as an act of self-presentation. To investigate this, tweets were coded by hand to establish the likely intended audience for each message. We found that in definitional tweets the imagined audience was usually non-specific, and could be largely defined as an abstract "everyone." As argued by Litt and Hargittai in 2016, this "abstract" audience is used when an individual wants to express a personal value instantly, usually as a form of self-presentation. Some examples of this form of expression are below.[21]

> *Fear is what stops you, and courage is what keeps you going.*

> *Humility is really important. Being able to humble yourself and just listen to advice and critiques will take you so much further than trying to do it by yourself.*

POLICING OR PRAISING THE BEHAVIOUR OF OTHERS

Another way in which people define virtues online is through describing other people, or their behaviour, as lacking in some important sense. These tweets tend to be targeted at specific individuals or groups. In these expressed opinions, virtue terms are often used to call into question the character of individuals, and people whose behaviour is not considered virtuous are often "called out" online for their actions. While these individuals are often in the public eye, these terms are also used in direct arguments between two users of the platform:

> *@[username] The way you lack empathy is incredible.. instead of saying anything kind or productive you feel like you have a right to tell victims they deserve to be physically hurt!!?!?! You need to look at yourself!! Lastly how on earth can you think people want attention*

> *All that you say is literally the precise opposite of being humble. Humility level is 0 = You think some "higher power" cares about JUST YOU and made the World entirely JUST for YOU. Believing you are humble while thinking that is the true irony.*

These tweets give an example of social media allowing people to impose their views on the importance of morality on others, with the value of moral behaviour continually reinforced through online conversations. This impulse is also expressed as a high form of praise:

19 A full list of these terms is included in the methodological annex to this document.
20 Goffman, E (1959) 'The Presentation of Self in Everyday Life'
21 Litt, Hargittai (2016) 'The Imagined Audience on Social Networking Sites' Social Media and Society volume 2 issue 1

@bbcquestiontime This woman is brilliant – the honesty she shows is worth thousands of the lies of self interested @Jacob_Rees_ Mogg and #RodLiddle

His humility is adorable #XFactor

QUOTES

We also found virtues commonly defined on Twitter through shared quotes from historical figures.

Courage is contagious. When a brave man takes a stand, the spines of others are stiffened. - Billy Graham #quote

Honesty is the first chapter in the book of wisdom.- Thomas Jefferson [link] #Quotes

Notably, we found there is a stark gender divide in the individuals being quoted, reflecting societal narratives which privilege the authority of men's voices. Of the 200 tweets we analysed, 24 were named quotes. Seventeen of these were men, and seven were women. For one term – empathy – all four of these quotes came from women. For the remaining terms, quotes were shared from 17 men and 4 women; of which 3 were the same Anne Frank quote. This shows how virtues themselves are gendered, acting as signifiers to constructions of masculinity and femininity; a division clearly reproduced on Twitter.

WORKPLACE CULTURE

Twitter provides an online space for networking and promotion of individual businesses, and this was reflected in definitions of virtue. Virtues are discussed as necessary traits for employees and employers, if they are to succeed in the workplace. This was particularly the case for describing the qualities of a good leader:

Of the many qualities required for truly great leadership, Humility is underrated the most. This may also be why it is so challenging to change organisations.

We have always believed it is essential to be open about business performance whether celebrating or tightening our belts. The key to a successful workplace culture is Honesty. [link]

Businesses also used claims of virtuousness as an act of self-promotion. This highlights how traits of virtue are still defined by people in a similar way to early philosophical thinkers, as the following quote from the Oxford Dictionary of Philosophy describes:

"A Virtue is a trait of character that is to be admired: one rendering its possessor better, either morally, or intellectually, or in the conduct of specific affairs."

On Twitter, traits of virtue are highly valued characteristics in a person, and often viewed as essential to being a good employee and team member.

A full breakdown of themes coded within each virtue term is included below

TABLE 12.1.

DEFINITIONS OF COURAGE ON TWITTER

Courage - Definitions	**% of sample**
Overcoming fear	28%
Describing someone as courageous	14%
Political courage	12%
Be a leader	10%
Writing is an act of courage	8%
Other	28%

The analysis of the most commonly used definitions of courage on Twitter shows that use on the platform corresponds to the sense in wider use. English dictionaries refer to courage as the ability to act despite the presence of fear, and 28% of coded tweets fell directly into this category. Praising another person's courage was the second most common, closely followed by describing an act of political courage. This highlights how individuals use Twitter to appraise the behaviour of others, often public figures.

Describing writing as an act of courage is interesting here, particularly in the context of Twitter's vast potential audience. When one composes a tweet, they take a personal risk because they cannot control its reach or response; a worry which we potentially see reflected above.

TABLE 12.2.

DEFINITIONS OF EMPATHY ON TWITTER

Empathy - Definitions	% of sample
Argument	22%
Workplace empathy	12%
Feeling another's emotions	12%
Empathy is not innate	6%
Declining empathy	6%
Other	42%

The most common definition of empathy we identified was its use in arguing with another person, or group of people, on Twitter. Most commonly, these were calling someone out for not having empathy. Of all the virtues, the ability to empathise is viewed as essential for understanding how other people behave, therefore describing someone as without empathy implies they are not able to listen to another person's views, effectively stopping arguments in their tracks.

The need for empathy in the workplace and discussion of declining empathy in society highlight its perceived high value as a personal characteristic. Furthermore, the identification of empathy as a learned skill, rather than innate, implies that the acquisition of virtues is believed to be an essential aspect of character development.

One category here, regarding whether empathy is an innate human characteristic, provides some intriguing evidence that classical and modern views of empathy have diverged. It is difficult to make a direct comparison - the specific meaning of 'empathy' as the capacity to imagine oneself in the position of someone else, rather than simply feeling 'sympathy' for their hardships, was developed in the early 20th century.[22] Kant, however, writes of sympathy in 1797 that "it is one of the impulses that nature has implanted in us to do what the representation of duty alone would not accomplish."[23]

In our tweet sample, however, 3 of the 50 tweets analysed argued that empathy was not innate, and must be learned – with two tweets mentioning a professional medical context.

> *"Empathy is something you can't write a prescription for."*
>
> *#doingthebestwecan #chronicpain #crps*

> *"Empathy is not in our genes", New Paper! Preprint here [link] and abstract beneath.*

While this smattering of messages is clearly not nationally representative, these tweets indicate a waning belief that others are naturally empathetic, which may start to make sense of declining levels of trust in society.

TABLE 12.3.

DEFINITIONS OF HUMILITY ON TWITTER

Humility - Definitions	% of sample
Describing someone else's humility	20%
Good leadership requires humility	10%
Being humble	10%
Thinking of yourself less	8%
Describing someone without humility	8%
Other	44%

Definitions of humility on Twitter have stayed close to their original meaning. Thinking less often about yourself and being humble is almost identical to how the Cambridge Dictionary describes humility. However, out of the four virtues we analysed humility was the most likely to be described as something outside of the most used definitions (44% other). Notable amongst these were three tweets were about humility being overrated, and statements of the need for individuals to have an ego in order to be successful. This reflects wider societal changes which have emerged as part of a neoliberal society based on individualistic notions of personal success.

22 See 'Empathy and Sympathy in Ethics", retrieved September 2019 from https://www.iep.utm.edu/emp-symp/#H2
23 Kant, Immanuel The Metaphysics of Morals 6:457

TABLE 12.1.

DEFINITIONS OF HONESTY ON TWITTER

Honesty - Definitions	**% of sample**
Honesty is of value	30%
Honesty is the best policy	20%
Questioning someone's honesty	8%
Describing someone as honest	14%
Other	28%

The second most common definition of honesty in our dataset consists of repurposings of the phrase "Honesty is the best policy." This can be traced back to Sir Edwin Sandy, a politician and colonialist, writing in 1599: 'Our grosse conceipts, who think honestie the best policie' - a proverb whose original meaning has remained unchanged, hundreds of years later, in its use on social media.

SECTION 2
VIRTUOUS ACTION ONLINE

The first section of this report focuses on the ways in which people use virtue language online, as part of everyday language; it examines definitions of virtue, as well as the topics which arise alongside them. This approach, however, ignores the fact that online space is not solely discursive. We have seen that social media encourages people to talk about virtues – crucially, it also gives them the means to act upon them.

The possibility of online virtuous action raises a substantial question – to what extent can behaviour which is confined to online space be considered truly virtuous? One stumbling block here is that we are confined in this analysis to observed behaviour. We cannot use tweets to examine that aspect, so important to Aristotle, of the character and motives of the individual sending them; at least, to do so would certainly be beyond the scope of this study. Some studies, however, have suggested a link between the use of social media and the development of virtues – Vallor et al have contended that SNS may support and strengthen real-life friendship by facilitating reciprocity, self-knowledge, empathy and the shared life.[24] To help answer this question below, we make an effort to connect our putatively virtuous actions with real-life outcomes which could be argued to increase human flourishing; either on behalf of the person sharing the message or a third party with which the message, or fundraiser, is concerned.

We wanted to measure the extent to which social media, and Twitter, in particular, allows people to carry out virtuous acts, using only a small part of human behaviour – the messages people share on Twitter. To do this, we developed three experimental methods to measure virtuous action in three separate senses:

1. *Fundraising and volunteering online*

   This examines the action of sharing links to fundraising sites, and discussion of volunteering, by Twitter users.

2. *Expressions of gratitude*

   This section concerns gratitude in action option, attempting to capture instances of 'genuine' gratitude expressed towards peers, strangers and public figures. It also examines gratitude expressed between companies and their customers online.

3. *Application to learning a skill*

   The final experiment examines the performed virtue of application, and the use of sharing skill development on Twitter as a means of self-motivation in learning and practicing skills.

Measuring behaviour alongside the use of language in this way affords a unique view into the connections between the use of virtue terms and virtuous action. Accordingly, we also look below for a relationship between the two – posing the question of whether familiarity with the language of moral virtues means people are more likely to act virtuously.

24 Vallor, S. (2009). "Social networking technology and the virtues." Ethics and Information Technologies 12: 157-170.

Experiment 1: Fundraising

Social networking is, in 2019, a crowded field. As once novel technologies such as the ability to use hashtags, upload video or control the audience for your messaging have become ubiquitous, platforms have moved towards offering increasingly similar feature sets. As a result, the differentiators between them have become cultural rather than simply technical – more about the people and communities one can reach on a space than the tools that platform offers to enable communication.

One important aspect of Twitter's cultural offering is a radical openness; the fact that anyone on the platform could, in theory, find and read your tweet. This lends the platform an aspect of the soapbox, promising people a public space, and the possibility of an audience for their causes and ideas. This ties to one of the early internet's most utopic ideas, and the premise upon which Facebook makes its argument for being a force for good – in connecting enough people, you enable entirely new forms of social organisation to take place. Communities spring up, campaigns are started, identities form. This has proven to be a powerful social force, if of contested social benefit. Arguably, many of the most turbulent political events, over the last few years, including the Brexit vote in 2016 and the election of Donald Trump, have been stoked by these new social groups, and the growing ability of politicians and businesses to target them precisely according to their interests.[25]

One area in which this ability to gather a crowd can be of unarguable social benefit is in charitable fundraising. Online sites such as justgiving.com allow people to set up cheerful, image-rich webpages through which people can sponsor them in running a marathon, or bathing in spaghetti hoops, and thereby donate to charity. Similar sites, notably GoFundMe and Crowdrise, have sprung up to allow people to raise funds for personal or political projects. These online donations are becoming an increasingly important source of income for charities; the Institute of Fundraising's 2019 benchmark report found that online donations increased over the last 3 years for 68% of the fundraisers surveyed and that 95% of those organisations communicated with their donors through social media (with 88% communicating on Twitter).[26]

To measure the extent and character of this fundraising on Twitter, we used the platform's streaming API to collect 125 thousand English language tweets containing a link to a fundraising campaign, sent from the UK between 1st October 2018 and 12th April 2019.[27]

These tweets broke down as demonstrated in Table 13.

Table 13 shows the percentage of each collection which were not original tweets, but retweets – users sharing someone else's message online. This shows that tweets concerning the two platforms which are primarily focussed around raising money for personal or political motives, rather than for charity or similar causes, are much more likely to be originally sent by a small number of people and then amplified by those sharing their message.

TABLE 13.

TWEETS BY TYPE OF LINK

Term	Tweets	Of which Retweets (%)	Users	Average Tweets per user
Gofundme	61,671	35%	25,314	2.4
Justgiving	59,941	15%	29,962	2.0
BT Mydonate	3,167	19%	1,839	1.7
Crowdrise	239	57%	190	1.3

25 For a discussion of the Leave vote on social media, see Vyacheslav Polonski's analysis in 2016, retrieved Sep. 19 from https://www.referendumanalysis.eu/eu-referendum-analysis-2016/section-7-social-media/impact-of-social-media-on-the-outcome-of-the-eu-referendum/

26 Institute of Fundraising (2019) 'The Status of UK Fundraising - 2019 Report' – retrieved September 2019 from https://hub.blackbaud.co.uk/npinsights/the-status-of-uk-fundraising-2019-report

27 As above, technical issues accessing the API mean that this collection is, unfortunately, not continuous. Rather, it spans three discrete periods: 10/10/2018 to 4/12/2018, 13/1/2019 to 15/2/2019 and 16/3/2019 to 12/4/2019; 127 days of 184 within this period in total.

HOW MUCH IS A TWEET WORTH? THE VALUE OF ONLINE INTERACTION IN FUNDRAISING.

Our hypothesis here is that the action of sharing a link to an ongoing fundraiser is in itself a virtuous act. In theory, by raising the profile of a cause online, you bring that cause to the attention of potential supporters. This attention may be fleeting, and its audience limited, but there are whole industries banking on the fact that it changes behaviour. Fundraising sites, and indeed any campaign trying to squeeze support, information or money from large groups of people, have long counted on social media to act as a 'force multiplier', hoping that the right posts from the right accounts might open the hearts and wallets of a new listening public.[28]

The real value of sharing links in this way is a subject of active debate. The activity has been decried as 'clicktivism', with detractors claiming that participating in online campaigns lulls citizens into a false sense of accomplishment; the instant feeling of righteousness and charity satisfying people's wish to change things before they have taken any action. This is, perhaps, more likely to be true in the act of signing a digital petition than of online fundraising; after all, participating in the latter requires people to actually spend some money. The question here, however, is whether tweeting a link to a campaign to which you may or may not have donated makes a difference to the amount that campaign ends up collecting.

Below, we set out to put some evidence behind this question, and to search for a link between the number of times a campaign is shared on Twitter, and the eventual amount of money raised. To do so, we used Twitter's API to collect 60 thousand links to fundraising pages on JustGiving.com, each of which had been tweeted by at least one user within the UK. We then used JustGiving's public API to collect further data about these campaigns, including the date of the fundraising event they related to, their fundraising target, and the total amount raised through online donations. These were filtered in the following ways:

- To line up with our collection period, we examined only fundraisers for events which had taken place between September 1st 2018 and December 3rd 2018, during our first continuous collection period of tweets.
- Some fundraisers involved campaigns lasting a long time, with fundraising targets in some cases of millions of pounds. As the impact of individual donors on these behemoth fundraisers is likely to be more difficult to measure, we took a more human view of scale: campaigns with targets of above £10,000 were removed.
- Any campaigns not raising money in GBP were also removed.

