Full Fact

Report on the Facebook Third Party Fact Checking programme

Jan–Jun 2019

Full Fact

Full Fact is the UK's independent fact checking charity.

Full Fact
2 Carlton Gardens
London
SW1Y 5AA

team@fullfact.org

+44 (0)20 3397 5140

@FullFact

fullfact.org

Published by Full Fact, July 2019

Registered Charity number 1158683

Contents

Executive Summary . . . 5
- Key recommendations . . . 6
- Recommendations for Facebook . . . 6
- Recommendations for government . . . 6
- The production of this report . . . 7
- Facebook's response . . . 7
- Editorial independence . . . 7

A brief overview of how the Third Party Fact Checking programme works . . . 9
- The queue . . . 9
- "Attaching" a fact check . . . 10
- What happens next . . . 10
- Ratings . . . 11

Overview of what Full Fact has done in Jan–Jun 2019 . . . 13
- Fact checking . . . 13
- Developing operating guidelines . . . 13
- Liaising with Facebook and other fact checkers working on the programme . . . 14
- Assessing and reporting on the Third Party Fact Checking programme . . . 14
- Building networks . . . 15
- Funding . . . 16

Observations from our work . . . 17
- Specific topics of interest . . . 17
- Health . . . 17
- Police . . . 17
- Some case studies that have informed our recommendations . . . 17
- Satire . . . 17
- Opinion . . . 18
- The burden of proof being on the claimant . . . 19

Our view of the Third Party Fact Checking programme . . . 20
- Tackling specific harms . . . 21
- 'Spam filtering' . . . 22
- The role of technology . . . 22

Recommendations for Facebook and others . . . 25
- Improving the information and tools available to fact checkers . . . 25
- Developing the Third Party Fact Checking programme ratings system . . . 27
- Resolving editorial questions around the programme . . . 30
- Making it easier to evaluate our work on the programme . . . 31
- Expanding and developing the programme . . . 31
- Recommendations for government . . . 32

Future work for Full Fact **34**
Appendix: Full Fact's Operating Guidelines for the Third Party Fact Checking programme . . . **36**
Background: general operating guidelines **36**
What we check, and why **37**
What Full Fact prioritises **37**
Fact checking other content from the queue **38**
Political actors **38**
Humour **39**
How we check **39**
We check claims, not people **39**
We present evidence to allow our readers to reach their own conclusions **39**
Health **40**
How we assign ratings **41**
True **41**
Mixture **41**
False **41**
Satire **42**
Opinion **42**
Ratings we have not yet used **42**
Major Incident procedure **43**
Major incident goal **43**
Triggering a major incident **43**
Active monitoring **44**
Prioritising official sources of information **44**
Reviewing **44**
Action over explanation **45**
Liaising with others **45**

Executive Summary

Since starting work with Facebook on the Third Party Fact Checking programme in January, the first three months were Full Fact's familiarisation period. The following three months were focused on trying to expand our coverage in specific areas that we identified as important, notably health information.

For this period our goal was to understand the nature of the challenges we would be facing from online misinformation on Facebook, to understand and discuss the kinds of editorial choices we need to make within the rules of the programme, and to develop operating guidelines to govern our future work on the Third Party Fact Checking programme.

This report sets out our experience so far and shares our draft operating guidelines. We welcome your feedback on them. We expect that future reports will be briefer.

Our overall view at this point is that –

- The Third Party Fact Checking programme is worthwhile, and it is likely that something similar may be needed on other internet platforms too.
- We have been encouraged by some signs that Facebook is continuing to develop the programme based on feedback, and we believe that further development is needed
- We believe that Facebook's current rating system for the Third Party Fact Checking programme needs to change, and we have made other specific recommendations about how the programme can be strengthened.
- Fact checking depends on access to authoritative expert information, and in a world with more information than ever, where it's hard to know what's true and what's not, we believe government should do more to ensure trustworthy sources are available, for example in areas like public health and the law.

However, we raise two major concerns –

- Scale. Facebook's focus seems to be increasing scale by extending the Third Party Fact Checking programme to more languages and countries (it is currently working with fact checkers across 42 languages worldwide). However, there is also a need to scale up the volume of content and speed of response.

This, again, is an industry-wide concern relevant to other internet companies too.

- Opacity. We want Facebook to share more data with fact checkers, so that we can better evaluate content we are checking and evaluate our impact.

Key recommendations

We make eleven recommendations based on our experience of the programme so far. Ten of these are recommendations for action Facebook should take; one is a longer term recommendation for government.

Recommendations for Facebook

- **Recommendation 1:** Continue developing tools that can better identify potentially harmful false content including repeated posts
- **Recommendation 2:** Provide more data on shares over time for flagged content
- **Recommendation 3**: Add a 'Mixture' rating which does not reduce the reach of content
- **Recommendation 4:** Add an 'Unsubstantiated' rating
- **Recommendation 5:** Add a 'More context needed' rating
- **Recommendation 6:** Add a rating for humorous posts other than satire or pranks
- **Recommendation 7:** Develop clearer guidance on how to differentiate between several claims within a single post
- **Recommendation 8:** Share more data with fact checkers about the reach of our fact checks
- **Recommendation 9:** The Third Party Fact Checking programme should expand to fully include Instagram content
- **Recommendation 10:** Be explicit about plans for machine learning

Recommendations for government

- **Recommendation 11:** The government should review responsibilities for providing authoritative public information on topics where harm may result from inaccurate information and fill gaps

The production of this report

This report was drafted by staff at Full Fact with input from everybody involved in our work under the Third Party Fact Checking programme. The contents are the responsibility of the Chief Executive. They may or may not reflect the views of members of Full Fact's cross-party Board of Trustees and they are not the responsibility of Facebook or any other organisation named in the report.

This report has not been shared in advance with other fact checkers who are part of Facebook's Third Party Fact Checking programme. However, we would be particularly grateful for feedback from other fact checkers.

According to the approach we agreed before joining the Third Party Fact Checking programme, this report was provided in draft to Facebook on 5 July 2019, with an invitation for Facebook to provide feedback and to respond publicly.

Facebook's response

Facebook have responded: "Our third-party fact-checking programme is an important part of our multi-pronged approach to fighting misinformation. We welcome feedback that draws on the experiences and first-hand knowledge of organisations like Full Fact, which has become a valued partner in the U.K.

We are encouraged that many of the recommendations in the report are being actively pursued by our teams as part of continued dialogue with our partners, and we know there's always room to improve. This includes scaling the impact of fact-checks through identical content matching and similarity detection, continuing to evolve our rating scale to account for a growing spectrum of types of misinformation, piloting ways to utilise fact-checkers' signals on Instagram and more. We also agree that **there's a need to explore additional tactics for fighting false news at scale**.

We look forward to continued collaboration with Full Fact and our more than 50 global fact-checking partners."

Editorial independence

Facebook has not sought to influence Full Fact's editorial choices. In particular, Facebook has never asked Full Fact to fact check or not to fact check any specific post, or to give or change any rating, or to treat

any publisher in one way or another. This notice will appear in all future quarterly reports unless there is any reason to modify it.

Facebook provides us with a queue of publicly-shared posts that Facebook has identified as potentially needing fact checking using its own systems. We do not know except in the broadest terms how these posts are chosen. What we have seen included in the queue so far strikes us as what you might reasonably expect such a system to include, although at this stage we have not formed a view on what it may be missing.

A brief overview of how the Third Party Fact Checking programme works

The queue

Fact checkers working on the Third Party Fact Checking programme are provided by Facebook with a "queue" of content (such as text posts, images, videos and links) that it has identified as possibly false. Each fact checker's queue is generated specifically for the territory they operate in; our queue is supposed to prioritise UK-centric content.

