Organisation for Economic Co-operation and Development

DSTI/DPC/DEMA(2024)4/FINAL

Unclassified

English - Or. English

28 June 2024

DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INNOVATION
DIGITAL POLICY COMMITTEE

Working Party on Digital Economics, Measurement, and Analysis

THE OECD TRUTH QUEST SURVEY: METHODOLOGY AND FINDINGS

JT03546919

Foreword

False and misleading content online poses significant risks to the well-being of people and society, but a lack of cross-country comparable evidence persists. This paper contributes to the statistical literature in this area by presenting the OECD Truth Quest Survey methodology and key findings. The cross-country comparable data from the survey will help policy makers better understand the mechanisms underlying the diffusion of false and misleading content online with a view to designing media literacy strategies, programmes and related policies to address the negative effects of such content.

This paper was written by Molly Lesher, Hanna Pawelec and Mercedes Fogarassy, under the direction of Audrey Plonk. The survey design benefitted from feedback from Achim Edelmann (SciencesPo médialab), Kinga Makovi (New York University Abu Dhabi), and Christian Mueller (University of Bern). Research by Worthy Cho and feedback from Camilo Umana-Dajud in the initial phase of the project are gratefully acknowledged.

The paper was approved and declassified by the OECD Digital Policy Committee on 4 April 2024 and prepared for publication by the OECD Secretariat.

Note to Delegations:

This document is also available on iLibrary as OECD (2024), "The OECD Truth Quest Survey: Methodology and findings", OECD Digital Economy Papers, No. 369, OECD Publishing, Paris, https://doi.org/10.1787/92a94c0f-en*.*

Table of contents

Tables

Figures

Boxes

Overview and key findings

The OECD Truth Quest Survey (Truth Quest) measures the ability of people to identify false and misleading content online across 21 countries. Truth Quest is a gamified, web-based survey in which respondents interact with both true and false content on an interface that resembles a "real life" social media site. Affordances such as avatars and scores encourage engagement. In total, 40 765 people completed Truth Quest across five continents.

Main findings

Perceived ability to identify false and misleading content online is uncorrelated with measured ability

- The overall Truth Quest score is 60%, indicating that respondents were able to correctly identify true and false content 60% of the time. True claims were on average more difficult to detect (56%) than false and misleading content (61%).
- In nearly all countries, respondents indicated that they are very or somewhat confident in their ability to identify false and misleading content online. Confidence tends to increase with education and income level and decrease with age. Across all countries, men are more confident than women.
- Importantly, respondents who do not feel confident in their ability to recognise false and misleading content online were as good at identifying such content as those who were confident. This suggests that confidence is not associated with ability to identify false and misleading content online, calling into question the use of perception surveys in this area.

The type of false and misleading content drives some differences within and across countries, but theme does not play a significant role

- Across all countries surveyed, satire (71%) is the easiest type of content to identify as false followed by disinformation (64%) and propaganda (58%). Misinformation (54%), contextual deception (56%) and true content (56%) were more difficult for respondents to correctly identify.
- Across the three main themes studied (the environment, health and international affairs), no major differences were observed in people's ability to correctly identify true or false and misleading content. Within each theme, the country scores ranged from 55% and 68% for content related to health and the environment and from 54% to 63% for international affairs, which has the lowest variation across countries.

AI-generated content is easier to identify than human-generated content

- On average, claims generated by artificial intelligence (AI) were correctly recognised as true or false 68% of the time. As with the scores for human-generated content, true claims were more difficult to correctly identify (63%) than disinformation (74%). Across all surveyed countries, AI-generated true claims were on average 7 percentage points (pp) easier to correctly identify as true compared to human-generated claims.

- AI-generated disinformation was 10 pp easier to correctly identify as false compared to human-generated disinformation. The gap between Truth Quest scores for AI- and human-generated disinformation ranges from 4 pp in Ireland to 18 pp in France.

Perceptions about AI affect people's ability to identify the veracity of content online

- People who agree AI will have a positive impact on their life are more likely to correctly identify the veracity of content with an AI label compared to people with a negative opinion of AI.
- On average, 59% of respondents with a positive perception of AI correctly identified a true claim with an AI label compared to 52% of those who do not agree that the impact of AI on their life will be positive. The gap between the two groups jumps to 12 pp when only respondents who strongly agree and strongly disagree are considered.
- There is no difference in the overall Truth Quest scores between respondents with a positive and negative perception of the impact of AI on their life. This suggests that it is not ability to identify false and misleading content online that is driving the difference between these two groups' scores for claims with an AI label, but rather the presence of the label.