This filtering process left us with 7.9 thousand pages on justgiving.com, each relating to a single fundraising campaign.

We then built a linear regression model to explore the relationship between the number of tweets sent linking to campaigns in the final months before the event, and the eventual amount which that campaign raised.

This statistical analysis suggested that tweets sharing links do indeed have a positive effect on the amount a fundraising campaign eventually raises, with each additional tweet linking to a campaign, the model predicts that campaign will receive around £56 of extra funding. While this relationship is statistically significant ($p<0.001$) the tweets themselves play only a very minor role in fundraising – the number of tweets sent explains a very low percentage (around 2%) of the eventual amount donated. This makes sense; we are using a simple heuristic here, and our model does not take into account the effect of the fundraiser's focus or goal, the profiles of the people publicising it or, indeed, promotion undertaken on any other social network. What this does show, however, is a connection – relatively weak and unpredictable, but nevertheless present – between posting about charity and the charitable act of donation.

28 See for example this recent article from 'Charity digital news'; one of many similar guides to fundraising on social media: https://www.charitydigitalnews.co.uk/2019/10/09/7-tips-on-leveraging-social-media-for-fundraising-campaigns/

Experiment 2: Gratitude

An ostensible benefit of the utopian power of social media to connect people -or at least allow them to ping red notification dots to each other's pockets - is the ability it gives members of the public to connect with people in public life. Large corporations, political representatives and law enforcement are increasingly expected to have an active online presence; and absorb (and sometimes respond to) the enquiries and criticisms sent their way. The availability of public figures online has in many cases exposed them to serious harassment and abuse. In 2017, a study by Demos found that one in twenty tweets sent to British MP's Twitter accounts in a four-month period were abusive, with some MPs receiving abuse in 60% of the tweets which mention them.[29] A follow up report authored with RUSI found that this form of abuse can result in serious psychological harm for those targeted, as well as having 'a wider cumulative impact on societal stability.'[30]

Less well studied, however, is the extent to which social media is used to express gratitude, both to those in the public eye as well as peers and strangers. Gratitude is important - the act of feeling and expressing thanks has been associated with increased subjective well-being, as well as strengthening social bonds and encouraging pro-social behaviour.[31] In Britain, it's also something we feel that we lack. The Jubilee Centre for Character and Virtues found in 2015 that 80% of British citizens surveyed felt there was a lack of gratitude in society, and 78% want to see more effort put into its promotion in workplaces and schools.[32] Below, we examine uses of terms to express thanks on Twitter, attempt to find the volume of this which express 'genuine' gratitude, and examine the popular figures who tend to be the subject of gratitude online.

In order to study this phenomenon, Twitter's streaming API was used to collect tweets containing a series of generic terms which can be used to express gratitude, especially in a personal context - 'thank you,' 'grateful for', 'made my day' etc. As above, this collection was filtered to remove non-English language tweets, as well as tweets sent from users not likely to be in the UK.

On examining this initial dataset, we found that many of the phrases used in our collection were being employed in ways which arguably did not constitute an act of gratitude. This included sarcasm ('thanks for nothing') and use of terms as a conversational nicety ('Thanks for confirming my point'). Accordingly, a classifier was trained to identify tweets which expressed 'genuine gratitude'. This included people thanking others for their support or actions, gratitude to celebrities and organisations for their work, and gratitude expressed towards customers by businesses, charities and organisations.

Deciding whether a 280 character message constitutes a 'genuine' act of gratitude is a non-trivial task for humans, let alone for an algorithm, and of all the classifiers trained in the course of this research, this one was the hardest to train and achieved the lowest overall accuracy (71%) However, we found that the classifier was useful in removing much of the irrelevant, phrasal discussion outlined above. A full description of this classifier, along with examples of tweets falling into each class, in included in the methodological annex to this document.

As mentioned above, a notable proportion of gratitude in our dataset was from organisations towards their customers and supporters. Since this is behaviour is arguably driven by commercial rather than moral imperatives, a further classifier was trained to identify and filter out this content. To further remove this discussion, tweets sent from accounts describing themselves as organisations rather than individuals were also classified and removed. The full classification pipeline applied to this dataset is illustrated in Figure 14.

29 Krasodomski-Jones, A. (2017) 'Signal and Noise', Demos

30 Babuta, A., Krasodomski-Jones, A. (2018) 'The Personal Security of Individuals in British Public Life', RUSI

31 For links between gratitude and increased subjective wellbeing, see Emmons, R. A., and McCullough, M. E. (2003) 'Counting Blessings Versus Burdens: An Experimental Investigation of Gratitude and Subjective Well-being in Daily Life', Journal of Personality and Social Psychology, vol. 84, no. 2, pp. 377–389.;

Froh, J. J., Sefick, W. J. and Emmons, R. A. (2008) 'Counting Blessings in Early Adolescents: An Experimental Study of Gratitude and Subjective Well-being', Journal of School Psychology, vol. 46, no. 2, pp. 213–233.;

Watkins, P. C., Woodward, K., Stone, T. and Kolts, R. L. (2003) 'Gratitude and Happiness: Development of a Measure of Gratitude, and Relationships with Subjective Well-being', Social Behavior and Personality, vol. 31,no. 5, pp. 431–451.

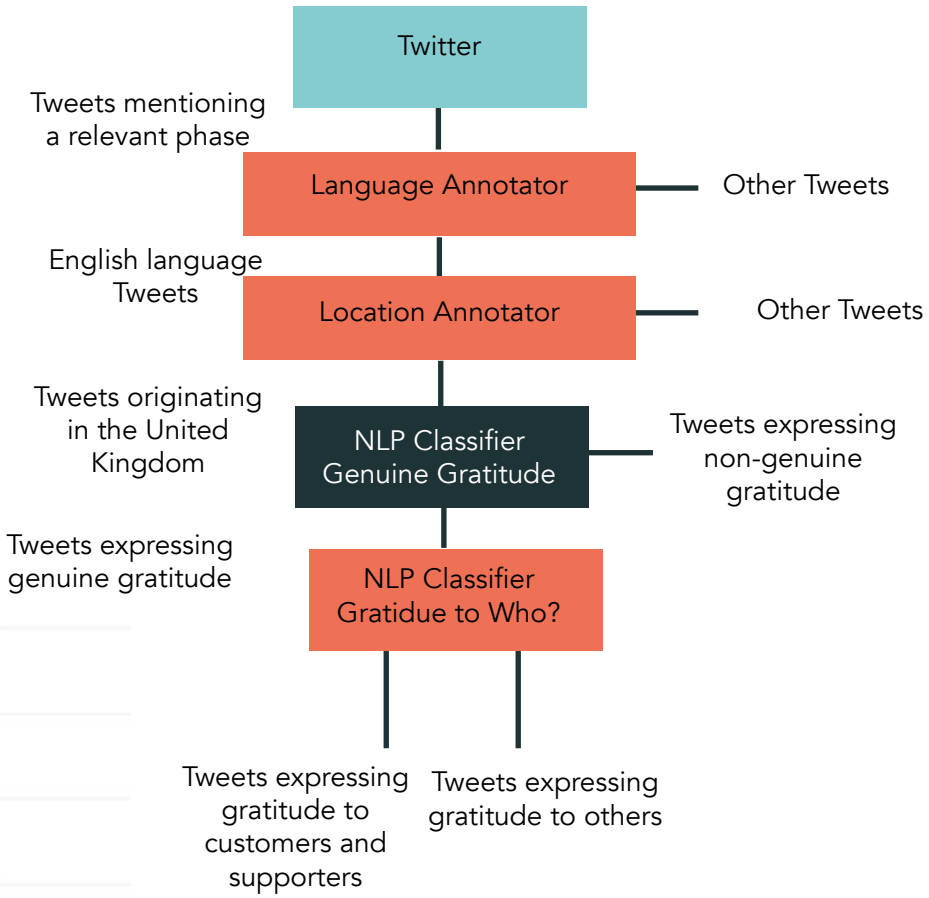
For links to pro-social behaviour, see Grant, A. M. and Gino, F. (2010) 'A Little Thanks Goes a Long Way: Explaining Why Gratitude Expressions Motivate Prosocial Behavior', Journal of Personality and Social Psychology, vol. 98, no. 6, pp. 946–955.

Algoe, S. B., Haidt, J. and Gable, S. L. (2008) 'Beyond Reciprocity: Gratitudeand Relationships in Everyday Life', Emotion, vol. 8, no. 3, pp. 425–429

32 Arthur, J., Kristjansson K., Gulliford L., Morgan B. (2015) 'An Attitude for Gratitude - How Gratitude Is Understood, Experienced and Valued by the British Public.' Jubilee Centre for Character and Virtues References from the above endnote are also taken, with thanks, from this report.

FIGURE 14.

CLASSIFIER PIPELINE USED TO LABEL GRATITUDE



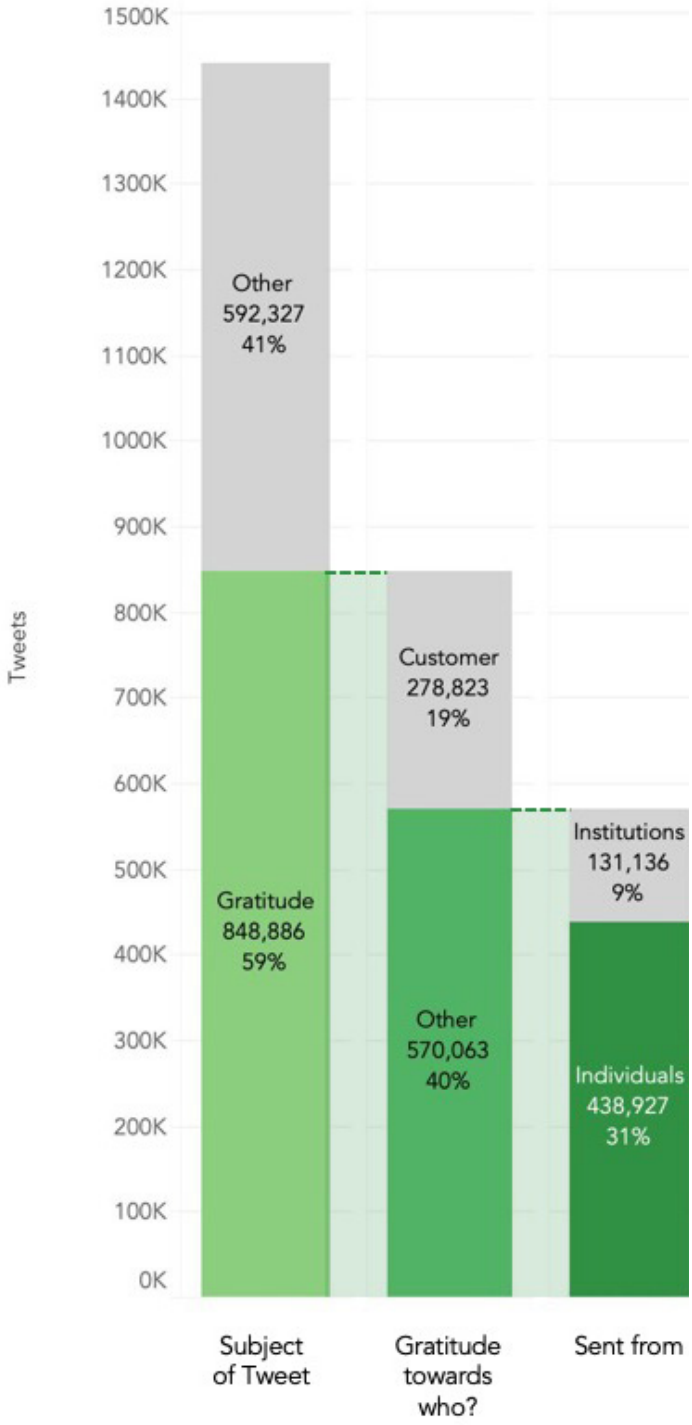


FIGURE 15.

BREAKDOWN OF TWEETS IN 'GRATITUDE' COLLECTION

Of all the experiments conducted during this project, this involved by far the largest collection of data, with 1.4 Million tweets sent using a relevant phrase from the UK, sent by 543,126 Twitter users likely to be from the UK. As shown in Figure 15, the majority of these (59%) expressed 'genuine' gratitude, with a substantial minority (40%) of all tweet thanking people other than customers.

This filtering process allowed us to concentrate on the expressions of gratitude we're most interested in here – the 31% of tweets which expressed genuine gratitude, not directed at a customer, and not sent from an account owned by an institution or organisation. It is this third of the dataset that we concentrate on below.

WHO GETS THANKED ON TWITTER?

This filtering exposed something intriguing – people on Twitter like to thank each other 'in person', by including another user's @screenname in the body of their tweet. An overwhelming 93% of those 440 thousand tweets, sent by an individual expressing gratitude to others, contained another user's screenname. This not only helps specify exactly who the tweet is talking about, but also brings that account into the conversation, causing your tweet to appear on their Twitter homepage and, under default settings, a notification to be sent to their account.

TABLE 16.

TYPES OF ACCOUNTS MENTIONED IN GRATITUDE TWEETS

Account Type	Tweets	% of top 100 mentions
Popstar	35,549	24.8%
Sports	27,572	19.3%
Politics	19,968	14.0%
Influencer	15,581	10.9%
Company	10,574	7.4%
Media	5,714	4.0%
???	5,478	3.8%
Charity	4,834	3.4%
Journalism	4,749	3.3%
Animals	3,275	2.3%
NHS	3,085	2.2%
Actor	2,998	2.1%
Individual	2,194	1.5%
Video Games	806	0.6%
Writer	712	0.5%

To investigate these further, we selected the 100 accounts most often mentioned in gratitude tweets, which were then annotated by hand to establish the type of account they represented – politicians, for example, or sports personalities. Table 16 shows the number of tweets corresponding to each type of account. Heading up this list are the people we might expect to be most popular in pop culture – music celebrities and those associated with the business receive the most gratitude, followed by sports stars. Interestingly, given the amount of abuse these accounts also receive, prominent accounts belonging to politicians and political parties were thanked in 14% of tweets mentioning a top 100 account.