We do not know exactly what metrics Facebook uses to determine what goes into the queue, but we do know that it is a combination of Facebook users flagging the content as suspicious, and Facebook's algorithms proactively identifying other signals that might suggest it is false (such as, for example, comments underneath saying "this is fake".)

The queue also includes information on the total number of shares each post has received, and the date it was first shared on. (Since the period this report covers, while it was being written, Facebook has also added information on the number of users who flagged the content, and the number of shares in the previous 24 hours.)

Fact checkers can bookmark items from the queue, to examine later and eventually attach any published fact checks to.

We are also able to proactively add posts to the queue which we have found through our own monitoring and fact checking, for example website links or Facebook posts. The posts we add must be rated either 'false' or 'mixture'. So far we have added one post on health: a Facebook status with almost 60,000 shares claiming a tampon could be put in a stab wound. We added another on whether there was a legal ban on British media reporting on the Yellow Vest protests in France.

From our experience so far, the majority of items in the queue are not things that we either would or could fact check: they may be statements of opinion rather than factual claims, news articles about widely accepted events, or random links that are nothing to do with

factual claims at all (there was a period when there were a surprising number of Mr Bean videos.) This does not seem unusual to us; it is roughly what we would expect at this stage since launch, especially as user behaviour in terms of flagging, and the precision of Facebook's algorithms in terms of identifying useful signals, may both need time to adjust.

"Attaching" a fact check

Once we have researched, written and published our fact check on our website, the Third Party Fact Checking programme enables Full Fact to "attach" the fact check article to the content on Facebook. This is through an interface on the queue in which we include:

- the URL of our article
- one of nine possible ratings chosen from a drop-down menu (see below)
- and a brief headline with a rating statement at the front (e.g. "FALSE" – although this text is not restricted to the exact wording of the ratings in the drop down menu, and we have occasionally used other phrases such as "context needed".)

The same fact check can be attached to more than one piece of content (for example, if the same claims appear in multiple posts).

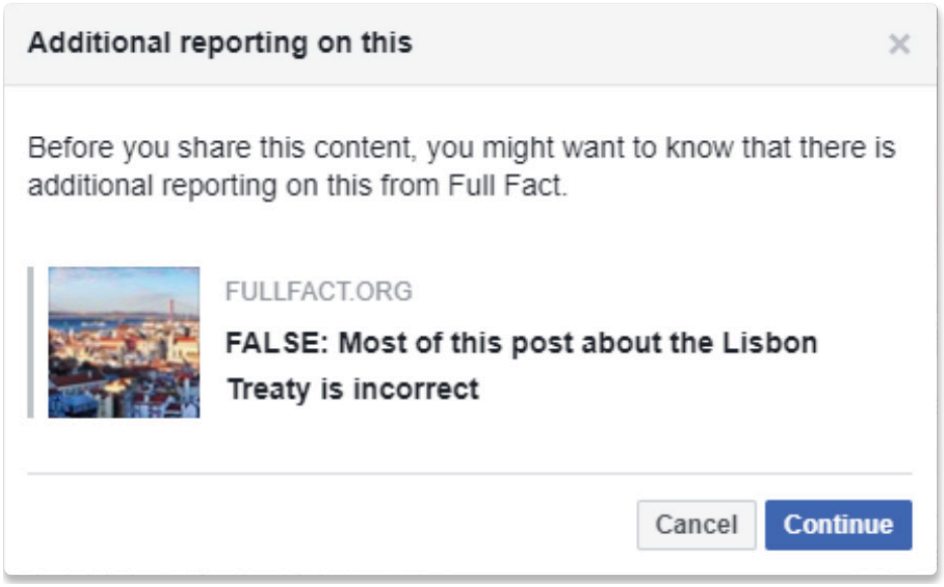
There is also an option to have Facebook apply the fact check automatically to "identical content", which we understand to mean only genuinely identical content – the exact same image or exact same text.

What happens next

Depending on the rating applied by us, Facebook may take additional action – for example, reducing the distribution of the post if it has been rated false.

Users who then want to share the post that we have fact checked will receive a notification about our "additional reporting" on the topic, which includes the short headline we added when attaching the fact check, and a link to the fact check on our website.

If they still want to share the post, they can click "continue" and will be able to share it.



Ratings

There are nine possible ratings fact checkers can apply to content under the programme: False, Mixture, False Headline, True, Not eligible, Satire, Opinion, Prank generator, and Not rated.

Of these, only False, False Headline and Mixture are used by Facebook to reduce the distribution of content, and to notify users if they have shared something that has been fact checked.

The following is how Facebook describe each of these ratings:

False: The primary claim(s) of the content are factually inaccurate. This generally corresponds to "false" or "mostly false" ratings on fact-checkers' sites.

Mixture: The claim(s) of the content are a mix of accurate and inaccurate, or the primary claim is misleading or incomplete.

False Headline: The primary claim(s) of the article body content are true, but the primary claim within the headline is factually inaccurate.

True: The primary claim(s) of the content are factually accurate. This generally corresponds to "true" or "mostly true" ratings on fact-checkers' sites.

Not eligible: The content contains a claim that is not verifiable, was true at the time of writing, or from a website or Page with the primary purpose of expressing the opinion or agenda of a political figure.

Satire: The content is posted by a Page or domain that is a known satire publication, or a reasonable person would understand the content to be irony or humor with a social message. It still may benefit from additional context.

Opinion: The content expresses a personal opinion, advocates a point of view (e.g., on a social or political issue), or is self-promotional. This includes, but is not limited to, content shared from a website or Page with the main purpose of expressing the opinions or agendas of public figures, think tanks, NGOs, and businesses.

Prank generator: Websites that allow users to create their own "prank" news stories to share on social media sites.

Not rated: This is the default state before fact-checkers have fact-checked content or if the URL is broken. Leaving it in this state (or returning to this rating from another rating) means that we should take no action based on your rating.

Overview of what Full Fact has done in Jan–Jun 2019

Fact checking

In January we attached ten fact checks to 16 pieces of content on Facebook's fact checking queue. In June, we attached 19 fact checks to 58 pieces of content. **All the content we've written as part of the Third Party Fact Checking programme can be viewed at fullfact.org/online**.

Of the 96 fact checks we've published as part of the Third Party Fact Checking programme up to 1 July, 59 rated the claim(s) as 'false', 19 were rated 'mixture', seven were rated 'opinion', six were rated 'satire' and five were rated true. None have been rated as 'false headline', 'not rated', 'not eligible' or 'prank generator' yet.

Over the six months, **one claim on the queue was deleted before we could attach our fact check to it**. That was a post on wind turbines not taking as much energy to build as they release.

There was no situation that we treated as a 'Major Incident' (a breaking news event such as a terrorist attack requiring urgent fact checking) in this period.

Developing operating guidelines

All of Full Fact's editorial work is governed by editorial guidelines to ensure we meet our charitable standards. The Third Party Fact Checking programme needed specific operating guidelines

During this period, every post has been reviewed through our normal review process, which involves the claim being fact checked and then the fact check being reviewed, including sources, methods, and for example calculations, by one or more other fact checkers.

Additionally, in these six months each post has also been reviewed by our Editor before publication and during the period in which we were developing our operating guidelines they were also reviewed by our Chief Executive. We have held regular discussions among the whole editorial team, and with the Chief Executive, to consider hard cases and lessons learned.

Although this has been time consuming, it has provided a solid basis for ensuring that we take a robust and consistent approach to the editorial challenges of the programme.

This work and experience has led us to develop robust operating guidelines that will allow us to work quickly while securing our charitable standards. The draft guidelines are included in this report and we welcome feedback.