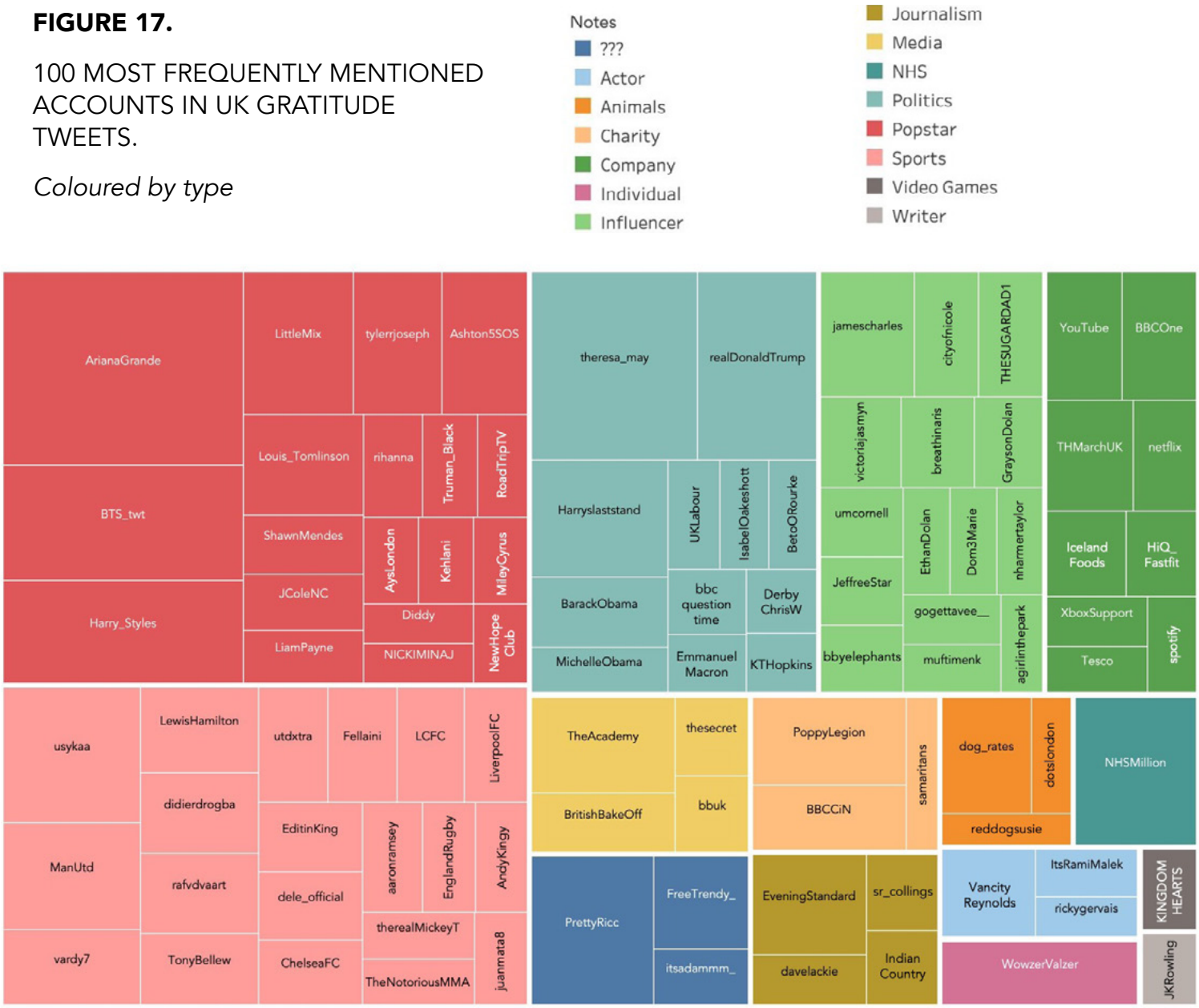
Figure 17 lists the names of each of these oft-thanked accounts, which are sized by the number of tweets expressing gratitude which they appear in, presenting us with a view of the most valued individuals and organisations in modern Britain, or at least on modern British Twitter. Much of this gratitude, of course, is situated in the world of social media. Alongside familiar names from global politics (and sometimes eclipsing them in volume) the 'influencers' category contains a new type of celebrity, whose notoriety stems entirely from the videos and images they post on social media. This is in some sense reassuring - these individuals, along with musicians like Ariana Grande, have substantial followings of people tuned to their every message, and the conversation surrounding them is likely to be particularly influential in determining the moral tone of discussions on social media.

This graph also shows that, while charities and institutions represent a small percentage of the overall category breakdown, accounts like @NHSMillion, which exists explicitly to allow people to express gratitude towards the National Health Service, is particularly often mentioned, with @PoppyLegion and the Samaritans also appearing in the top 100 most thanked accounts. The fact that these accounts are overshadowed by celebrities reflects perhaps on the UK's societal priorities, but also on the fact that these individuals tend to be particularly active on social media.

FIGURE 17.

100 MOST FREQUENTLY MENTIONED ACCOUNTS IN UK GRATITUDE TWEETS.

Coloured by type



Experiment 3: Application

Our final performed virtue is perhaps the most unusual, but certainly the most direct. It focuses on people using social media as a way to help them apply themselves to tasks, and share their achievements. To do this effectively, people first need an audience – ideally one interested in the domain in which they've made progress. One aspect of social media which is critical in attracting these specific audiences is through using a 'hashtag', a method for #labelling and #organising content which originated on Internet Relay Chat in 1988.[33] After its reincarnation on Twitter, hashtags have become near universally adopted across social media platforms. Since including a hashtag in a tweet will allow anyone searching for that term to find and view it, their use can been seen as an attempt by the sender to better tap into Twitter's global intended audience.

Below, we study one specific form of hashtag use; as a daily motivator to practice and perfect a talent. We examine the use of 11 hashtags – for example, #100daysofart - intended to be posted daily by Twitter users as they apply a skill, thereby enacting the performative virtues of application and perseverance. Posting one of these hashtags, often alongside a photo or video showing progress, acts as a public statement of intent; an implicit commitment to see the 100 or 30-day period through, under the gaze of a putatively interested online public.

In total, 7,923 tweets using one of the 11 hashtags selected was collected in the year long period between 1st July 2018 and 1st July 2019. This is obviously a tiny dataset compared to those collected above, but is not negligible – on average, it represents 22 tweets sent each day. As we were primarily interested in studying tweets sent by real people applying themselves to a programme of practice, this collection was filtered in two ways. Firstly, any tweets consisting only of retweets of another user's content were removed. Secondly, all tweets were classified to establish whether the account sending them was likely to be a corporate or institutional account, rather than one run by an individual, and non-institutional tweets removed. After this filtering, we were left with 4,956 tweets (63%) sent by 432 individuals in the UK.

The collection focused on a range of skills, from generic terms like '#100dayspractice' to highly specific exercises, such as '#300kettlebellswingsfor30days'. Some terms are highly popular, their implicit challenge taken up by hundreds of people in the UK. Others only garnered a handful of tweets, or in the case of the ill-fated '#100squatschallenge', just one. What makes these hashtags particularly interesting in the context of this research is their location in the intersection between posting on social media and taking action. As long as people are honestly playing the game, these tweets provide often documented evidence of real-world French practice, sketches and swung lumps of iron.

TABLE 18.

APPLICATION HASHTAGS IN COLLECTION

Hashtag	Tweets	Users	Average Tweets per user
Total	4,956	432	11.5
#30daychallenge	2,095	241	8.7
#100daysofpractice	1,485	73	20.3
#100daychallenge	973	66	14.7
#30dayschallenge	153	33	4.6
#100dayschallenge	152	22	6.9
#100daysofexercise	40	6	6.7
#300kettlebellswingsfor30days	36	4	9.0
#100daysofrunning	35	6	5.8
#100daysofart	25	6	4.2
#100daysofanimation	23	3	7.7
#100daysofbirds	3	1	3.0
#100squatschallenge	1	1	1.0

33 Salazar, E. (2017) Hashtags 2.0 - An Annotated History of the Hashtag and a Window to its Future, Icono 14, volumen 15 (2), pp. 16-54. doi: 10.7195/ri14. v15i2.1091

A testimony to the motivational effect which these hashtags can have was provided by a violinist, who used #100daysofpractice to tweet regular videos of her practicing pieces, scales and tricky musical passages. In a video explaining her motivation for the project, she explains that the videos themselves were useful, once made, in identifying areas which she needed to focus on. However, the reaction of people to her videos was also key in helping her to keep the project going:

> *"I didn't think people would be impressed by watching me practice… but once I started doing it, it felt like there was a community of people practicing, and it wasn't so solitary anymore."*

This speaks to the power of Twitter to enable people to reach for encouragement to the world in general, and for the response to encourage personal development and application. This effect is studied in more detail below.

#6andabitdaysofpractice – PERSEVERANCE IN APPLICATION HASHTAGS

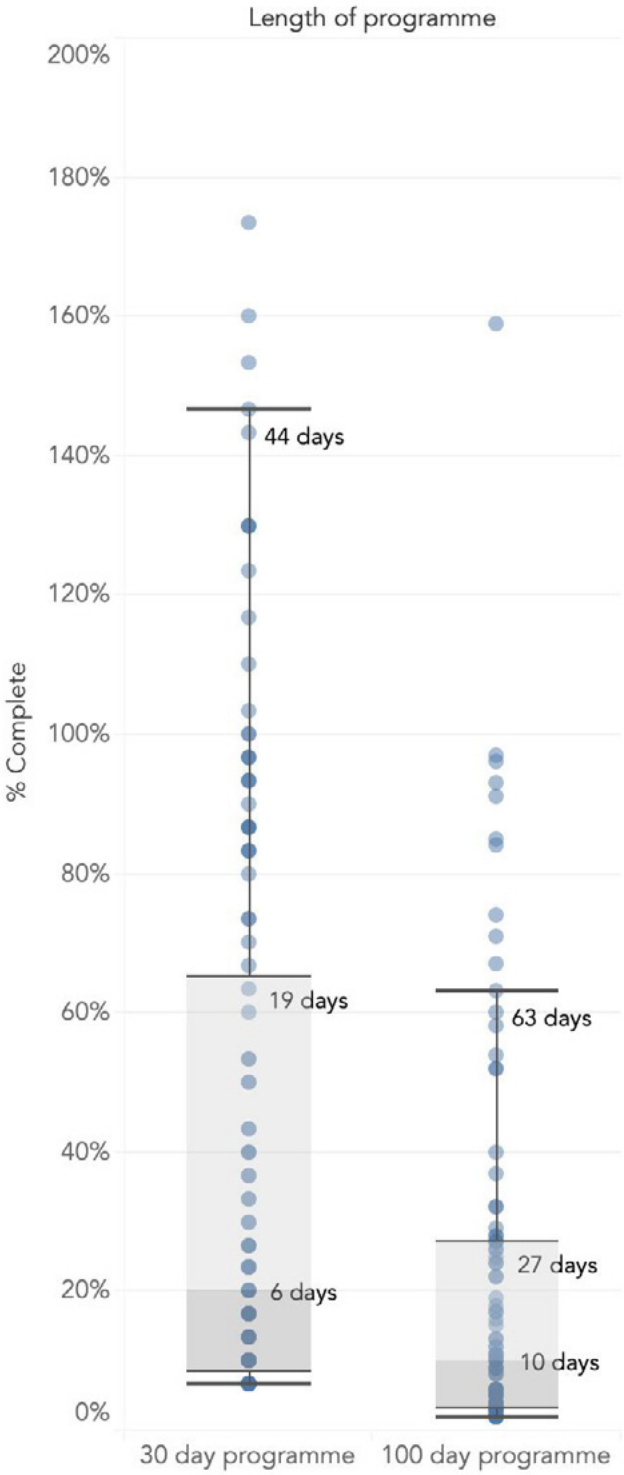
In order to measure how effective social media hashtags were in encouraging people to learn a skill, we counted the number of (not necessarily contiguous) days in which each user persevered to a hashtag by sending a tweet containing it. In order to remove users who were commenting on, rather than taking part in, one of these courses, 212 users who only posted on a single day were excluded from this analysis (49% of the dataset)

Figure 19 shows the distribution of how long people tended to persevere with their committed programmes. Each dot above represents one or more users, with position on the Y-axis reflecting the number of days for which that user persevered with a relevant hashtag. The box-and whisker plots show quartiles of the data, and the location of the median point - a quarter of the accounts in our dataset lie in each of the shaded grey areas, and the line between these shaded areas shows the median number of days. The whiskers, extending out from either side of the box, represent 1.5 times the shaded range between the first and third quartiles (the IQR) – points lying outside of these whiskers can be considered outliers.

Figure 19 shows that most people stop posting on a 30-day hashtag after 6 days, and a 100-day hashtag after 10. Those that make it past this point, however, are likely to persevere for around three times as long, with the next quartile extending to 19 and 27 days for 30- and 100-day courses respectively. A handful

FIGURE 19.

TWITTER USERS BY PERCENTAGE OF PROGRAMME COMPLETED



of people also go beyond the implicit promise of their hashtags, and post for longer than the length of the programme. A sample of tweets from these eager participants shows some people taking up a second 30 day challenge having completed their first, or, in one case, using '#100daysofpractice' and '#365daysofpractice' in the same tweet in an attempt to get more exposure for their son's year of playing the piano.

These data show that motivating hashtags are being used, and persevered with, by a more or less committed group of users on Twitter. Although a relatively small number of users made it through an entire programme – overall, 14 users (6% of those tweeting for more than one day) completed their courses, and 77 (36%) stuck to them for more than two weeks – for some at least, posting on social media seems to be a useful motivator in learning a skill.

#100daysofpeersupport – THE INFLUENCE OF OTHERS ON PERSEVERANCE

The analysis above ignores a crucial part of the experience of Twitter – the second to second feedback to messages gained from followers, likes and comments. In order to investigate the influence which the reaction of a Twitter user's wider social circle has on their likelihood to persevere with their practice, we looked for a relationship between the number of likes and retweets someone received during their first five days of posting, and the length of time for which they continue to use that hashtag. A scatter graph showing these two measures is shown in Figure 20.

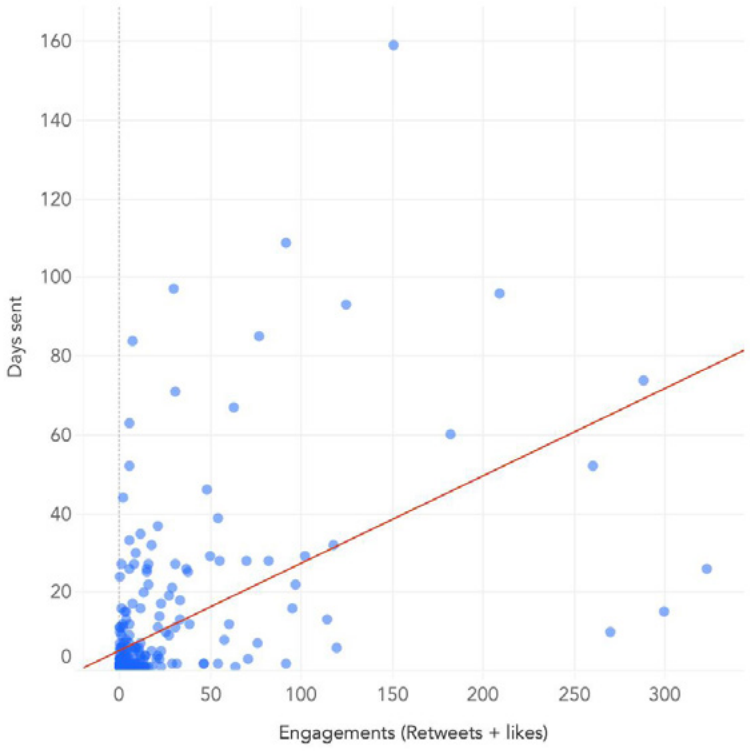
A linear model was constructed to investigate this relationship. It predicts that there is a positive, statistically significant ($p<0.001$) relationship between the number of engagements, constituting likes and retweets, received in the very first days of a user committing to a programme, and the number of days they keep using that hashtag. The model predicts that for every ten extra engagements received in those crucial first days, people will spend just over 2 more days posting their practice online.