We have also developed an initial Major Incident procedure, which is included in this report.

Liaising with Facebook and other fact checkers working on the programme

Full Fact takes part in calls with Facebook and other fact checkers working on the programme, organised by Facebook. Facebook also organises regional meetings for the same purpose. We attended a meeting with European fact checkers in April.

We regularly liaise with other fact checkers separately from Facebook to discuss our experiences and learn from one another as well.

Assessing and reporting on the Third Party Fact Checking programme

Full Fact committed to reporting regularly on the operation of the programme when we began work in January 2019. The first of these reports was unavoidably delayed due to staff absences, with the result that we took the decision to combine the reports on the first two quarters into one. We will be releasing reports quarterly from this point on.

We have devoted time to internal discussions of what we are learning and to producing this report, which we hope is of value to Facebook, to other internet companies, and to anyone seeking to scrutinise their or our work.

We anticipate that future reports will be shorter once Full Fact's work on the programme is in more of a steady state. However, we hope and expect to see continuing improvements in the operation of the programme.

We are grateful to Facebook for agreeing to this condition of our participation. It is important and necessary for Full Fact as a charity

that exists for the public benefit to be transparent and accountable about our assessment of the public benefit of the work.

Building networks

Full Fact's experience of fact checking is that our work is most effective when we work closely with people and organisations with deep subject expertise. This allows us to be faster, more rigorous, and more comprehensive. In other contexts we have worked closely with leading academic experts in different policy areas such as the Institute for Fiscal Studies, Oxford University's Migration Observatory, and the UK in a Changing Europe project from the Economic and Social Research Council.

Fact checking online content, including but not limited to the content we see under the Third Party Fact Checking programme, has taken us into subject areas where we need to broaden our networks.

Fact checking issues of public health, of the kind that often arise on Facebook (rather than claims about health policy, the NHS and so forth) goes beyond Full Fact's established in-house expertise. We have therefore begun setting up meetings with different expert organisations who might be able to help ensure our content is relevant, timely and that we're targeting the biggest problem areas for health misinformation.

We would welcome contacts from any organisation that might be interested in working with us, particularly in the field of public health.

So far we have had conversations with among others Alzheimer's Society, Anthony Nolan, and the Vaccine Confidence Project at the London School of Hygiene and Tropical Medicine. We've also reached out to dozens of other organisations and are working to identify more we can approach to help us in our work.

We had an exploratory meeting with representatives from the Association of Police Communicators (APCOMM) to discuss how we might establish lines of communication in the event of a major incident and in due course we may update our Major Incident procedure if we create any formal mechanism for doing so. This would be reported in our quarterly report.

We are concerned that we are finding areas where it is hard to find sources of impartial and authoritative expert advice, especially from organisations that are capable of responding in time to be relevant to modern online public debate. We address this in the recommendations.

Funding

The total fees Full Fact has earned from Facebook for work on the Third Party Fact Checking programme during Jan–Jun 2019 is $171,800.

The amount of money that Full Fact is entitled to depends on the amount of fact checking done under the programme.

After completing our first three months of work on the programme, and having developed our editorial approach to the project, in April Full Fact hired one new fact checker to add to our existing team's work.

Observations from our work

Specific topics of interest

Health

At least 18 pieces have come under the general banner of health as part of the Third Party Fact Checking programme: on subjects from **side effects of the pill** and **whether chemicals in bath products can induce labour**, to emergency scenarios like **whether cough CPR works**, and **whether a tampon can help someone who's been stabbed**.

We have often found it difficult to get answers on these health claims, and had a particular case where **we were bounced between 13 different press offices trying to get to the bottom of the Radox and labour claim**.

Vaccine-related claims have been the most numerous health-related claims in the queue. These often require specific expertise which goes beyond Full Fact's in-house expertise, so in the first six months we focused on building up connections with experts in relevant area. This should improve the quality and speed with which we can fact check vaccine-related claims going forward.

Police

Several claims appearing multiple times on the queue (**this 999 call image is misleading** and **two pieces on speed limits**) involved contacting the police to fact check claims circulating online with limited evidence. We suspect more of the Third Party Fact Checking work will involve research of a similar nature.

Some case studies that have informed our recommendations

Satire

One common problem we had was around humorous posts, which many people may have misunderstood as being real. **At the launch of our participation in the programme**, we had said in multiple blog posts that "We'll only be checking images, videos or articles presented as fact-based reporting. Other content, like satire and opinion, will be exempt." This was badly phrased: we should have said we'd be checking all these types of content, but satire and opinion are

exempt from having their distribution on news feeds impacted because of our ratings.

We fact checked one post that claimed (as a joke) that **the BBC was adding Arabic subtitles to EastEnders** – many readers had seemingly interpreted it as a real news story. While in the end we were actually unable to attach our fact check to the content on the queue because there were some technical issues with the queue, **we nonetheless received some push back from the original piece's publisher** who felt we should not have fact checked it at all. (There may be a need to communicate more clearly that the satire rating does not reduce a post's distribution – indeed it is a signal to Facebook that they should not take action against the content.)

Most people wouldn't call **the video purporting to show a police officer taking drugs** satire, but that is how we rated it. The video was filmed as a joke, so giving it a rating that would damage its distribution seems inappropriate. Some commenters and the person who'd posted it (who wasn't the original creator) did seem to think it was legitimate, and it had been shared over 34,000 times. Satire seemed the best rating, as its distribution would be unaffected and it would acknowledge in some way that the content was created for humour rather than to mislead. Going forward, rating jokes (or more widely people messing around online to be funny) as satire, is not ideal. We discuss this further in our recommendations.

Opinion

We rated **a claim comparing the population of Iceland and the number of homeless people in the UK** as opinion. This was due to lack of a better rating, rather than us thinking the statement itself was what Facebook probably intended the "opinion" rating to be used for. The fact check itself addressed the claim that "there are now more UK citizens homeless than the entire population of Iceland".

Our conclusion, in short, was that the two numbers are likely in the same ballpark. (The best available estimate from Shelter on the number of homeless people puts homelessness in Great Britain at 320,000 while the population of Iceland is around 360,000; the Shelter estimate is likely a low estimate due to the difficulty of collecting robust data on this issue). Therefore, it's not possible to state definitively that the claim is true— but because it's based on a likely underestimate, doesn't include Northern Ireland, and the numbers are in the same territory, we felt it was a case where it was possible to have different reasonable

interpretations of the same evidence. As such, it would have been disingenuous to give it a false or mixture rating and see its distribution reduced as a result.

So we went with opinion, which means the post doesn't get reduced distribution and users trying to share don't get prompted with our reporting. Our fact check would appear in 'related articles' below the post, with the message "CONTEXT: The number of homeless people in Britain is broadly comparable to the population of Iceland". We discuss the need for a rating that reflects such situations more in the recommendations.

We used the opinion rating in another piece, which looked at **whether the NHS is "free for all 500 million EU citizens"** because, as we wrote, whether the claim is correct or not comes down to whether you interpret "free for all" as meaning in certain circumstances or in all circumstances.

We used opinion again for a piece where the rating came down to **whether or not votes for Labour in the 2019 EU elections could be interpreted as votes to Leave the EU**.

The burden of proof being on the claimant

We fact checked **an image claiming a woman in Sweden had been attacked by a Muslim migrant**, which we rated as false. While an attack did take place, we established – after speaking with journalists in Sweden – that the identity of the attacker remains unknown, and there was no evidence that he was a Muslim or a migrant.

The Facebook guidelines suggest rating unproven claims as "mixture", and we are naturally wary about describing a claim as "false" when we do not have positive evidence of its falsity. But in this case, especially given the harm that can result from this type of misinformation, we decided that the burden of proof should be on the claimant. In effect, in stating the identity of the attacker with certainty despite there being no evidence behind that part of the claim, the claim is falsely asserting knowledge where no such knowledge exists: in the end we decided on a "false" rating.

Our view of the Third Party Fact Checking programme

This section represents Full Fact's view of the Third Party Fact Checking programme. We do not speak for Facebook, who may take a different view, or for any other fact checker participating in the programme.

In brief –

- The Third Party Fact Checking programme includes some work of clear social value that can at its best help to save lives, if it can achieve the necessary scale.
- A lot of the work has at least clear value to Facebook in creating better environments for its users.
- The Third Party Fact Checking programme may play an important role in generating the data to make new technologies for reducing harms from inaccurate information possible, but at the moment we know too little about plans for using that data. We call below for Facebook to make more data available to fact checking partners.

Full Fact recognises that there are multiple different ways in which fact checking can be beneficial. It may be that it reduces the immediate spread of false information (as seems to be the primary goal of this programme). But it could also – for example – reduce people's belief in false information that has already spread, it may improve broader understanding of issues, it may reduce the likelihood of similar misinformation circulating in the future, it can reduce long term incentives for actors to spread misinformation (the "they know we check" effect), and it should perform an educational role in giving people a toolkit to make assessments of information themselves.

We feel that all these modes of action should be considered when assessing the possible impact of the programme.

Full Fact sees an important distinction between intervention on specific topics where there is clear harm associated with inaccurate information (such as elections, health, and during emergencies) and what could be described as a wider 'spam filtering' function covering inaccurate content more generally. Both of these are valuable but the task and benefits are different in each case.

Tackling specific harms

We believe that the Third Party Fact Checking programme can be valuable in helping to tackle specific harms from inaccurate information.

We have already seen cases in the first six months of our work in the programme where we have helped to address posts circulating that represent potential risks to life, or to people's health and wellbeing.

As we have said before, we also believe there is a clear, specific and valuable role for the programme in responding to emergency situations, and in tackling attempted election interference.

Some of the content most clearly addressing specific harms includes –

- A claim wrongly suggesting that, if you cannot speak on a 999 call, pressing 55 will allow the police to track your location (a misunderstanding of the "silent solutions" scheme to help police distinguish genuine emergency calls from accidental dialling.)
- A claim promoting the idea of "Cough CPR" – that if you are suffering a heart attack, you should cough repeatedly in order to keep your heart beating (medical authorities do not recommend this).
- A claim saying that if you are stabbed, you should "whack" a tampon into the stab wound, as this will stop the bleeding (first aid experts we spoke to said that it likely would not be effective at this, and could lead to further problems).
- A claim saying that taking a pregnancy test could be used to "check for testicular cancer if you are unsure of lumps and bumps". Cancer Research UK told us they definitely wouldn't recommend relying on a pregnancy test to self-diagnose testicular cancer, as it wouldn't come up positive in all cases of the disease.
- A claim saying that type 1 diabetes is listed as a side effect of the MMR vaccine. It's listed as an adverse reaction of the vaccine used in the US (which isn't the same as a side effect, it can refer to conditions developed by chance after someone was vaccinated but not caused by the vaccine).
- A comprehensive guide to some of the main claims made about the ingredients in vaccines, the countries they may or not be banned in, whether they are harmful and in what amounts.

However, we have two important points for further work.

The first is an operational point, that we suspect that there must be more of this kind of content than we are currently seeing or able to fact check under the Third Party Fact Checking programme. We hope that we can work with Facebook to identify and prioritise more of this kind of valuable work under the programme. Recommendation 1 reflects this.

The second is a longer-term strategic point, that we need to develop a plan for taking this kind of work to internet scale. We are keen to work with Facebook and others to achieve this while maintaining high standards of accuracy, balance, and accountability for the public benefit.

'Spam filtering'

There is another category of content which we regularly see as part of the Third Party Fact Checking programme, which is content which may be inaccurate or misleading but where the stakes are not so high as to risk life. It may be a nuisance or simply content that reduces the quality of experience on Facebook. It may even be inaccurate content which is harmless, and obvious, and which people enjoy.

We would not prioritise fact checking this kind of content within Full Fact's fact checking, but we recognise that it has value in creating better environments for internet users, particularly as Facebook and others seek machine learning approaches to tackling content quality questions at scale. Our operating guidelines discuss how we will approach these fact checks but in brief while of course we will publish them to be transparent we will not normally promote them through Full Fact's own channels.

The role of technology

Full Fact has pioneered the use of technology and AI to make fact checking more effective. Our work on automated fact checking has been described as "seminal" and our tools have been used on three continents, and with our partners AfricaCheck, Chequeado, and the Open Data Institute, Full Fact won the Google AI Impact Challenge for our work in this area to use AI for social good.

We understand the need for Facebook (and other internet companies) to be able to make decisions about how all content is displayed within their products. One of the factors influencing these choices needs to be the likelihood of spreading inaccurate information, whether that

information is harmful or simply in this context a nuisance.

Understandably, internet companies are looking for technologies that can identify inaccurate information at internet scale. Facebook has publicly suggested that "we're going to shift increasingly to a method where more of this content is flagged up front by A.I. tools that we develop", as Mark Zuckerberg said before the US Congress. Other internet companies are certainly working in the same area.

These systems do not yet exist in any general sense. Creating these technologies involves solving some very hard problems, including ethical as well as technological problems. And attempts to do so need to be carefully scrutinised, which is one role Full Fact plays in this area.

The Third Party Fact Checking programme may play an important role in generating the data to make these new technologies possible, but at the moment we know too little about plans for using it.

We believe that AI can be useful in identifying content and patterns of inaccurate content that may lead to specific harms. The queue Facebook provides to fact checkers under the Third Party Fact Checking programme is an early example of this. Effective and ethical technology could in time help to make human efforts to tackle specific harmful inaccurate information more effective by identifying and classifying it at scale.

However, machine learning depends on the data it learns from and we doubt that the existing ratings system is likely to produce a high quality outcome from machine learning. The categories are too broad for us to be confident that they have specific statistical qualities that distinguish them. Computers do not understand language or images and it is not obvious that what makes one post on a subject true and another on the same subject false is something a computer can pick up from the data the programme is generating.

It is possible that Facebook has information that we are not aware of that makes it confident that it can generate effective machine learning approaches without serious negative side effects. For example, they might be using data about the actors behind particular posts or groups of posts as well as data on the content of the post itself.

We would welcome a clearer statement from Facebook of the potential avenues they see for developing machine learning tools based on the Third Party Fact Checking data. We believe that our domain expertise could help make those efforts more effective and help to avoid

negative side effects or unintended consequences. We recognise that this discussion might have to be private because revealing details of plans to develop technology to prevent abuse can help people bypass those safeguards. However, at the moment no such discussion has taken place in public or in private.

Full Fact is glad to be part of a group of platforms, academics and practitioners organising **a conference called Truth and Trust Online** in October, working with all parties working on automated approaches to augment manual efforts on improving the truthfulness and trustworthiness of online communications. The organising committee includes representatives from Full Fact, Amazon, Facebook, Google, Microsoft, and Twitter, as well as from academia and elsewhere. The call for papers is now open.

Recommendations for Facebook and others

Improving the information and tools available to fact checkers

In deciding which posts to fact check, we have access to Facebook's "queue". This provides a list of posts which have been flagged by users or Facebook's algorithm as potentially inaccurate. It indicates when a post was first shared, when it was flagged to the queue, and how many shares it has received. All these factors feed in to what we decide to fact check.