There are clearly a number of influential factors which we do not account for here – this is eloquently demonstrated in Figure 20 by the distance many points lie from the model's line of best fit. The effect which early engagement has, however, is not insignificant – it accounts for 28% of the total variance displayed in the dataset.

This relationship between application and feedback online is interesting for two reasons. Not only does this suggest that social media can be useful in encouraging people to learn and perfect new skills, but it also shows that encouraging others in their endeavours to apply themselves, even if only by clicking 'like' on one of their posts, is likely to have an effect on their success. There is evidence, then, to suggest that encouragement on social media may in itself be a virtuous act.

FIGURE 20.

VOLUME OF ENGAGEMENTS (RETWEETS + LIKES) RECEIVED BY A TWITTER USER IN THEIR FIRST FIVE DAYS OF A PROGRAMME AGAINST A PERSEVERENCE WITH THE PROGRAMME



Offline links – evidencing the connection between virtuous speech and virtuous action

In the course of this project, we have developed methods for analysing virtue online, in speech and in action. We do not yet know, however, whether these expressions of morality are linked. Below, we investigate whether someone's proclivity for using virtue terms online affects their likelihood to conduct virtuous actions on social media.

To examine this, we first assembled a dataset roughly representative UK Twitter users at large. To this end, Twitter's 1% random sample endpoint was used to collect a large sample of tweets – this was filtered this down to those sent from 3,688 random users likely to be tweeting from the UK. We then collected all tweets sent by these users during our collection period, and counted the number of times each user has used a virtue term, shared a link to a fundraising campaign, or sent a tweet expressing gratitude. (Due to the relatively low percentage of the population using one of our 11 'application' hashtags, this action was not investigated here.)

The resulting data was subjected to a statistical analysis to determine whether there was a link between how often someone talks about moral virtues in a non-neutral sense, and how often they perform the virtuous actions of gratitude and sharing links to fundraisers online.

To this end, two separate multiple linear regression models were trained in an effort to establish the relationship between virtue language and acts of gratitude, and virtue language and sharing of fundraiser links. In order to make each model as meaningful as possible, the following filtering processes were applied to the data:

- Tweets containing virtue terms were filtered to remove 'neutral' uses, which are more likely to indicate conversational use ('in all honesty' etc) than positive or negative tweets.
- Tweets expressing gratitude were filtered to remove those not classified as expressing 'genuine' gratitude. Similarly, those containing a fundraising term were filtered to remove any tweets not classified as relevant to fundraising.
- 'Virtuous action' tweets which also contained a virtue term - which thanked people for their honesty, for example – were removed from the dataset.
- Intuitively, we might expect people who use Twitter more often in general to send more virtue related tweets of all types. To control for this, a count was taken of the number of tweets sent by each user, on any subject, during our collection period. This variable was included, and thus controlled for, during each regression.
- Clear outliers in each case were inspected. Where these were judged highly likely to be automated accounts they were removed from the dataset. For example, an account tweeting every hour to promote the works of a romance novelist was excised from the dataset. In total, four outlying accounts were removed in this way.

The distribution of virtue tweets against these two behaviours are shown in in Figures 21.1 and 21.2 on the following page. As in Figure 20 above, these show that the relationship in both cases is fairly weak. Again, this makes sense – it is clearly possible for people to use Twitter to express thanks, or to fundraise, without habitually using one of our four virtue terms. The model for each of these relationships, however, suggests that there is a statistically significant positive relationship ($p < 0.001$ in both cases) between use of virtue language and both types of behaviour. The models predicts that for every 10 more tweets using virtue language person sends, they are likely to send 5 more tweets genuinely thanking people, and 1 more tweet about fundraising. Full results for all of the regressions carried out in this paper are included in the methodological annex.

Clearly, neither relationship here was going to be straightforwardly correlative; there are many factors which will affect people's likelihood to thank each other and share links to charities, entirely independent of their likelihood to discuss morality online. We see this in the model too - in each case, this relationship, though significant, is fairly weak, with virtue term use accounting for 10% of the variance seen between virtue language and gratitude, and 5% between virtue language and fundraising. What this analysis does tell us, however, is that familiarity with virtue terms does make a difference - there are gentle links, however gentle, between people's use of virtue terms on Twitter and their likelihood to perform virtuous actions on the platform. In general, the more virtue is discussed, the more it is enacted.

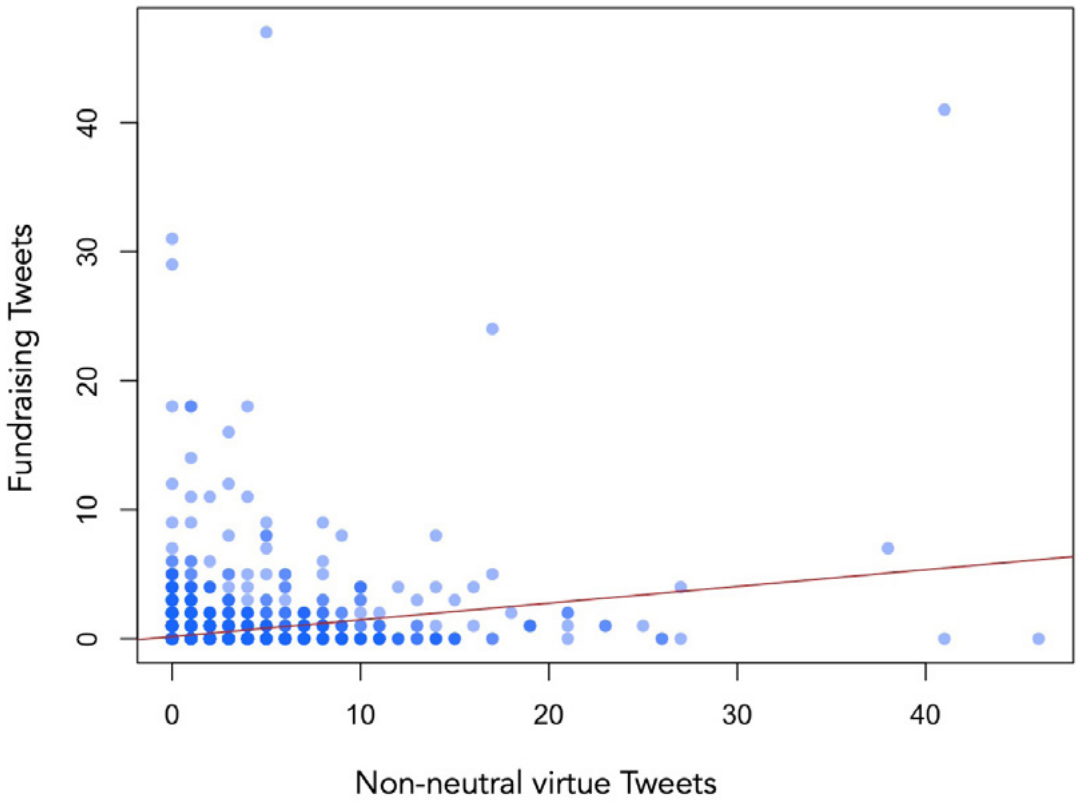
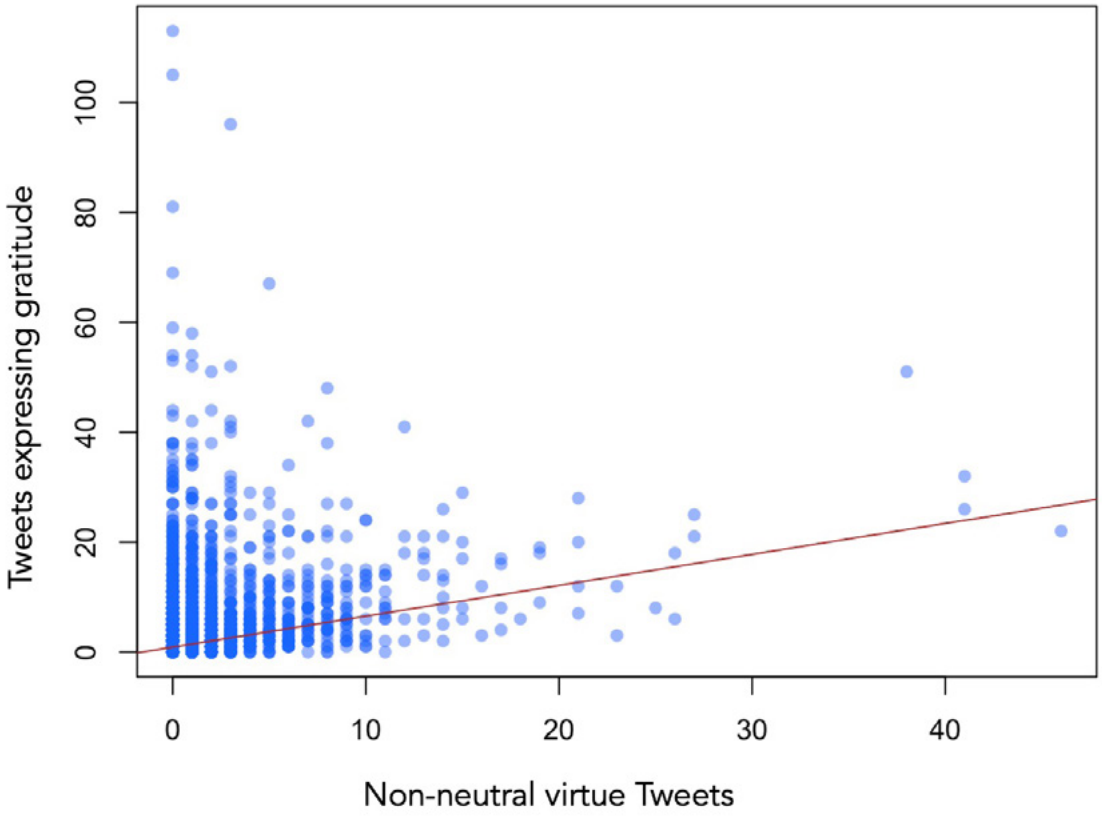


FIGURE 21.1. (Above)

NON-NEUTRAL VIRTUE TWEETS AGAINST LINKS TO FUNDRAISERS SHARED ONLINE

FIGURE 21.2 (Below).

NON-NEUTRAL VIRTUE TWEETS AGAINST EXPRESSIONS OF GRATITUDE ONLINE



APPENDIX 1

Results of regressions

CHARITY

Linear regression analysis was used to test if the number of tweets sent linking to a fundraiser on justgiving.com significantly predicted the eventual amount in pounds sterling that fundraiser raised. The results of the regression indicated the predictor explained 2.1% of the variance (R^2 =.0207, $F(1,7909)=167.1$, $p<0.0001$). It was found that tweets containing links strongly predict the total amount raised (β = 55.84, $p<.0001$), though the significance of this should be interpreted in light of the very low explanatory power of this relationship.

APPLICATION

Linear regression analysis was used to test if the number of engagements – defined as the sum of retweets and likes received on a given tweet – received by a user during their first five days of posting an application hashtag – had a significant effect on the number of days for which that user continued to use that hashtag in tweets. The results of the regression indicated the predictor explained 28% of the variance (R^2 =.2770, $p<0.0001$). It was found that engagements positively predict days of perseverance (β = 0.2134, $p<.0001$)

VIRTUE TERMS

Multiple regression analysis was used to test if the number of tweets containing positive or negative language around virtue sent by a Twitter, as well as the number of tweets that user sent overall, significantly predicted the number of tweets sent expressing gratitude. The results of the regression indicated the two predictors explained 9.7% of the variance (R^2 =.0967, $F(2,3659)=195.8$, $p<0.0001$). It was found that virtue terms sent significantly predicted gratitude (β = .56, $p<.0001$), but that number of tweets sent overall had almost no effect (β = 0.004, $p<.0001$).

We also tested the effect of use of virtue terms and overall tweet volume on people sharing tweets containing links to online fundraisers. In this case indicators predicted 4.8% of the variance observed (R^2 = .0484, $F(2,3659)=93.01$, $p<0.0001$). It was found that virtue terms sent significantly weakly predicted (β = .13, $p<.0001$), and that again number of tweets sent overall had almost no effect (β = 0.0001, $p<.5$).

Natural Language Processing and the NLP Classifier

Building algorithms to categorise and separate tweets forms an important part of the research method for this paper. This responds to a general challenge of social media research: the data that is routinely produced and collected is too large to be manually read.

Natural language processing classifiers provide an analytical window into these kinds of datasets. They are trained by analysts on a given dataset to recognise the linguistic difference between different kinds of data, in this case between tweets. This training is conducted using a technology called 'Method 52', developed by the project team to allow non-technical analysts to train and use classifiers. These were built using Method 52's web-based user interface to proceed through the following phases:

PHASE 1: DEFINITION OF CATEGORIES.

The formal criteria explaining how tweets should be annotated is developed. Practically, this means that a small number of categories – between two and five – are defined. These will be the categories that the classifier will try to place each (and every) tweet within. The exact definition of the categories develops throughout the early interaction of the data. These categories are not arrived at a priori, but rather iteratively, informed by the researcher's

interaction with the data – the researcher's idea of what comprises a category will often be challenged by the actual data itself, causing a redefinition of that category. This process ensures that the categories reflect the evidence, rather than the preconceptions or expectations of the analyst. This is consistent with a well-known sociological method called 'grounded theory'.

PHASE 2: CREATION OF A GOLD-STANDARD TEST DATASET

This phase provides a source of truth against which the classifier performance is tested. A number of tweets (usually 100, but more are selected if the dataset is very large) are randomly selected to form a gold standard test set. These are manually coded into the categories defined during Phase 1. The tweets comprising this gold standard are then removed from the main dataset, and are not used to train the classifier.

PHASE 3: TRAINING

This phase describes the process wherein training data is introduced into the statistical model, called 'mark up'. Through a process called 'active learning', each unlabelled tweet in the dataset is assessed by the classifier for the level of confidence it has that the tweet is in the correct category. The classifier selects the tweets with the lowest confidence score, and these are presented to the human analyst via a user interface of Method52. The analyst reads each tweet, and decides which of the pre-assigned categories (see Phase 1) that it should belong to. A small group of these (usually around 10) are submitted as training data, and the NLP model is recalculated. The NLP algorithm then looks for statistical correlations between the language used and the meaning expressed to arrive at a series of rules-based criteria, and presents the researcher with a new set of tweets which, under the recalculated model, it has low levels of confidence for.