We have made recommendations for how the queue could develop to improve decision-making processes for fact checkers. In addition, we have one recommendation for how to increase the reach of fact checks we publish.

On a practical note, we have had some issues with posts we've 'bookmarked' on the queue, then fact checked, later disappearing so we cannot attach our fact checks to them. This has happened in three cases.

Recommendation 1: Continue developing tools that can better identify potentially harmful false content including repeated posts

We suspect that there must be more potentially harmful false content than we are currently seeing or able to fact check under the Third Party Fact Checking programme. We hope that we can work with Facebook to identify and prioritise more particularly harmful content, such as that relating to public health, under the programme.

Once we submit a rating for a piece of content in the queue, there is an option to allow Facebook to automatically apply that rating to other, identical, posts (for example, identical images). This is valuable, but limited by the tendency of content to subtly change as it goes viral. The viral process often sees the same text or image shared in varying ways - where the language and layout of a post is similar to the original, but not identical. In its most literal sense, this includes people sharing different screenshots of a post on one social media site onto other sites.

Take this post about **the Lisbon Treaty**. We received an unprecedented number of reader requests to check this claim, which appeared all over Facebook (and other social media) but often with slight variations in wording or layout. We rated two **posts** (one of which has since been deleted) with around 2,000 shares between them, yet we know that there are many other versions on Facebook, **some with far more shares**. But the process of identifying these manually is time consuming and imperfect.

This is a repeated pattern we see with online misinformation (we observed the same thing in posts about **Shamima Begum**, **999 calls** and **harmful dog treats**).

Although there is no quick fix in identifying similar but not quite identical content, we suggest that Facebook continue to make developing the tools to do this a priority. We were pleased to see that in the second quarter of the year Facebook did introduce a feature that suggests possibly related content for posts that have already had fact checks applied to them. While its effectiveness is currently limited (we will assess it more fully in our next report) it is a positive step.

We hope it will improve, and that Facebook will continue to develop more tools to enable fact checkers to search for and surface similar content. In addition to discovering content related to that which they have already fact checked, it would be valuable to have tools to better search for prior examples of identical or similar content during the research phase (knowing where and when a claim originated is often important context for fully understanding it, and may in fact change our conclusion – for example in the case of claims that are now outdated but may have been accurate when they first started).

Without such tools, the Third Party Fact Checking programme risks only addressing the tip of the iceberg. The reach of our content could grow rapidly with effective tools in place for better identifying similar posts.

Recommendation 2: Provide more data on shares over time for flagged content

We recently checked a post claiming that a bath product was harmful for pregnant women. It had **an exceptionally high number of shares (over 100,000)**, which was a primary reason for checking it. But it was also around a year old which means it may have no longer been getting very much reach online. Often things go viral in waves, or simply stop circulating after a while; so it would be highly valuable to have data on not just the number of shares, but when those shares

happened. We were pleased that in the second quarter of the year, Facebook rolled out changes based on user feedback that do provide some more insight into this (showing how many shares the post has received in the past day, in addition to total shares), which is a very welcome and positive step.

However, both for fully understanding the context of a post's history and how rapidly it is currently spreading (and thus being able to prioritise what to check better), and for being able to assess the impact that our fact checks have on a post's virality, we would need fuller data on how the post accrued shares over time, provided in a usable – and ideally downloadable – format.

Developing the Third Party Fact Checking programme ratings system

For a number of the posts we fact checked, we found the existing rating system to be ill-suited. Below are four ratings we recommend adding, with case studies to explain why they are necessary. Three of them are related to a central observation about the inadequacy of the 'Mixture' rating; the fourth to the fact that the 'Satire' rating is the only way of labelling much humorous content.

The 'Mixture' rating is not fit for purpose (encompasses recommendations 3-5)

The 'Mixture' rating – which Facebook suggests should also be used to cover cases that could be described as 'unproven' – is insufficient for all the purposes it is being used for. As the only rating that currently sits between the poles of unambiguously 'True' or 'False', it could potentially be applied to a majority of the posts we check, but fails to accurately describe many of these situations. We also feel it can be over-punitive, as we understand that content rated as 'Mixture' will have its distribution significantly reduced.

Recommendation 3: Add a 'Mixture' rating which does not reduce the reach of content

Facebook defines the "mixture" rating as "a mix of accurate and inaccurate, or the primary claim is misleading or incorrect". In some cases the overall message of a post is broadly correct, but some of the finer details are not, to the extent that we would not feel comfortable as a fact checking organisation endorsing it as "true". This means it should technically be categorised as mixture, but the reduction in

circulation of a post that comes with this rating seems excessive given that much in the post is correct.

Given that there are certain circumstances in which "mixture" is the only reasonable rating to apply, but that it would not seem appropriate for the post to have its distribution reduced as a result, we recommend that Facebook introduce a rating akin to "mixture", but which doesn't reduce the reach of that content.

Case study: The post (since deleted) that said "There are more UK citizens homeless than the entire population of Iceland." As noted before, we felt that **because it's not possible to state definitively that the claim is true we could not rate it as such**, but as we felt it was a case where it was possible to have different reasonable interpretations of the same evidence, and the best evidence suggested that the numbers were in the same ballpark, rating it as mixture or false also seemed wrong.

Recommendation 4: Add an 'Unsubstantiated' rating

In some cases, we cannot definitively say something is false, but equally can find no evidence that it is correct. Facebook suggests that the "mixture" rating can be applied to "unproven" claims, but this is an insufficient response in cases where there is absolutely no substance to a claim (as opposed to cases where the evidence is genuinely mixed or unclear). A rating of "mixture" gives such baseless claims more credibility than they deserve by implying that there is some degree of truth in them.

In such cases the burden of proof should rest with those making the claim. This is particularly the case in situations when evidence should be findable if the claim were true. In these situations, if there is no evidence for the claim, it should effectively be considered as being close to, or even equivalent to, false.

To this end, the definition of "false" could possibly be expanded to include unevidenced assertions (even when they cannot be definitively disproved), although retroactively changing definitions may be a problem for consistency. But we believe a better option is to introduce a new "unsubstantiated" category, which Facebook can treat as a signal akin to a "false" rating. The additional merit of a separate "unsubstantiated" category is that it would allow users to better distinguish between content that has been debunked as false, and content for which there is simply no evidence. This rating seems particularly relevant in cases of terror attacks or other emergencies, where a lot of unsubstantiated rumours quickly start circulating online.

Case study: The post claiming that a Swedish woman was attacked in a nightclub by a Muslim migrant. Despite the guidelines suggesting an unproven claim should be rated as "mixture", we rated it is "false" due to **the complete lack of evidence for the claim that he was a Muslim or a migrant**, and a consideration of the harm that can result from this type of misinformation.

Recommendation 5: Add a 'More context needed' rating

In some cases we cannot definitively rate something as true, false, or even mixture, but we could still add more context to help a reader. This would make them more informed before they choose whether or not to share the piece. There is currently no category for this purpose.

The highly specialist – as well as occasionally ambiguous or provisional – nature of much medical evidence is one reason why we are recommending to Facebook that a "context needed" rating might be necessary. For example, we often see posts that discuss the listed side effects of various medicines, in a way that implies they are inherently dangerous. These may be technically accurate, but potentially misleading without the context of relative risks and regulatory processes.

Case study: This post lists **potential side effects of one brand of contraceptive pill**. Most of them are accurate, in the sense that they are listed as potential side effects, but it could well be interpreted in ways that overstate the risk. We rated it as "true", as we did not feel it was inaccurate enough to justify even a "mixture" rating; however, we believe that a "more context" rating would have been more appropriate.

Case study: This post **claims to have calculated the total size of the People's Vote March in London**. The **assessment of the expert we spoke with** was that the total number is likely to be higher than their estimate, but we cannot say this definitively. Due to the lack of appropriate rating, we did not rate it on Facebook, even though we could add valuable context for a reader.

Recommendation 6: Add a rating for humorous posts other than satire or pranks

Facebook's definition of satire is "a page or domain that is a known satire publication, or a reasonable person would understand the content to be irony or humour with a social message". But a lot of the time Facebook posts are quite simply jokes, or more generally just messing about, intending to be funny without any social message.

These can then get picked up by people who miss the point of the joke, or encounter it out of context, and share it believing it to be real. It would be helpful for Facebook users to be able to distinguish these kind of jokes—which don't have a satirical message but get misconstrued online as real—from actual satire.

Facebook also has a rating of "prank generator" for websites that allow users to create their own humorous fake news stories, which likewise is a specific instance of the more general category of "jokes". (We have not seen any examples of these in the queue to date.)

Given that satire is important in a democracy, we can see the value in having a specific rating for it – both to enable Facebook to better identify it and protect it from being treated as false news, and to give better information to Facebook users who may have taken it as real. But that means there should also be a rating for the broader category of non-serious, lighthearted or humorous posts that people might misunderstand. Like the "satire" rating, this should not reduce the reach of the post. It is not our job to judge the quality of people's senses of humour.

Case study: this **viral video of a man dressed up as a police officer and appearing to snort drugs**. The video was originally posted as a joke, but **many people sharing it thought it was real**. We rated it as 'satire', but that seems like quite a stretch.

Resolving editorial questions around the programme

Recommendation 7: Develop clearer guidance on how to differentiate between several claims within a single post

The current ratings system offers little guidance on how to prioritise a single/the most important claim within a post. In some cases, there is a risk that a post which contains a complete falsehood—with the potential to cause harm—could end up being rated "mixture" on the grounds that it got some less important details correct. We strongly feel that it is advisable to focus on the most prominent/harmful claim in such cases, and clearer guidance on how to differentiate between several claims within a single post would be welcome.

Case study: This same post claiming to show a Swedish woman who was "savagely beaten by a Muslim migrant" after asking him to stop groping her. The post is correct in as much as **it does show a Swedish woman who was beaten up in a nightclub after stopping a man from groping her** - but we don't know if he was a Muslim or a migrant. In

rating it false (for reasons outlined above), we decided to focus on the claim about the attacker being a Muslim migrant, as this was clearly the most notable claim and the main reason for its online circulation. However, we could have rated it "mixture" on the grounds that much of the information about the woman and the attack was correct.

One possible approach in the future might be to enable fact checkers to apply multiple ratings to content, so that individual claims can be better separated out. Currently the rating can only be applied to the content as a whole (be it a link, a text post or an image).

Making it easier to evaluate our work on the programme

Recommendation 8: Share more data with fact checkers about the reach of our fact checks

Currently, the only sense we have of how many people our fact checks are reaching comes from data on visits to our own website. But the Third Party Fact Checking programme brings with it a number of new ways in which people can read our content. In addition to the traditional ways we reach people—on our site, via our social media feeds, and via search engines—Facebook users may also see our fact checks if they engage with a post we have rated, and may for example get a notification linking to our fact check before they try and share something we have rated as misleading or false.

It would be helpful to understand how effective the additional ways that Third Party Fact Checking programme fact checks reach Facebook users are. Does the notification stop many people from sharing? What percentage of people who view a post we have rated click on our fact check beneath it? Are there cases in which our content gets many more interactions from Facebook users, and what does this tell us about how to effectively get the attention of Facebook users in future?

Other fact checkers, speaking to the BBC, have said **they want more data about the reach of their work**, so they can assess its value.

Expanding and developing the programme

Recommendation 9: The Third Party Fact Checking programme should expand to Instagram

We believe the Third Party Fact Checking programme should be expanded to other platforms: most immediately, Instagram (which

is owned by Facebook). The potential to prevent harm is high here, particularly with the widespread existence of health misinformation on the platform. Facebook have already taken some steps towards using the results of the programme to influence content on Instagram, or Instagram images that are shared to Facebook. However, directly checking content on Instagram is not yet a part of the programme.

Recommendation 10: be explicit about plans for machine learning

We would welcome a clearer statement from Facebook of the potential avenues they see for developing machine learning tools based on the Third Party Fact Checking data. We believe that our domain expertise could help make those efforts more effective and help to avoid negative side effects or unintended consequences. We recognise that this discussion might have to be private because revealing details of plans to develop technology to prevent abuse can help people bypass those safeguards. However, at the moment no such discussion has taken place in public or in private.

Recommendations for government

Recommendation 11: The government should review responsibilities for providing authoritative public information on topics where harm may result from inaccurate information and fill gaps

As we argued in our paper "**Tackling Misinformation in the Open Society**", we believe that public bodies should be given a clear mandate to inform the public, in order to build resilience against misinformation.

In our work on the Third Party Fact Checking programme already, we have seen multiple examples of a related issue: major areas of public interest in which no body has primary responsibility for providing accurate and useful information.

One obvious area is matters of public health. In one example, our attempts to fact check a claim about the safety of a bathroom product for pregnant women saw us bounced repeatedly between the press offices of 13 different public bodies, all of whom believed that providing such information was somebody else's job. Similarly, we've had inquiries regarding the introduction of 5G technology in the UK, and there's a distinct lack of official guidance properly addressing some public concerns. In a recent debate, an MP expressed dismay at

Public Health England's "standard reply" to questions about 5G.

We have seen this in multiple cases relating to health issues.

Another area is law, in which there is no public body with a clear duty to provide information on the functioning of the legal system.

The lack of such authoritative sources has practical consequences: notably, it dramatically slows down the speed with which organisations such as ours can respond to misinformation (some of these fact checks can end up taking weeks). It also means that the final product may be less authoritative and useful to the reader.

Most importantly, the absence of reliable and trustworthy information can create a vacuum in which misinformation is better able to spread.

Establishing bodies with clear duties for providing impartial information in areas of public concern would have clear benefits. This kind of public service could potentially be provided by a wider range of public service institutions depending on the topic. It could be government itself (for example, when it comes to the law this could **build on the work on public legal education** already overseen and supported by the Solicitor General); trusted and independent public bodies such as the NHS (their **Behind the Headlines** service is a good example); or academic initiatives with a specific communications role and resources (where successful models include the **Institute for Fiscal Studies**, the **Migration Observatory** at Oxford University, and **the UK in a Changing Europe** initiative).

Future work for Full Fact

Our priorities are to increase our output under the Third Party Fact Checking programme and to further develop our links with relevant expert organisations to ensure that our work on the programme has the greatest possible public benefit.

As mentioned in 'Our view of the Third Party Fact Checking programme', we are keen to work with Facebook and others to find ways to help increase this work to internet scale.

One relevant question – as discussed briefly in the recommendations – is why any of Facebook's programmes, including the Third Party Fact Checking programme, should be restricted to Facebook alone? It is clear to us that this work could have value on other platforms, including (but not limited to) other platforms owned by Facebook.

Facebook have already said that they are **testing using ratings applied to images** under the Third Party Fact Checking programme to influence the discoverability of identical images on Instagram. In March, Facebook announced that content from other media sites (ie Twitter, YouTube) is now eligible to be checked as part of the Third Party Fact Checking programme. That means we can check tweets, Youtube videos, Instagram posts, etc, but our supporting articles will only appear (or impact a post's distribution) if links to these are shared on Facebook.