PHASE 4: PERFORMANCE REVIEW AND MODIFICATION

The updated classifier is then used to classify each tweet within the gold standard test set. The decisions made by the classifier are compared with the decisions made (in Phase 2) by the human analyst. On the basis of this comparison, classifier performance statistics – 'recall', 'precision', and 'overall' (see 'assessment of classifiers', above) - are created and appraised by a human analyst.

PHASE 5: RETRAINING

Phase 3 and 4 are iterated until classifier performance ceases to increase. This state is called 'plateau', and, when reached, is considered the practical optimum performance that a classifier can reasonably reach. Plateau typically occurs within 200-300 annotated tweets, although it depends on the scenario: the more complex the task, the more training data that is required.

PHASE 6: PROCESSING

When the classifier performance has plateaued, the NLP model is used to process all the remaining tweets in the dataset into the categories defined during Phase 1, using rules inferred from data the algorithm has been trained on. Processing creates a series of new data sets – one for each category of meaning – each containing the tweets considered by the model to most likely fall within that category.

PHASE 7: CREATION OF A NEW CLASSIFIER (PHASE 1), OR POST-PROCESSING ANALYSIS (PHASE 8)

Practically, classifiers are built to work together. Each is able to perform a fairly simple task at a very large scale: to filter relevant tweets from irrelevant ones, to sort tweets into broad category of meanings, or to separate tweets containing one kind of key message with those containing another. When classifiers work together, they are called a 'cascade'. Cascades of classifiers were used for both case studies. After Phase 7 is completed, a decision is made about whether to return to Phase 1 to construct the next classifier within the cascade, or, if the cascade if complete, to move to the final phase - post-processing analysis.

PHASE 8: POST PROCESSING ANALYSIS:

After tweets have been processed, the new datasets are often analysed and assessed using a variety of other techniques.

CLASSIFIER PERFORMANCE

No NLP classifier used on this scale will work perfectly, and a vital new coalface in this kind of research is to understand how well any given algorithm performs on various measures, and the implications of this performance for the research results. Each classifier trained and used for this paper was measured for accuracy. In each case, this was done by:

1. Randomly selecting 100-300 tweets to comprise a 'gold standard'.
2. Coding each of these tweets by hand, conducted by an analyst.
3. Coding each of these tweets using the classifier.
4. Comparing the results and recording whether the classifier got the same result as the analyst.

There are three outcomes of this test. Each measures the ability of the classifier to make the same decisions as a human in a different way:

Recall

Recall is a measure of the correct selections that the classifier makes as a proportion of the total correct selections it could have made. If there were 10 relevant tweets in a dataset, and a relevancy classifier successfully picks 8 of them, it has a recall score of 80%.

Precision

Precision is a measure of the correct selections the classifier makes as a proportion of all the selections it has made. If a relevancy classifier selects 10 tweets as relevant, and 8 of them actually are indeed relevant, it has a precision score of 80%.

Overall F-score

The 'overall' score combines measures of precision and recall to create one, overall measurement of performance for the classifier. All classifiers are a trade-off between recall and precision. Classifiers with a high recall score tend to be less precise, and vice versa.

The scores obtained for each classifier used in this research are listed below:

TABLE 22.1.

CLASSIFIER 1A: VIRTUE LANGUAGE - INSTITTUTIONS VS OTHER

Label	Precision	Recall	F-Score
Institutions	0.759	0.647	0.698
Other	0.831	0.894	0.861
Overall accuracy			0.81

TABLE 22.2.

CLASSIFIER 1B: VIRTUE LANGUAGE - NEUTRAL VS OTHER

Label	Precision	Recall	F-Score
Neutral	0.459	0.531	0.493
Non-Neutral	0.867	0.831	0.848
Overall accuracy			0.767

TABLE 22.3.

CLASSIFIER 1C: VIRTUE LANGUAGE - SENTIMENT

Label	Precision	Recall	F-Score
Positive	0.806	0.763	0.784
Negative	0.707	0.756	0.73
Overall accuracy			0.76

TABLE 23.

CLASSIFIER 2: FUNDRAISING - RELEVANCE

Label	Precision	Recall	F-Score
Volunteering	0.635	0.783	0.701
Non-Volunteering	0.635	0.763	0.813
Overall accuracy			0.77

TABLE 24.1.

CLASSIFIER 3A: GRATITUDE - GENUINE VS OTHER

Label	Precision	Recall	F-Score
Gratitude	0.714	0.724	0.719
Non-Gratitude	0.712	0.703	0.707
Overall accuracy			0.713

TABLE 24.2.

CLASSIFIER 3B: GRATITUDE - EXPRESSED TO WHOM

Label	Precision	Recall	F-Score
Customer	0.643	0.643	0.643
Non-Customer	0.861	0.861	0.861
Overall accuracy			0.8

CLASSIFIERS – DESCRIPTION AND EXAMPLES

To train each of the classifiers described above, researchers made a series of decisions as to the types of tweets which fell into each category – for example, whether tweets mentioning large supermarket chains should be labelled as 'institutions' (they were).

To explain these decisions, we outline below examples of categories which were considered relevant to each label when building each classifier, along with a small number of illustrative tweets. These are designed to give concrete examples of the decisions our classifiers are trained to make; however, these categories are not exhaustive, and there will be other discussions at play within each of these labels.

In order to protect the anonymity of the authors of these tweets, each of those below has been 'bowlderised' – the wording of the tweet has been altered in a way which preserves the meaning. Links and usernames have also been removed.

Classifier 1a: Virtue language – Institutions / other

This classifier applied two labels to tweets: 'Institutions' and 'Other'. Examples of the topics contained within these labels are below:

'Institutions'

This label captured tweets discussing institutions in public life, including political actors and parties, large corporations and the press. Themes included:

Discussion of large companies and brands

> *Good on Nike. Supporting Colin Kaepernick when most brands lack political courage altogether is something generations in the future will remember as a part of the company's legacy. Being on the right side of history is bigger than the bottom line.*

Honesty, courage etc (or lack thereof) of the press

> *what a state our mainstream media has sunk into. on crucial issues for a free press, honesty, fairness, integrity and balance even the Murdoch's sky outscores bbc*

Politics and politicians

> *[username] [username] [username] [username] [username] [username] Your faulty logic is being corrected here - that's not abuse, irony or stupidity. You're trotting out discredited Project Fear britnat lines against Pro-indy scotland cos you've been caught out. Have some humility - apologise!*

> *I know Twitter does not cope well with nuance but on #McCain: Surely we can acknowledge his courage personally, his civility and integrity without ignoring his US exceptionalism, Russophobia, myopic conservatism and dedicated warmongering.*

Countries and geopolitics

> *They have the courage to claim Koreans are being brainwashed but Americans were fed enough orientalist myths to justify dropping atomic bombs on Japan.*

Social media and tech companies

> *Silicon Valley changed a lot from its 1970s socialist hippy origins and has mutated to a unsavoury Ayn Rander cabal fed on podcasts by the Daily Stoic. The infosphere is shaped by Jack and Zuck, robots with all the empathy of a T-1000© and the moral fortitude of a Hedonismbot.*

'Other:'

All tweets not mentioning an institution as defined above.

Conversations between users

> *[username] [username] But I have presented you with facts about being employed and black in the UK? Anyway, your excuses do not invalidate the conclusions in the article, they just highlight your lack of empathy and comprehension skills.*

Praise of individuals and movements

> *Solidarity with the #FrackFreeFour and thank you for your courage*

> *[username] Incredibly inspiring. I admire her honesty. But it is important for graduates to realise that "the academic life" doesn't stop when we walk across that stage.*

Discussions around mental health

> *see, the thing about anxiety is there's so much fear about Doing The Thing but when you finally find the courage to do it, you feel DUMB after it is all done because it took you so much energy to do an everyday thing for normal people & then you get anxiety about having anxiety.*

Generic advice

> *X20. honesty is always the best policy [link]*

> *Sometimes courage is the quiet voice at the end of the day that says, tomorrow I'll try again...*

Tributes

> *Jarrod Lyle (1981-2018) The Australian competed with a combination rarely seen, of grit and gratitude. He leaves us with his legacy of immense courage. [link]*

Pop culture references (Here impersonating Georgia Steel from the 2018 edition of Love Island)

> *I love a bit of tea with my breakfast. I don't just drink any tea though babes: I drink a pint of Loyalty and if there's none of that then I'm partial to a bit of Honesty. That's just who I am, that's me babes. Inside and out.*

Marketing

> *[brand]- The TRUE story behind [brand]At last I have the courage to tell my story. It took me a while to be able to do this, but now, it feels right. For all those of you who are… [link]*

People sharing personal experience

> *picture it and what i would end up doing or that i could even have days where i'd be happy. i found comfort in little things like making new friends and videos on youtube and art and music. i never had enough courage to do anything but i'm happy i didn't and happy i managed to get through*

Quotes and poetry

> *[username] In peace there's nothing so becomes a man As modest stillness and humility But when the blast of war blows in our ears Then imitate the action of the tiger Stiffen the sinews, summon up the blood Disguise fair nature with hard-favour'd rage. . . I am one of heart with [username]*

Classifier 1b: Virtue language - Neutral / non-neutral; and

Classifier 1c: Virtue language – Positive / Negative

Working in parallel with the classifier above on institutions, a pair of classifiers was trained to determine, of tweets using virtue terms, whether these terms were being used in a positive, negative or neutral sense - whether they were, for example, praising an actor's empathy, decrying their lack of courage or using terms as part of a phrase (e.g. 'in all honesty'). This was carried out in two stages. Classifier 1b first divided tweets into neutral and non-neutral uses of terms. These non-neutral terms were then divided into positive and negative tweets. Figure 5 above provides an illustration of this classifier pipeline – an exploration of what we mean by 'positive', 'negative' and 'neutral is provided below.

'Neutral:'

This label aimed to remove from the dataset 'dispassionate' uses of terms whwhich do not express an opinion on virtues or lack thereof – for example, uses of terms in book titles. These included:

Use of terms in the title of books and other media.

New Book Club Book! This month it is 'The Empathy Problem' by @GavinExtence, a story of a man whose tumour gave him a new lease of life... The book club is on the second Wednesday of this month at PointBook. [link]

So, courage the cowardly dog is 120x scarier now... [link]

Marketing for events and workshops

At a loose end on Friday 13th July? In the Staffordshire area? Join us for our new cognitive empathy workshop! 10am-12pm (noon). Message me here for details [link]

Tweets discussing lack of empathy caused by autism, for example.

[username] I did, I think, but I might've trying to fit in. I agree with your point though. All autistic people need to be accepted including those with no or little empathy.

Tweets not classified as neutral were then classified according to whether they used virtue terms in a positive or negative sense. Examples of topics within these labels are below.

'Positive':

Tweets discussing the benefits of virtue:

Carl Rogers believed that GENUINENESS, EMPATHY, and UNCONDITIONAL POSITIVE REGARD facilitate all relationships; at work, at home, everywhere. [link] [link]

Advice to others to emulate these virtues

Dear New Doctor, There will be times when you've run out of the doc things you can offer the patient. Always remember the human things you have to offer too: 1.Kindness 2.Empathy 3.Compassion #NewDocsTips

Tweets recognising virtue in others

He is a principled person who never compromised on his principles for sake of power. I salute his courage with which he resigned from the National Assembly: Yousaf Raza Gilani #ServingHumanityDrQadri

The place of virtue within religion

The very first qualification of true religious devotion is humility. You cannot practice or progress in religion until you understand where your weaknesses lie. Only your humility can show you that. Religion sets right every wrong element of your mind and body, head & heart. OM [link]

Well-wishing

[username] Incredibly frustrating – so much empathy xx

'Negative':

Tweets claiming other people are not deserving of empathy

it's pretty wild how hard it is to acknowledge that manipulative pieces of shit from your past do not actually deserve of any of your empathy or time lol

Messages alleging a lack of virtue in others

Based on the amount of wilful ignorance or dishonesty Elizabeth has shown since this referendum. I think people should help Elizabeth by teaching her the importance of honesty, respect for others, and an education. How are you going to help her with that, Jack? [link]

Climate journalism is in crisis in the English-speaking world. There is not enough coverage. There are too few trained, experienced climate reporters; so little awareness of climate impacts across the full range of news stories; not enough courage to even mention climate in disaster/weather stories.

Discussion of suicide being a cowardly act

Claims that virtue within institutions is under attack

2/ As Trump does, the Brexit right from the Prime minister down is attacking those practices, rights and institutions that form the bedrock of democracy - basic honesty, the right to oppose, legal checks on the government, and impartiality of civil services & electoral commission.

Classifier 2: Fundraising – relevance

This classifier aimed to separate discussion relevant to fundraising and volunteering from other discussions concerning fundraising sites. Labels and examples are below.

"Relevant"

Discussions relevant to fundraising and volunteering, including:

Mentions of charitable acts which people are taking part in

> *Excited to be heading to [username] for my charity supper club as part of [username].. there are amazing prizes from [username], [username], [username], [username], [username] and so much MORE! £2 donation gets you a chance of winning! #givefoodlovefood [username] [link]*

Encouraging others to volunteer

> *Here's a friendly reminder: if you're feeling disheartened by the state of the world, just about anything is more effective than being sat in front of your computer pouring fuel on the dumpster fire of social media politics. Volunteer for a cause you believe in... 1/3*

Tweets requesting volunteers

> *Come join Bolton Mini Creator Fair's Volunteer Crew! Have fun, share your skills, gain experience and participate in the greatest show {and tell} on Earth! To find out more or to #volunteer, go to: [link]*

Mentions of others volunteering

> *Thank you [username] for volunteering this afternoon to support our school community with distributing flyers for our upcoming PTA/SLT Meeting scheduled for November 9th. #ParentsRock #WeMakeTheDifferenceTogether #PTAMeeting 8:45am #SLTMeeting 4:30pm [link]*

"Irrelevant"

All other tweets, including:

Users fundraising for themselves

> *Hello everyone, I have 2 GoFundMe pages for help to become self-sufficient. Hope you donate, it will help me lots!!!*

Fundraising for party-political causes

> *I donated to the GoFundMe page for Judge Kavanaugh... [link]*

Hunger Games references

> *"I volunteer as tribute! [username]? [link]"*

Phrasal uses of 'volunteer' – for example, 'volunteering information'

Classifier 3a – Gratitude – genuine vs other

This classifier was designed to determine, where possible, if a tweet represented a 'genuine' instance of an expression of gratitude, as opposed to a conversational or formulaic use of terms related to gratitude. It used two labels, 'gratitude' and 'non-gratitude'

"Gratitude"

Tweets thanking others for their support, for reading their work, for their response, etc; including those from charities or organisations

> *[username] [username] [username] Thank you so much, Lucy - I really appreciate your support! Fingers crossed x.*

> *That small comment made my day. It was exactly what I needed after a particularly challenging few weeks and felt I was failing. Thank you NHS*

> *#beheard has come to its' end! Wow, what an incredible response. We wanted to thank every one of you who watched, shared your stories, and got involved because… [link]*

> *Let us mark the beginning of #BCAM with a big THANK YOU to our #Walkers, #Volunteers & #Supporters… making a difference & helping us to end breast cancer. Thank you, thank you.. THANK YOU!! [link] [link]*

> *Thanks to musicians and other celebrities (including professional wrestlers) for their performances*

> *@Kelly_WP I love that you write thanks for the support on your parcels, but it's more like thank you for entertaining us everytime you wrestle [link]*

Gratitude towards named individuals or groups

I have the most loveable and supportive boyfriend in the world and I am so grateful for him

Happy birthday vic! Always take care! Always keep safe. God bless you! I'm so lucky to have you in my life Also thank you for everything! Enjoy your day vic! Always loving you and I'll always be there for you   [username] [link]

I'm so grateful to the colon cancer support group. The individuals there understand you like no one else. They know you do not want pity or sympathy, you just want acceptance. I've... [link]

Thankful for all the great locals that kept smaller fires out round my property. I love u lads. Love you Malibu. Thank you to all hero firefighters around California. It is going to be a journey to rebuild. Stay strong everyone.