However, as we've said, the ability to directly check content on Instagram directly is not yet a part of the programme.

Facebook have also recently said that vaccine misinformation will no longer appear on Instagram Explore or Hashtag pages. This may prevent users inadvertently coming across antivax content initially, but will do little to help those already in the community.

We do not see why the Third Party Fact Checking programme cannot be fully expanded to Instagram. The potential to prevent harm is high here, and there are known risks of health misinformation on the platform.

We have noted Facebook's public discussion of increasing the role of crowdsourcing in understanding information quality on its platform. We will be studying their ideas carefully and engaging with Facebook in those discussions.

Finally, we will continue to work on technology to tackle harmful inaccurate information for the public benefit, and to scrutinise work in this field.

Appendix: Full Fact's Operating Guidelines for the Third Party Fact Checking programme

These operating guidelines are an evolving document; we may change them over time as we learn more about how the Third Party Fact Checking programme works, and as we encounter difficult or edge cases that challenge our thinking. We will discuss these changes in future quarterly reports.

In all cases, when we encounter a situation that the guidelines do not cover, staff should consult the Editor (or in the Editor's absence, the Chief Executive). The Editor may consult the Chief Executive at any time, and the Chief Executive is ultimately responsible for upholding Full Fact's standards.

Any changes to these guidelines will follow discussions between the Editor, the editorial team, and the Chief Executive. They must ultimately be agreed by the Editor and the Chief Executive.

Background: general operating guidelines

We have a set of standards for our pre-existing fact checking work, and most of these have translated across to our work as part of the Facebook Third Party Fact Checking programme. They underpin these operating guidelines, which should be read in that context.

As with all charities, Full Fact is legally required to work for the public benefit and to be politically non-partisan. Our legally-binding charitable objectives go a step further than this, requiring us to work "in an impartial, objective, balanced and independent manner observing strict political neutrality". These principles apply equally to our work on the Third Party Fact Checking programme. We monitor our work to ensure that both our processes and our output meet these criteria; that includes our selection of which claims to fact check.

The Third Party Fact Checking programme is also governed by systems and guidance set down by Facebook, for example the choice of ratings that Facebook provides. We must operate within these and we will publish quarterly reports on our experience of the programme and how it might continue to develop.

What we check, and why

In addition to our balance and impartiality requirements, when selecting claims to check normally we have a rule of thumb—that what we check should be some combination of important, influential and interesting.

- "Important" here means that the issue has real-world impact — something that can affect people's lives and choices.
- "Influential" means that the claim is likely to reach and affect a large number of people, and potentially influence their beliefs (which could include, for example, if it was said by a public figure, if it appeared in the national media, or if it was widely shared online).
- "Interesting" means just that: that the question of whether the claim is accurate should be something that will engage an audience, or illuminate a broader issue. (For example, we generally avoid checking statements that are trivially true.) One possible guide for this is the volume of requests from our readers to fact check a particular claim, but we must take care to maintain our independence when considering any external requests.

Not everything we check will necessarily hit all three of these, but (in our work outside the Third Party Fact Checking programme) if a claim doesn't register on any of them then we would not normally check it.

What Full Fact prioritises

These rules of thumb inform our prioritisation of work in the Third Party Fact Checking programme. Analogously with the "important" measure, we prioritise false or misleading **claims that have the potential to cause harm if they are believed** (such as health misinformation).

The "influential" measure translates into the number of shares a post has received, and also factors such as whether influential pages have shared it, and whether there are multiple versions — we will prioritise **claims that have spread widely**.

The "interesting" measure has slightly less weight here as an independent factor, due to the fact that we also consider the number of shares a claim has as being reflective of the level of interest in the topic, and the presence of the claim in the dashboard queue suggests that some users may have flagged it as suspicious (which for these purposes we treat as equivalent to a reader request). In effect, the expectation that a claim be both interesting and influential are somewhat merged in the online context.

However, there is another context in which the "interesting" measure may influence our prioritisation: we may choose to check some relatively trivial claims if we think that they have value as an engaging way to educate people on techniques for spotting false information online (for example, **a claim about a horse that allowed us to point readers towards our guide on how to spot misleading images online**).

Fact checking other content from the queue

The existence of content in the queue is sufficient evidence that it is useful to Facebook to have that content fact checked, even if Full Fact might not have fact checked it outside the Third Party Fact Checking programme, and is sufficient to justify fact checking and rating that content.

All fact checks under the Third Party Fact Checking programme must be published on **the dedicated page for these fact checks**. However, the extent to which fact checks of this kind are promoted elsewhere on Full Fact's own channels should be determined by our own views of what is interesting and useful to our audiences.

Political actors

According to Facebook's guidance, the Third Party Fact Checking programme is not intended to be applied to "a website or Page with the primary purpose of expressing the opinion or agenda of a political figure". We do not include in the Third Party Fact Checking programme fact checks of claims made on Facebook by politicians, political parties, or non-party national political groups (we may, of course fact check these as part of our general fact checking work). Political opinions are also not subject to fact checking, as is the case with our general work.

Beyond these exclusions, however, there are a range of political actors on Facebook (such as activists, local party accounts or interest groupings) whose posts we should treat sensitively, with a mind to protecting freedom of speech. We do not believe that simply being involved in politics should make you exempt from fact checking or the Third Party Fact Checking programme, nor that simply appending a political opinion to a central factual claim should exclude it from consideration. If a claim originates from a political source but is primarily a factual claim that can be checked, we may do so. We should however be cautious when applying ratings that may reduce the distribution of a post in a situation where the factual claims are not

plainly false (see below for further discussion of how we apply ratings). If in doubt, this should be checked with the Editor before publication.

In addition, inaccurate claims originally made by politicians but that are being shared by third parties (for example, screenshots of a tweet from a politician) are eligible to be fact checked through Third Party Fact Checking. This reflects our principle that we check the claim not the person.

Humour

Much false information online originates from attempts at humour. We don't believe it's our job to judge how funny someone's joke is. We should only prioritise humorous posts in a situation where there is compelling evidence (e.g. from comments or shares) that a significant number of people have mistakenly taken it seriously, and also when doing so would satisfy our other standards for selecting it to check (such as potential harm, or educational potential). Other fact checks of humorous posts for the Third Party Fact Checking programme should not normally be promoted through Full Fact's own channels.

How we check

We check claims, not people

The core of what we check is individual, identifiable factual claims; it is not the people who make them, or the broader positions or opinions they may be advocating. Our conclusions about claims should not normally comment on the motives, intent or character of the person or institution that made the claim. When analysing the spread of specific unsubstantiated claims it may sometimes be appropriate to comment on the actors involved, and it may be necessary to discuss the broader positions they advocate in order to properly contextualise how a claim is likely to have been understood by its audience. If in doubt this should always be checked by the Editor before publication.

We present evidence to allow our readers to reach their own conclusions

We present our own conclusions on the accuracy or otherwise of factual claims, but we always back this up by providing the evidence we have based our conclusions on to the reader (in the form of links to primary or secondary sources). We should always seek out the most authoritative source for any factual statement we make. We should provide sufficient evidence for Full Fact's readers to make up their own

minds and reach their own conclusions from our work. Where there is insufficient quality evidence to reach a firm conclusion, we should tell the reader that.

If we must use evidence that is—for whatever reason—not publicly available, we should say so clearly and explain why; this should be checked with the Editor before publication.

Our work means that we frequently have to make judgements about the reliability of sources in a manner that reflects our commitment to impartiality. In many cases it will be useful to explain those judgements clearly to the reader.

Standards of evidence will, by the nature of things, vary depending on the nature of the claim. For some types of claim (for example those of a statistical nature) there may be independently quality-assessed sources such as national statistics; in other cases (such as claims about historical events) evidence may be harder to come by; particularly in matters around health claims, evidence may be partial or tentative. We should always be cautious, question our sources, and avoid over-interpreting evidence. However, we should not let over-fussy philosophical rigour deter us from reaching clear, useful conclusions: absence of evidence may not technically be evidence of absence, but in many cases it may be close enough for our purposes.

In all cases, we believe that it is the responsibility of the person or institution making the claim to provide the evidence to support it. If they cannot do so and we can find no evidence to support their claim then we should say so.

Health

Misleading health information has clear potential to cause severe harm. The nature of medical evidence is such that it is often impossible to state definitively that something is unambiguously true or untrue. Despite this, we should still aim to give clear advice to our readers and to present conclusions that reflect the best possible current knowledge. This includes assigning ratings such as “True” or “False” when the weight of evidence supports that interpretation. If multiple expert bodies with competency in a particular medical field tell us the same thing, then we should be comfortable passing that on to our readers. However, if there is more than one responsible body of professional opinion, our fact checks should reflect that in a balanced and proportionate way.

How we assign ratings

In our general fact checking, Full Fact is relatively unusual among fact checking organisations in that we do not use any kind of rating system in our published fact checks, as we tend to believe that they can often obscure more than they illuminate, and can be hard to apply in a consistent manner. However for the Third Party Fact Checking programme **we are required to apply one of the following ratings**; what follows is our current thinking on how these should be applied. In all cases, if there is a question about the rating being applied, it should be discussed with the Editor before publication.

True

We have only checked a small number of true claims, as our prioritisation of potentially misleading claims that could cause harm means that they are not our top priority. We would apply this in situations where we are confident the central claim or claims are unambiguously correct, or are close enough to being accurate that a reasonable person would not feel it necessary to correct them. (For example, minor imprecision on figures, or information that might be slightly out of date but is still substantially true.)

Mixture

This is a complex rating: it applies to posts that contain both true and false claims, and also claims that some fact checkers may rate as "unproven". As a blend of truth, untruth and uncertainty, you could make a case that a large proportion of all human communication falls into this category; we try to use it more sparingly than that, although it still accounts for a substantial portion of our ratings. We will usually apply it if a post includes multiple claims of equal prominence, some of which are accurate and some of which are inaccurate; we may apply it if the claims have insufficient evidence to support them, or if they are presented in a significantly misleading way. If a post includes multiple claims of varying accuracy, but there is an identifiable central claim of greater prominence than the others, then we may choose not to apply the mixture rating.

False

We apply the false rating in situations where we are confident the central claim or claims are categorically false or highly misleading. We may apply it in situations where we are confident there is no evidence to support the claim; while on a strict interpretation it's not possible definitively to say that such a claim is false, a false rating

may sometimes be justified if the claim is asserting knowledge where no such knowledge is possible or where there is no reasonable basis for the claim.

This is particularly true in the case of claims that relate to, for example, specific events or historical information.

Satire

We have used this rating for both articles that are clearly intended to be satirical, but which have been misunderstood by readers, as well as more broadly for humorous content (see above for a discussion of why). Applying this rating does not affect the distribution of a post, which is why we use it in this broad manner — we don't think the distribution of a post should be affected simply because some people missed the joke.

We appreciate that "satire" is not a good descriptor of this broad a category of posts, and as such (see above) one of our recommendations to Facebook is that they introduce a new rating to cover humour more broadly.

Opinion

This rating is obviously intended to encompass (for example) political opinion, such as newspaper columns. We have also used it in a different sense, as an alternative to the "Mixture" rating in cases where the truth of a claim is ambiguous or has insufficient evidence (such that we could not rate it "True"), but where we nonetheless feel that it was based on a defensible set of assumptions and thus should not have its distribution affected. In other words, we may use 'Opinion' where it is possible to have different reasonable interpretations of the same evidence and the claim we are fact checking is clearly one of those interpretations.

Ratings we have not yet used

False Headline, Not eligible, Prank generator, and Not rated. We will update these guidelines as and when we use them.

Major Incident procedure

One of the areas where we believe the Third Party Fact Checking programme can play a useful role is in responding quickly to emergency situations where rumours and inaccurate information may be spreading online, for example after terrorist attacks or during natural disasters. In these situations the risk of harm from misleading information can be very high.

Major incident goal

To act quickly to reduce harm.

The focus on harm is critical: misunderstandings and inaccurate early reports are a constant feature of breaking news situations. We will not seek to resolve every misunderstanding or example of inaccurate information, but instead to prioritise what could be harmful.

Examples of potentially harmful content might include –

- Inaccurate health or safety advice
- False information about who has been affected
- False claims about what official sources have said

Triggering a major incident

Major incidents will often appear as breaking news and can be spotted by any member of staff (whether or not a fact checker) or flagged to us by Facebook or another outside source such as the emergency services. Major incidents may well occur outside working hours so a member of staff who believes they have spotted one should alert colleagues promptly through all internal channels.

Speed is essential and, if necessary, any member of the editorial team can declare a major incident. Usually to ensure coordination we would expect the decision to be made formally by the Editor, or the Chief Executive, or else the most senior member of the editorial team available.

When a major incident occurs we should –

- Tell all staff
- Ensure enough editorial staff (a minimum of two) are online for us to publish in line with our processes

- Tell Facebook through our main contact
- Consider notifying any relevant emergency service through their communication team

Active monitoring

During a major incident, Full Fact will not wait for potentially false or misleading information to appear in the Third Party Fact Checking programme queue.

We will actively monitor online sources and respond to what we believe is having an impact. The exact nature of monitoring will depend on the situation but is likely to include monitoring trending and fast-emerging posts.

Prioritising official sources of information

We recognise that during a major incident official bodies such as the emergency services will often be the most reliable sources of information.

Usually it is Full Fact's role to scrutinise, be sceptical of, and fact check the work of any public body.

During a major incident, we will use our judgement based on the context and nature of the incident, but will generally start with the presumption that official statements from the emergency services or other public bodies are the best source of reliable information that can minimise harm to the public. This approach would change if there was, for example, verifiable primary evidence that contradicted official claims.

Reviewing

In normal circumstances, Full Fact's work always involves two or three fact checkers: one (the reviewer) independently checking the work done by the first, with a third often performing a final check before publication.

During a major incident, we will adopt a triage approach. Some fact checks may need extra care, while others (such as flagging demonstrably fake images) may need to be published rapidly in line with the major incident goal to act quickly to reduce harm. We currently do not envisage a situation in which a single fact checker would publish without any extra review, but we would likely drop the third review and speed up the second review.

Action over explanation

During a major incident, it may not be possible to publish detailed fact checks at the speed necessary to reduce harm.

It is, however, important to maintain transparency. At minimum we will publish a single post with a list of actions taken and broad explanations such as 'manipulated images'.

Liaising with others

Any actions taken by Full Fact must always be taken independently and within our charitable remit and operating guidelines.

During a major incident, Full Fact's charitable goal of informed public discussion is shared by many other organisations, including the emergency services. We understand that situations can become operationally more difficult due to inaccurate information circulating.

We are therefore open to liaising with the emergency services or other relevant bodies to ensure that we can rapidly obtain reliable information from them, both about what is happening and about any specific concerns about harms from inaccurate information.

Actions taken by Full Fact based on this information will remain entirely Full Fact's responsibility and independent decision.

Full Fact
2 Carlton Gardens
London
SW1Y 5AA

team@fullfact.org

+44 (0)20 3397 5140

@FullFact

fullfact.org

Published by Full Fact, July 2019

Registered Charity number 1158683

Full Fact