Gratitude towards organisations and institutions

Almost 29 yrs of this invisible monster illness kicking my ass. Today, I have hope. Thank you to all those honourable MPs that have taken time to listen to their constituents that are effected by Myalgic Encephalomyelitis #MEDebate #MyalgicE #missingmillions [link]

Gratitude from organisations towards their members, customers, staff or students.

NCs PE would like to thank all those pupils that have represented the school this half-term. Have a fantastic half term!

"Non gratitude"

Thanks as a type of politeness – e.g. 'thank you for letting me know'

[username] Thank you for confirming what I said. The EU would push for ever closer union and for expansion.

Passive aggressive use of terms – e.g."Thanks in advance."

Hello. Any update on the 1835 Stornoway to Dingwall sailing? We're hearing in Ullapool that there will be no livestock allowed. Appreciate your confirmation. Thanks!

Thanks for retweeting / sharing / following - other social media conversation

Greetings [username]! Thank you for following me on Twitch! Have a biscuit! [link]

Tweets promoting the value of being thankful

be grateful for those who are around you. who love you. who care for you. who show even the slightest positivity to you. you never know what might happen

Sarcasm, jokes and memes

Content harvesting is the equivalent digitally of fracking, thank you for coming to my ted talk.

[username] Oh Susan, nearly missed you there. You've unblocked me then? Thanks for your article from a journal described by the Commons as pro-Hamas. Always interested in terrorist publishing. Love that they quote Haaretz, the most loathed publication in Israel, famous for anti-semitic opeds. Great.

Marketing

THE NEW SINGLE 'SOAK IT UP' IS HERE! So excited for this to finally be out there for all! Click below to watch the video and please leave us a comment to tell us what you think! https://t.co/1tUnOfQ9X2 We want to thank EVERYONE who gives a crap about us xx

Classifier 3b – Gratitude – expressed to whom

This classifier was built to examine tweets which had been labelled as 'gratitude' by classifier 3a, above. It separates gratitude expressed towards customers, fans and supporters from other types of gratitude - . Examples of themes are below.

"Customer"

Customer service, both in general and to specific individuals

We are sincerely sorry about recent customer service issues some of you have reported. We take this issue very seriously and are taking steps to correct it. Please see attached for info. Thank you for your patience, your feedback and your continued support. <3 [link]

[username] Hey, Thanks for getting in touch

and apologies for this [username]! Could you kindly DM your username to us so we can to look into this further? [link]

Thanking those donating or supporting to the tweeter

Big thank you to all of you who have donated & supported the [username] in Indonesia already [link]

I absolutely agree, [username] has grown hugely in confidence. Thank you everyone who supported [username] at Abergavenny and St Mary's. [link]

Tweets thanking fans and other supporters

It's a bit late but I just saw my account reached the 20 000 milestone, I just wanted to thank you all for the huge support! So glad to have so many of you enjoying my art and animations, thank you for all of your kindness and love [link]

To say thank you to all our fans who supported our supporters' buses regularly this season, everyone on the bus this Saturday will receive different secret Club Shop gifts. An envelope will be placed on every seat, inside will be the details of your gift. Merry Christmas.

"NonCustomer"

Other expressions of gratitude, including:

Gratitude to organisations

Thank you to @g2fireworks @stocktoncouncil for a great display at Stockton riverside tonight. It had my children's undivided attention. Also like to thank @ ClevelandPolice, @NEAmbulance and @ ClevelandFB for being there https://t.co/ VikaSi0bwy

A big thank you to @BBCRADIOKENT for the opp to talk all things #fertility related Begins 2:10 @FertilityNUK @CARE_ Fertility & @gatewaywomen - I was able to mention you guys and the great work you do! #FertilityWeek18 #WorldFertilityDay #Childlessness https://t.co/YPR0yxhi0D*

Tweets thanking specific individuals

I absolutely agree, [username] has grown hugely in confidence. Thank you everyone who supported [username] at Abergavenny and St Mary's. [link]

Massive congratulations to our groundswoman Paulina Chapman who is moving to exciting new pastures. Thank you for all your hard work and we wish you all the best. [link]

Undirected gratitude

Disappointment and heartbreaks are some of the most valuable lessons in life. They teach you so so much, when you recover you realise just how much they taught you and you will actually be thankful for them because you gained so much wisdom

Keywords used

'Gratitude'
my thanks to
express my gratitude
grateful for
thank you for
thank u for
thanks for
want to thank
like to thank
so grateful
thank you to
you made my day
have my gratitude
sincere gratitude
endless gratitude
express my appreciation
forever be grateful
I am blessed to
wanted to thank
thankful for
my appreciation for
appreciate your

APPENDIX 2

Statement of Purpose

As with much of the changing technological landscape, our empirical understanding of social media's effects on our society, culture and individual characters is still in its nascence. Adolescents and young adults are the most prolific adopters of social networking services (SNS), and the effect of social media on developing minds has been the subject of much speculation. In 'The Moral Web', a previous report exploring the behaviour of young people on SNS, Demos found evidence that young people engage in an array of unethical and risky behaviours online. However, we also found that SNS provide opportunities to display moral and civic virtues such as honesty and empathy in communication, encourage new forms of civic participation and enable acts of courage in countering online abuse.

Public anxiety surrounding the impact of SNS on the development of young people has led many researchers to focus on the negative aspects of SNS. The purpose of this literature review is to redress this imbalance by collating some of the existing research on the ways in which SNS may promote virtuous behaviour online and offline. Understanding how "virtue terms" are used in SNS and whether and how they translate to real world behaviour is crucial in informing policies that encourage pro-social behaviour online at a time when patchy regulation of the internet presents new and unique challenges to educators and parents alike.

Defining Virtue

VIRTUE AS A POSSESSED CHARACTER TRAIT

Virtue is frequently employed as a term defining a positive trait possessed by an individual and constitutive of a person's moral character, but it is ambiguous to define. The concepts of character and virtue begin with Aristotle's Nichomachean Ethics. For Aristotle, achieving eudaimonia - "human flourishing" - and good life for oneself and society as a whole requires living in accordance with perfect virtue (arete). Aristotle defines virtue as a state of character concerned with choice and deliberation which enables a man to be good and do his work well. Since virtue is a matter of choice, it is in the power of every man to pursue it.

Virtue is not an end in itself, but a means to avoiding the two excesses of vice - excess and deficiency - and enabling individuals and society to flourish. For instance, Aristotle sees courage as a method for navigating between the excesses of confidence and the deficiency of fear. Aristotelian virtues include courage, temperance, pride, generosity, magnificence, honesty, wit or charm. These virtues can be learnt, but must be formed into habits through repeated engagement in activities that lead to their deep internalisation.[34]

Drawing on Aristotle's work, a recent collaborative research piece by Demos and the Jubilee centre at the University of Birmingham defines virtues "as a set of personal traits or dispositions that produce specific moral emotions, inform motivations and guide conduct in any area of experience".[35] They can be divided in four categories:

- Moral virtues (such as courage, honesty, humility, empathy and gratitude),
- Intellectual virtues (such as curiosity and critical thinking),
- Performance virtues (such as resilience, application and self-regulation),
- Civic virtues (such as acts of service and

34 Aristotle Nicomachean ethics
35 Birdwell and Reynolds (2015) Character Nation. Available at https://www.demos.co.uk/project/character-nation-2/

volunteering)

The complexity and diversity of spheres of human experience mean that there can be no exhaustive list of the virtues constituting good character; indeed, Aristotle emphasises the contextual nature of ethical dilemmas and virtue. However, the broad categorisation of prototypical virtues outlined above offers a flexible working framework which can be applied to a variety of cultural and situational contexts. Ultimately, a necessary condition for a habit to be a virtue is that it contributes to the flourishing of the possessor of virtue and of social and institutional condition in which all human beings can flourish.[36] [37]

PERFORMING VIRTUE

While virtue can be thought of as a character trait possessed by individuals, the notion of virtue is also linked to action. As already mentioned above, virtues guide people's conduct in all spheres of human experience. Thus, virtue is not only a matter of understanding a set of rules and principles, it is about developing the ability and the will to act in a virtuous way in real-life contexts.[38] [39] In other words, virtue is embodied and developed in action- it is enacted or performed in a variety of real-world contexts, including on SNS. In fact, the term Virtue comes from the Greek word arete which can mean both "moral goodness" and "success or excellent action".[40] These two facets of virtue--character and action-- cannot be separated from one another. With that in mind, virtue is perhaps best conceptualised as a form of "performative knowledge"or "practical wisdom" - an ability or a skill to make reasoned moral judgements and act upon them.

VIRTUOUS ACTIONS

In everyday language, we use the concept of virtue to refer both to a person's character traits and their actions. For example, one might say that a person is courageous, has a standing trait of courage or that a particular act was courageous. In The Right and the Good, W.D. Ross argues that virtuous actions stem from any of the three following motivations: a desire to do one's duty, to do something good, to elicit pleasure or prevent pain for another person. Therefore, in order to assess whether an act is virtuous or not, one must be able to identify a person's motivations at the time of action.[41]

According to Aristotle, however, for an act to be virtuous, it must proceed not only from a person's virtuous feelings and motivations at the time of acting, but also from that person's firm and stable character. In this view, an act is only virtuous if it can be linked to a person's virtuous traits independently of the person's act or motivations. In other words, virtuous acts must derive from virtuous dispositions and not only a person's feelings and motivations at the moment of acting.[42] Therefore, for an action to be virtuous, it is not enough that it promotes social good or has a positive impact on society.

The role of SNS use in promoting virtue and virtuous behaviour

Research has definitively shown that overall usage of SNS has grown exponentially among both youth and adults in recent years. According to a recent survey by the National Office of Statistics approximately 65% of people aged above 16 living in the UK used SNS such as Twitter or Facebook in 2018 - a significant increase since 2011 (45%).[43] Among the 16-24 and 25-34 year-old demographics however, social media usage rose markedly to 93% and 88% respectively. Moreover, the survey shows that SNS were more popular among women than men, with respectively 69% and 60% respondents reporting SNS use.

A 2018 Ipsos Mori survey found similar results. On average, they found that 67% of British adults aged over 15 use social media platforms, with no significant variation between men and women. Furthermore, among the 15-24 and 25-34 year-old demographics they found that social media usage rises to respectively 89% and 88%.[44] Facebook seems to be the most popular social media platform across all age groups and was used regularly by

36 Swanton (2013) The definition of virtue ethics. The Cambridge Companion to Virtue and Ethics, 315.
37 Jubilee (2017) A Framework for Character education in Schools https://www.jubileecentre.ac.uk/media/news/article/5514/New-A-Framework-for-Character-Education-in-Schools-Published
38 Eiesle (1987) Must Virtue be Taught? Faculty Articles and Other Publications. Paper 30.
39 Jubilee (2017) A Framework for Charter Education in Schools
40 Eisele (1987) p.6.
41 Office of National Statistics (2018) Internet Access Households and Individuals. Available at https://www.ons.gov.uk/peoplepopulationandcommunity/ (Accessed 16.08.2018)
42 Hurka (2006) Virtuous act, virtuous dispositions, Analysis 66 (1), 69-76.
43 Office of National Statistics (2018) Internet Access Households and Individuals. Available at https://www.ons.gov.uk/peoplepopulationandcommunity/ (Accessed 16.08.2018)
44 Ipsos Mori (2018) Technology Tracker. Available at: https://www.ipsos.com/ (Accessed 16.08.2018).

approximately 61% of respondents - 59% male and 63% female respectively. Instagram was the second most popular platform (28%) followed by Twitter (21%). The research also found that young people particularly 15-24s have a more diverse mix of social media use.

SNS, then, are increasingly popular among young people and adults, and will almost certainly continue to be in the long term. This raises the question of how SNS use can promote virtuous behaviour online and offline.

If virtues are engrained end developed through repeated engagement in activities that lead to a habit of excellence, then the increasing use of SNS as a platform for social interaction will likely shape the character and habits of current and future generations. Studies by Harrison and Vallor have shown growing concern amongst parents and policy-makers that the Internet, and particularly social media, may provide more opportunities for breaches of morality and present users with behaviours and viewpoints that conflict with the value messages they receive elsewhere.[45] [46]

However SNS may also promote virtuous behaviour such as charitable actions, honesty or empathy by providing structural social opportunities and pressures to act virtuously.[47] A Demos report conducted in partnership with the Jubilee Centre for Character and Virtue found that the vast majority (88%) of 16-18 year-olds polled said they had given emotional support to a friend on SNS.[48] In the following section we review some of the existing literature on the role of SNS in promoting virtue and virtuous behaviour.

ENABLERS OF VIRTUES

According to Aristotle, achieving good life is an inherently social activity and friendship is important in nurturing virtuous behaviour because it provides emotional support and positive reinforcement along the difficult path to perfect virtue. In a theoretical paper exploring how ethical theories apply to social media, Vallor contends that SNS may support and strengthen real-life friendship by facilitating reciprocity, self-knowledge, empathy and the shared life. She argues that SNS such as Facebook offer opportunities for reciprocity that surpass prior online forms of self-expression - these include, liking, comments, friendship invitation, tagging,etc. Furthermore, contrary to the common fear that the use of SNS may decrease face-to-face interaction, she argues that they actually increase the opportunities for such interactions. In fact, SNS help facilitate collective social action, joint endeavours and the kind of civic friendship that Aristotle sees as enabler of virtuous action.[49]

There is a wealth of research which suggests that the Internet has had a detrimental effect on the quality of relationships - see, for example, Morgan's 2017 study on empathy and authenticity online. There may be some call for optimism here, however - a recent survey by the Pew Research Centre suggests that, overall, the Internet and SNS provide more social benefits than negatives by providing new opportunities to create, enhance and rediscover social ties and lowering the traditional communication constraints of cost, distance, and time. They found that a large majority of respondents (85%) agreed that the Internet had been a positive force in nurturing and enhancing their social relationships and that this would continue to be the case in the future.[50]

Moral Virtues

HONESTY AND AUTHENTICITY

While the anonymity of the Internet and many online contexts is often seen as a threat to the moral virtues of honesty and authenticity, SNS may provide structural opportunities to encourage honesty and the presentation of one's 'authentic' self. Online authenticity can be defined as the "consistency between one's behaviour and expressions online and their experiences, thoughts, feelings and actions offline".[51] Vallor hypothesizes that SNS may provide spaces which enable people to interact authentically. Some social networks have been explicitly set up to facilitate this, such as outSMACK.com, which provides support for gay and transgender youth who

45 Morgan, B., et al. (2017). Empathy and Authenticity Online: The roles of Moral Identity and Moral Disengagement in Encouraging or Discouraging Empathy and Authenticity, Jubilee Centre for Character and Virtue. p. 17
46 Harrison, T. J. (2014). Does the Internet Influence the Character Virtues of 11 to 14 year olds in England. School of Education. Birmingham, University of Birmingham. PhD. p.11
47 Vallor, S. (2009). "Social networking technology and the virtues." Ethics and Information Technologies 12: 157-170.
48 Harrison-Evans, P. and A. Krasodomski-Jones (2017). The Moral Web: Youth Character Ethics and Behaviour, Demos.
49 Vallor, S. (2012). "Flourishing on facebook: virtue friendship & new social media." Ethics and Information Technology 14(3): 185-199.
50 Anderson, J. Q. and L. Rainie (2010). The Future of social relations, Pew Internet and American Life project.
51 Morgan, B., et al. (2017). Empathy and Authenticity Online: The roles of Moral Identity and Moral Disengagement in Encouraging or Discouraging Empathy and Authenticity, Jubilee Centre for Character and Virtue.p.9.

may otherwise suppress their authentic selves.[52]

Stern's findings seem to support this authenticity hypothesis. She shows how blogging and the publication of personal content online can help young people refine their sense of self and as a result portray themselves more honestly online and offline. For instance, she quotes Lisa, a young online author:

> *"(My Blog has) made me more comfortable with myself… Instead of having to do things to please other people, to put on different ,asks for everyone, it's sort of made me say " hey, this is who I am! And you want to write in your comments, go ahead, but read this - This is me, Either you like it or you don't….".*[53]

Lisa describes how the validation and acceptance received by her online audience was a catalyst for this change.

Boyd's ethnographic study of young users of MySpace discusses the ways in which SNS profile act as a medium for individuals to explore their identity and write themselves into being. However, she finds that the version of the self portrayed by young users of SNS online is heavily influenced by an individual's perceived audience, and their ideas around what constitutes "cool" behaviour.[54]

EMPATHY AND COMPASSION

There has been concern among academics and educators that SNS and the online world prevents the full experience of empathy because of the absence of face-to-face contact and the fact that the communication of empathy is often largely non-verbal. However, research shows that SNS can provide people with opportunities to act in empathic and compassionate ways.

Vallor discusses the ways in which SNS may promote empathy - the ability to feel with and for others. Empathy is a virtue, as opposed to simply a feeling, as it requires cultivation to become habitual: empathy requires balancing between openness to the other and emotional preservation of the self. She argues that these are demonstrated on a plethora of SNS - such as Cancer Survivors Network, CaringBridge or Daily Strength - which allow people with serious illness, caregivers and people who have experienced traumatic events to share their experience and receive words of wisdom and support.[55] [56] Vallor argues that we need to explore further the way in which these platforms create opportunities to express and receive empathy.[57]

In his doctoral thesis, Harrison examined the influence of the Internet on the character virtues of honesty and compassion of 11-14 year olds in the UK. In his large n survey he found that a wide majority of respondents (71%) reported having helped someone else on the Internet. Furthermore, 66% of respondents agreed to the statement "I have helped other people on Facebook" and an overwhelming majority (65%) said that they were respectful of other people's views on Facebook, suggesting that SNS have the potential to foster compassionate behaviour. On the other hand, 43.3% of the respondents believed that Facebook have the potential to increase non-compassionate behaviours such as saying nasty things to people and nearly a third (31%) admitted being unkind to somebody online.[58] Harrison's finding that young people engage in both compassionate and non-compassionate behaviour in SNS seems to indicate that the Internet is neither a "good" or "bad" influence per se on compassionate behaviour online, and that more research is needed to understand the circumstances under which virtuous and non-virtuous behaviour emerge online.

In a 2001 paper, Preece and Ghozati analysed the content of two thousand messages drawn from a hundred online communities to determine how common empathy is online, whether it is as widespread online as in face-to-face communication and whether it is more common in some online communities than others. Their sample included patient support communities as well as a range of other groups centred around cultural issues, pet ownership, religion, politics and societal issues and sports. They defined empathic messages as those conveying feelings of compassion, knowing, feeling

52 Vallor (2009)
53 Stern, S. (2008). Producing Sites, Exploring Identities: youth Online Authorship. Youth, identity and digital media. D. Buckingham. Cambridge, MA, MIT Press. p. 110.
54 Boyd, D. (2008). Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life. Youth, Identity and Digital Media. D. Buckingham. Cambridge, MA, MIT Press.
55 Leonard, K., et al. (2015). "Moderated Social Media Support Groups for Patients." Journal of Consumer Health on the Internet 19(3-4): 219-232.
56 Wilkerson, D. A., et al. (2018). "Friendsourcing Peer Support for Alzheimer's Caregivers Using Facebook Social Media." Journal of Technology in Human Services: 1-20.
57 Vallor (2009)
58 Harrison, T. J. (2014). Does the Internet Influence the Character Virtues of 11 to 14 year olds in England. School of Education. Birmingham, University of Birmingham. PhD. p. 119-140

and shared experience. They found that 81% of the communities contained some empathetic messages, and that in 18% of the communities more than half of the messages were of empathic nature. However, they also found that hostile messages occurred in 1 in 3 online communities.[59]

Performance and Intellectual virtues

SNS may also provide opportunities for the growth of performance and intellectual virtues such as critical thinking, curiosity and self-application. For instance, Stern argues that SNS increase opportunities for youth authorship of creative content, although she raises concerns about risks - such as exposure to abuse or trolling - associated with publishing personal content online. Online youth authorship on SNS or blogs may promote the development of performed virtues such as self-application. In fact, in a series of interviews with young online authors, Stern found that young people were often motivated to set up a personal page through a desire to demonstrate autonomy and to master new skills. In this way, blogging sites may provide young people with a space to enact self-application, through time spent writing and the diligence required to maintain a regularly updated personal page. Young authors seem to feel a sense of obligation and pressure to maintain their sites, which may promote development of a habit of self-application. In addition, by allowing people to document and witness or monitor their personal growth through time, SNS can provide a motivation for perseverance.[60]

Stern's study also shows that youth authorship on SNS platforms provides young people with opportunities for self-reflection; to question previously held assumptions and reevaluate values and beliefs. In other words, SNS can promote or enable the exercise of intellectual virtues such as critical-thinking, curiosity and self-examination. For instance, one respondent said that: "My blog has helped me to centre my feelings and realise that I need to take things one step at a time. It forces me to think about who I am, what I like and who I want to be. I can think about one of the problems I am going to face, but writing about it allows me to work through the problem and start to look at solutions".[61]

CIVIC VIRTUES

SNS has the potential to foster civic virtues by encouraging young people to participate in civil society. The Internet may provide a vehicle for young people to connect with charitable causes, make a positive difference to other's lives and become more socially conscious. Harrison argues that the strong sense of community that can be developed on SNS through online participation, civic engagement and social and political action may lead SNS users to better understand their obligation to their communities and foster civic action both online and offline.[62]

A Demos report entitled Service Generation found that SNS have created new digital spaces significant for those willing to get involved in forms of civic participation. The report shows that approximately 340,000 young British Facebook users have interests related to social action, defined as 'practical action in the service of others', and between 7 and 14th November 2013 around 150,000 tweets were identified as discussing social action. Young people used SNS to discuss experiences or attitudes toward social action, share information and stories about actions they had taken, campaign and raise awareness.[63]

Another Demos report, entitled Introducing Generation Citizen, found that many young people used SNS as a tool to engage with social issues online. For example 38% of those surveyed reported having signed a petition online and 29% had used Facebook or Twitter to raise awareness about a cause.[64] Charity fundraising platforms also allow people to use SNS to donate to or crowdfund for causes they care about. In fact, crowdfunding campaigns using internet and social media to mobilise people quickly around causes are a growing market, with around £81m raised for good causes in 2015.[65]

While there is growing anxiety and evidence that SNS can promote illicit and risky behaviour,

59 Preece, J. and K. Ghozati (2001). Observations and Explorations of Empathy Online. the Internet and Health Communication: Experience and Expectations. R. Rice and J. Katz, Sage Publications Inc: 237-260.
60 Stern (2008)
61 Stern (2008)
62 Harrison, T. J. (2014). Does the Internet Influence the Character Virtues of 11 to 14 year olds in England. School of Education. Birmingham, University of Birmingham. PhD.p.53
63 Birdwell, J. and C. Miller (2013). Service Generation: A Step-Change in Youth Social Action. London, Demos.
64 Birdwell, J. and M. Bani (2014). Introducing Generation Citizen. London, Demos.
65 Nesta (2016). Crowdfunding Goodcauses: Opportunities for charities, community groups and social entrepreneurs, Nesta.

previous research also shows that SNS can provide opportunities for the development of moral, intellectual, performance and civic virtues and act virtuously. To understand this, we need to understand which factors promote virtuous behaviour online.

Factors promoting virtuous behaviour online

There is emerging evidence that mechanisms promoting accountability on SNS, such as the presence of moderator or strong community ties, promote virtuous or positive behaviour online. According to Preece and Ghozati's study of online communities, patient support groups and moderated communities were more likely to show high levels of empathic communication than other online groups.[66] This seems to suggest that communities with established norms and rules and where a moderator is present to enforce them may promote forms of virtuous behaviour such as empathy. Similarly, James and colleagues argue that accountability online (and the moral behaviour associated with it) depends on the strength of ties within a given online community: the stronger the ties, the greater the accountability and vice versa.[67]

In a 2001 paper, Preece and Ghozati voice their concern that features of the online world such as anonymity and asynchronicity enable individuals to feel less restrained and act in ways they would not normally act offline (eg. use impolite language, criticise more harshly and display anger and hatred) - the online disinhibition thesis.[68] However, there is emerging theoretical and empirical evidence that any disconnect between offline and online behaviour is unlikely to be complete. The co-construction theory suggests that young adults construct their online worlds as extensions of their offline world, and that both worlds are psychologically connected.[69] [70] For instance, in their survey of the relationship between young adult's online and offline lives, Subrahmanyam and colleagues found that participant's use of SNS was integrated with the people and concerns from their offline lives. Specifically, they found that young adults used SNS to connect with friends that were part of their offline social network. However, while a young adult's online and offline worlds are connected, they do not perfectly mirror each other. Rather, online contexts offer opportunities and limitations distinct from those offered by offline contexts, and these shape the way interactions are conducted.[71]

The co-construction model suggests that people's online and offline behaviour are importantly related, and studies by Wang and Wang (2008) and Wright and Li (2011) seem to suggest that pro-social behaviour offline and online are indeed connected. Wang and Wang's study of the link between pro-social behaviour offline and on online gaming sites showed that altruistic gamers were more likely to exhibit altruistic behaviour online - to offer to help to fellow players, for example - compared to less altruistic gamers.[72] Similarly, Wright and Li (2011) examined the link between pro-social behaviour offline and on SNS and found that face-to-face prosocial behaviour significantly predicted the display of prosocial behaviour on SNS.[73]

These studies support the co-construction theory and indicate that young adults behave and socialise similarly in the online worlds and offline. There is also some evidence that anti-social behaviour online is related to anti-social behaviour offline. For example, a survey on the link between cyberbullying and traditional bullying by Sourander and colleagues found that many cyberbullies were also traditional bullies.[74]

A recent research paper from the Jubilee Centre, based on a survey of 11-18 year olds, shows that character traits such as moral identity or moral disengagement could predict levels of online empathy and online authenticity. Moral identity - "having moral traits as an important part of one's sense of self" - was positively related to online empathy and online authenticity and could

66 Preece, J. and K. Ghozati (2001). Observations and Explorations of Empathy Online. the Internet and Health Communication: Experience and Expectations. R. Rice and J. Katz, Sage Publications Inc: 237-260.
67 James, C., et al. (2009). Young people, ethics and the new digital media: a synthesis form the goodplay project. Cambridge, MA and London, MacArthur Foundation.p.13
68 Suler, J. (2004). "The Online Disinhibition Effect." CyberPsychology & Behaviour 7(3): 321-326.
69 Suler, J. (2004). "The Online Disinhibition Effect." CyberPsychology & Behaviour 7(3): 321-326
70 Wright, M. F. and Y. Li (2011). "The associations between young adults' face-to-face prosocial behaviors and their online prosocial behaviors." Computers in Human Behavior 27(5): 1959-1962
71 Subrahmanyam, K., et al. (2008).
72 Wang, C., & Wang, C. (2008). Helping others in online games: Prosocial behavior in Cyberspace. CyberPsychology, Behavior, & Social Networking, 11,344–346.
73 Wright, M. F. and Y. Li (2011).
74 Sourander, A., et al. (2010). "Psychological Risk Factors Associated With Cyberbullying Among Adolescents: A Population Based Study." Arch Gen Psychiatry 67(7).

positively predict 8.7% of the variance in online empathic behaviour and 12% of the variance in online authenticity for the study's sample. On the other hand, moral disengagement - a character trait enabling individuals to disengage from their moral selves without feelings of shame and guilt - negatively predicted online authenticity and empathy.[75] Therefore, it seems as highlighted by Harrison that online behaviour is influenced by both the moral character of individuals and the features of the unique internet such as anonymity.[76]

In sum, virtuous behaviour online seems to be a factor of both the features of online platforms (eg. moderation, community) and the character traits of SNS users. In addition, there is some evidence that external events, such as natural disasters, can promote the viral spread of compassionate messages on SNS.[77]

Impact of SNS on offline behaviour

The impact of SNS use on offline virtuous behaviour is difficult to evaluate. To date, much research has focused on how SNS may facilitate illicit behaviour offline. For example, Huang and colleagues' study of Californian 10th graders found that while frequency of SNS use did not contribute to adolescent smoking and drinking, "exposure to friend's risky displays online significantly contributed to adolescent smoking and drinking". This points to the possibility of the internalization of bad behaviour through SNS networking.[78] There has also been much concern about the role of SNS and extreme content online in promoting acts of violence offline. A study by Hawdon and colleagues revealed the existence of networks of fans online, and particularly on Youtube, idolizing school shooters such as the Columbine Killers.[79] This is concerning in that it shows that those with extreme views can easily find legitimation and support online. Nevertheless, it is difficult to establish whether exposure to extremist content online is a causal factor engendering extremist-related violence offline. In fact, research by the counter-extremism think tank Quilliam found that in the vast majority of cases, individuals were not radicalised solely online in isolation of other contexts.[80]

Recent research also suggests that SNS may encourage civic behaviour offline by providing more opportunities to engage in forms of civic participation (eg. volunteering, giving to a charity). In their recent survey of civic participation in the UK, NCVO concluded that the internet and social media had made it easier for people to access opportunities such as volunteering.[81] Evidence from empirical studies seems to support this conclusion. For example, a study by Kim and Lee found that 74% of surveyed American college students who had volunteered for a non-profit organisation had joined it through a social media platform. They also found that the perception social pressure on social media was an important factor motivating people's decision to volunteer.[82]

Several studies have found a positive relationship between SNS use, volunteering and donation to charity. Using the PEW Internet and American Life Project data set, Mano found that participation in social media increases the level of online donations to charity, though it does not impact offline contributions.[83] Similarly, Farrow and Yuan found that alumni who actively used social media alumni groups were more likely to donate to their alma mater.[84] In addition, Valenzuela and colleagues found through their online survey of college students across Texas that Facebook users were significantly more likely to engage in form of civic participation such as volunteering or raising money for charity than non-

75 Morgan, B., et al. (2017). Empathy and Authenticity Online: The roles of Moral Identity and Moral Disengagement in Encouraging or Discouraging Empathy and Authenticity, Jubilee Centre for Character and Virtue.
76 Harrison, T. (2016). "Cultivating cyber-phronesis: a new educational approach to tackle cyberbullying." Pastoral Care in Education 34(4): 232-244.
77 Boulianne, S., et al. (2018). "Does compassion go viral? Social media, caring, and the Fort McMurray wildfire." Information, Communication & Society 21(5): 697-711.
78 Huang, G. C., et al. "Peer influences: the impact of online and offline friendship networks on adolescent smoking and alcohol use." Journal of Adolescent Health 54.5 (2014): 508-514
79 Hawdon, J., et al. (2015). "Online Extremism and Online Hate: Exposure among Adolescents and Young Adults in Four Nations." Nordicom-Information 3(4): 29-37.
80 Hussain, G. and E. M. Saltman Jihad Trending: A Comprehensive Analysis of Online Extremism and How to Counter it, Quilliam. Available at: https://www.quilliaminternational.com/.
81 NCVO (2017). Getting Involved: How people make a difference. London, NCVO.
82 Kim, Y. and W. N. Lee (2014). "Networking for philanthropy: increasing volunteer behavior via social networking sites." Cyberpsychol Behav Soc Netw 17(3): 160-165.
83 Mano, R. S. (2014). "Social media, social causes, giving behavior and money contributions." Computers in Human Behavior 31: 287-293.
84 Farrow, H. and Y. C. Yuan (2011). "Building Stronger Ties With Alumni Through Facebook to Increase Volunteerism and Charitable Giving." Journal of Computer-Mediated Communication 16(3): 445-464

Facebook users.[85] A Demos survey also found that 28% of the 16 to 18 year-olds they surveyed had encouraged others to take action on political or social issues over Facebook.[86] This evidence indicates that the Internet and SNS can provide a vehicle for people to connect to and participate in charitable causes and may lead them to become more socially conscious.

Rather than seeing SNS use as exerting a simple positive or negative influence on behaviour offline, a more productive approach may lie in studying the conditions under which SNS use promotes virtuous behaviour offline. Although more research is needed in this area, a study by Coyne and colleagues shows that social networking with parents, by adding them as friends on Facebook, for example, increases the likelihood of adolescent acting pro-socially offline and promotes feelings of connection between parents and children. Conversely, they found that social networking without parents was associated with negative outcomes such as increased relational aggression, delinquency and decreased feelings of connection.[87] Therefore, it is possible that the social makeup of one's online social network online influences offline behaviour - a point also evidenced by Huang and colleagues' study on social media and alcohol consumption.[88]

In addition, certain events such as disasters may increase the likelihood of SNS use translating into civic virtuous behaviour offline such as donating to charity, volunteering, and caring for victims. A study by Boulianne et al found that in the months following the McMurray wildfires in Canada, Albertans who used SNS were twice as likely to help than those who didn't. Furthermore, they found that the most popular tweets in the month following the event were expressions of support and concern, and invitations to help victims of the fires. They argue that viral tweets expressing compassion and the "spirit of care" they can engender in a community help promote acts of caring offline.[89]

Use of charitable fundraising sites

SNS provide low-cost opportunities for charitable organisations to reach and connect with new and wider audiences for fundraising purposes, as well allowing people to discover and contribute to charitable causes. A wide range of charitable fundraising sites (CFS) are available, including Facebook Causes, GoFundMe, Crowdrise and BT Mydonate (thought this latter has now ceased operation). To date, however, there is very little literature available on CFS, online donors and their motivations.

There is some evidence that SNS use increases the likelihood of people giving to charity. As noted above, studies by Mano, Farrow and Yuan have found that participation in social media increases the level of online donations to charity, and that alumni who were active in relevant social media groups were more likely to donate to their alma mater.[90] [91] While more research is needed to understand how SNS encourage charitable giving (particularly online) theoretical evidence indicates that SNS can increase people's motivations to give to charity in a variety of ways. For instance, Boulianne and colleagues argue that because donors are often prompted to post messages about donations on social media, SNS can normalise charitable giving among social network members.[92] In fact, research on motivations for charitable giving indicates that people are more likely to give - and give more - to charitable causes when they receive information that other people have donated, and when their giving is announced in public.[93]

In a study about SNS fundraising campaigns in the US, Saxton and Wang (2014) found that people who donated to charitable causes through CFS were motivated by different factors than people who gave through traditional offline channels. For example, while donations to charitable causes offline were largely motivated by the organisation's efficiency ratio (the percentage of donations which is turned into programmatic output) this factor was less

85 Valenzuela, S., et al. (2009). "Is There Social Capital in a Social Network Site?: Facebook Use and College Students' Life Satisfaction, Trust, and Participation." Journal of Computer-Mediated Communication 14(4): 875-901

86 Harrison-Evans, P. and A. Krasodomski-Jones (2017). The Moral Web: Youth Character Ethics and Behaviour, Demos.

87 Coyne, S. M., et al. (2014). "A friend request from dear old dad: associations between parent-child social networking and adolescent outcomes." Cyberpsychol Behav Soc Netw 17(1): 8-13.

88 Huang, G. C., et al. "Peer influences: the impact of online and offline friendship networks on adolescent smoking and alcohol use." Journal of Adolescent Health 54.5 (2014): 508-514

89 Boulianne, S., et al. (2018). "Does compassion go viral? Social media, caring, and the Fort McMurray wildfire." Information, Communication & Society 21(5): 697-711.

90 Farrow, H. and Y. C. Yuan (2011). "Building Stronger Ties With Alumni Through Facebook to Increase Volunteerism and Charitable Giving." Journal of Computer-Mediated Communication 16(3): 445-464

91 Mano, R. S. (2014). "Social media, social causes, giving behavior and money contributions." Computers in Human Behavior 31: 287-293.

92 Boulianne,S. et al (2018)

93 Bekkers, R. and P. Wiepking (2011). "Understanding Philanthropy: A Review of 50 Years of Theory and Research."

important for online donations. The organisations with the most successful online campaigns were rather those with a large network of "fans", allowing them to reach expanding circles of people through each fan's online network. In this sense, online donations are less a factor of a charitable organisation's internal characteristics (e.g. financial capacity, efficiency, size) but of their social network online. Saxton and Wang suggest as a possible explanation that online donations are strongly motivated by the pressures deriving from one's social network, and the desire to improve one's standing in that network.[94]

The negative impact of social media on character

As with any novel medium of communication, a great deal of attention has been paid to the potential detrimental effects of SNS on adolescents behaviour and character. A survey by the Jubilee Centre at the University of Birmingham found that more than half of UK parents think that SNS hamper the moral development of their children; parents thought that anger, arrogance, ignorance, bad judgment and hatred were the top negative character traits displayed on SNS.[95] This work suggests a high level of concern among parents that SNS have a negative and potentially harmful impact on youth character development.

There is substantial evidence that illicit behaviour is common online. A Demos report on moral behaviour online found that just over a quarter (26%) of 16-18 year-olds polled reported having bullied or insulted someone else on SNS, and 69% had experienced some form of cyberbullying. Furthermore, in their a cross-national comparison of incidences of cyberbullying in the United States and Finland, Näsi and colleagues found that approximately 17% of US respondents and 19% of Finnish respondents reported being victims of cyberbullying.[96] The recurring prevalence of cyberbullying in these 2 countries raises the possibility that SNS facilitate or even drive this harassment.[97]

The prevalence of illicit behaviour online has led many to warn that SNS may distort the character of adolescents. John Suler coined the influential phrase "online disinhibition effect" to describe the tendency of people to act differently in the cyberspace than they would normally do face to face. He argues that six features of SNS and the digital world contribute to this phenomenon:

- Dissociative anonymity (the general anonymity of the online world creates fewer checks on behaviour)
- Invisibility (People do not see each other face to face gives people courage to do things they otherwise would not do)
- Asynchronicity (people do not communicate with each other in real time online)
- Solipsistic introjection (absence of facial cues means that people rely on their own imagination to represent the interlocutor and his or her intentions)
- Dissociative imagination (internet users may think of their online person as different from themselves and their online lives as different from reality)
- Minimization of status and authority (absence of physical cues of authority and and status online reduces ability for authority figures offline to replicate this online)

Suler's "online disinhibition effect" thesis suggests that SNS and digital media provides more structural opportunities for illicit behaviour than the offline world, and may as a result have a negative influence on character.

There is some empirical evidence backing the online disinhibition thesis. A 2010 study comparing the moral justification of traditional bullying and cyberbullying among school children found that traditional forms of aggression seemed to require a higher level of rationalisation or justification than online aggression. The authors argued that the anonymity of SNS, combined with the distance from both the victim and the consequences of aggression, allow perpetrators of online aggression to disengage morally more easily and escape feelings of guilt,

94 Saxton, G. D. and L. Wang (2014). "The Social Network Effect: The Determinants of giving Through Social Media." Nonprofit and Voluntary Sector Quaterly 43(5): 850-868

95 Morgan, B. (2016). The Virtues and Vices of Social Media Sites. Virtue Insight: Conversation on Character. Birmingham, University of Birmingham, Jubilee Centre for Character and Virtue. Retrieved September 2018 from: https://virtueinsight.wordpress.com/2016/07/14/the-virtues-and-vices-of-social-media-sites/)

96 Harrison-Evans, P. and A. Krasodomski-Jones (2017). The Moral Web: Youth Character Ethics and Behaviour, Demos.

97 Näsi, M., et al. (2014). "Association between online harassment and exposure to harmful online content: A cross-national comparison between the United States and Finland." Computers in Human Behavior 41: 137-145.

shame and self-condemnation.[98] Further to this, based on in-depth interviews with young people between 10 and 25, James found that young people were often blind to the moral implication of their online action. For example she found that hostile speech online could be met with a belief that content posted online was "just a joke" and that half of the participants to the study asked to play an online game and presented with a hypothetical scenario said they would chose to scam another new player for their own benefit.[99]

98 Pornari, C. D. and J. Wood (2010). "Peer and cyber aggression in secondary school students: the role of moral disengagement, hostile attribution bias, and outcome expectancies." Aggress Behav 36(2): 81-94.

99 James, C. (2014) Disconnected: Youth, New Media and the Ethics Gap.

Licence to publish

Demos – License to Publish
The work (as defined below) is provided under the terms of this licence ('licence'). The work is protected by copyright and/or other applicable law. Any use of the work other than as authorized under this licence is prohibited. By exercising any rights to the work provided here, you accept and agree to be bound by the terms of this licence. Demos grants you the rights contained here in consideration of your acceptance of such terms and conditions.

1 Definitions
a 'Collective Work' means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this License.
b 'Derivative Work' means a work based upon the Work or upon the Work and other pre-existing works, such as a musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work or a translation from English into another language will not be considered a Derivative Work for the purpose of this License.
c 'Licensor' means the individual or entity that offers the Work under the terms of this License.
d 'Original Author' means the individual or entity who created the Work.
e 'Work' means the copyrightable work of authorship offered under the terms of this License.
f 'You' means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from Demos to exercise rights under this License despite a previous violation.

2 Fair Use Rights
Nothing in this licence is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3 License Grant
Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) licence to exercise the rights in the Work as stated below:
a to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;
b to distribute copies or phono-records of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works; The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4 Restrictions
The licence granted in Section 3 above is expressly made subject to and limited by the following restrictions:
a You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this License, and You must include a copy of, or the Uniform Resource Identifier for, this License with every copy or phono-record of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this License or the recipients' exercise of the rights granted hereunder. You may not sublicence the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this License Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this License. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any reference to such Licensor or the Original Author, as requested.

b You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.
c If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Collective Works, you must keep intact all copyright notices for the Work and give the Original Author credit reasonable to the medium or means You are utilizing by conveying the name (or pseudonym if applicable) of the Original Author if supplied; the title of the Work if supplied. Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5 Representations, Warranties and Disclaimer
a By offering the Work for public release under this License, Licensor represents and warrants that, to the best of Licensor's knowledge after reasonable inquiry:
i Licensor has secured all rights in the Work necessary to grant the licence rights hereunder
and to permit the lawful exercise of the rights granted hereunder without You having any obligation to pay any royalties, compulsory licence fees, residuals or any other payments;
ii The Work does not infringe the copyright, trademark, publicity rights, common law rights or any other right of any third party or constitute defamation, invasion of privacy or other tortious injury to any third party.
b Except as expressly stated in this licence or otherwise agreed in writing or required by applicable law, the work is licenced on an 'as is' basis, without warranties of any kind, either express or implied including, without limitation, any warranties regarding the contents or accuracy of the work.

6 Limitation on Liability
Except to the extent required by applicable law, and except for damages arising from liability to a third party resulting from breach of the warranties in section 5, in no event will licensor be liable to you on any legal theory for any special, incidental, consequential, punitive or exemplary damages arising out of this licence or the use of the work, even if licensor has been advised of the possibility of such damages.

7 Termination
a This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Collective Works from You under this License, however, will not have their licences terminated provided such individuals or entities remain in full compliance with those licences. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.
b Subject to the above terms and conditions, the licence granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different licence terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other licence that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8 Miscellaneous
a Each time You distribute or publicly digitally perform the Work or a Collective Work, Demos offers to the recipient a licence to the Work on the same terms and conditions as the licence granted to You under this License.
b If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.
c No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.
d This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of Demos and You.

DEMOS

Demos is a champion of people, ideas and democracy. We bring people together. We bridge divides. We listen and we understand. We are practical about the problems we face, but endlessly optimistic and ambitious about our capacity, together, to overcome them.

At a crossroads in Britain's history, we need ideas for renewal, reconnection and the restoration of hope. Challenges from populism to climate change remain unsolved, and a technological revolution dawns, but the centre of politics has been intellectually paralysed. Demos will change that. We can counter the impossible promises of the political extremes, and challenge despair – by bringing to life an aspirational narrative about the future of Britain that is rooted in the hopes and ambitions of people from across our country.

Demos is an independent, educational charity, registered in England and Wales. (Charity Registration no. 1042046)

Find out more at **www.demos.co.uk**

DEMOS

PUBLISHED BY DEMOS DECEMBER 2019

76 VINCENT SQUARE, LONDON, SW1P 2PD
T: 020 3878 3955
HELLO@DEMOS.CO.UK
WWW.DEMOS.CO.UK