Social media is a popular news source, and those with relatively lower Truth Quest scores trust social media the most

- Countries with the highest shares of respondents that source their news from social media have lower overall Truth Quest scores. Inversely, countries with the highest Truth Quest scores also have the lowest shares of people sourcing their news from social media.
- While people often source their news from social media, it is also the least trusted source of news with 57% of people on average not trusting it too much or at all as a reliable source of news. In contrast, only 9% trust news on social media a lot.
- On average across countries, those who trust social media a lot had a relatively lower Truth Quest score (54%) compared to those who trust news on social media some (59%) and not much or not at all (62%).

1 Introduction

False and misleading content online poses significant risks to the well-being of people and society. While such content is not necessarily illegal, it is harmful because it contributes to political polarisation, reduces trust in democratic institutions, and negatively impacts fundamental human rights (Lesher, Pawelec and Desai, 2022[1]). The Internet has become the primary conduit for spreading false and misleading content quickly and at scale.

The dissemination of false and misleading content online has played a prominent role in many recent crises (e.g. presidential elections around the world, the COVID-19 pandemic and the Russian Federation's (hereafter "Russia") war of aggression in Ukraine), harming people and societies in ways that are not yet entirely understood (Wall Street Journal, 2023[2]). Disentangling the mechanisms that underpin the circulation of such information and its impacts is challenging, in large part because of a lack of robust and cross-country comparable evidence. Expanding the evidence base in this area is fundamental to designing effective public policies to tackle this pernicious and widespread problem.

The circulation of false and misleading content online leads to harmful consequences in part because it can be hard to distinguish facts from falsehoods. Research on deception and deception detection shows that people find it very hard to recognise duplicity in others (Bond and DePaulo, 2006[3]; Luke, 2019[4]). Some research suggests that the average person discerns lies from facts at an accuracy of just over 50%, only slightly better than the rate one would achieve by chance (Bond and DePaulo, 2006[3]). More specifically, this research has found that on average, people identify truths as non-deceptive at a rate above 50%; however, they identify falsehoods as deceptive at an accuracy rate below 50%.

While it is difficult to measure the circulation of false and misleading content online, the quantitative literature in this area is expanding. Researchers have measured different dimensions of this phenomenon in certain countries and contexts, including but not limited to the prevalence of false and misleading information on online platforms (Guess, Nagler and Tucker, 2019[5]), the rate at which false information circulates online (Allcott and Gentzkow, 2017[6]), the consumption of false and misleading content (Fletcher et al., 2018[7]); (Allen et al., 2020[8]), the effects of false information on behaviour or perceptions (Southwell et al., 2022[9]), and susceptibility to false and misleading information (Pennycook and Rand, 2019[10]).

Despite these and other efforts to understand false and misleading content online and its effects, a lack of cross-country comparable evidence persists. This represents an important gap in the evidence base needed to understand and address the negative impacts of false and misleading content on people and society. This is the case in part because relevant data is dispersed across different platforms, maintained in different formats, and controlled by private companies (Watts, Rothschild and Mobius, 2021[11]), and also because gathering data for a large number of countries requires a co-ordinated approach among multiple stakeholders based on a consistent methodology.

This paper discusses the OECD Truth Quest Survey which measures the ability of people to identify false and misleading content online in a real-life setting across 21 countries. Truth Quest contributes to the statistical literature on measuring false and misleading content by providing cross-country comparable evidence on media literacy skills by theme, type and origin (generated by humans or AI). It assesses the effect of AI labels on people's performance and offers insights into where people get their news, as well as people's perceptions about their media literacy skills, among other issues.

2 Approaches to measuring false and misleading content online

False and misleading content is an amorphous concept without clearly established definitions or boundaries, making comparing across studies difficult. Coupled with the fact that people are generally not very good at detecting false and misleading content online, measurement becomes especially challenging. Nonetheless, efforts to measure various aspects of false and misleading content online have increased in tandem with concerns about its prevalence and negative impacts. This section surveys the statistical literature in this area with a view to putting Truth Quest and its findings into context. It identifies the most common phenomena measured and key modalities for administering surveys about false and misleading content online.

Measurement of false and misleading content online focuses on four main phenomena

Attempts to measure various aspects of false and misleading content online primarily cover four areas: the content of the material and its circulation, the origin of the content (human- versus AI-generated), the behaviour of people in relation to it, and people's perceptions about various aspects of such content. Measurement efforts take various forms, and most centre around topical issues such as the COVID-19 pandemic and vaccines, climate change and emotionally driven topics often related to identity, such as race, religion and political affiliation.

The content of false and misleading content online and its circulation

Research to date has often concentrated on information content itself, often through claim-based surveys. In such studies, researchers compile databases of headlines, news articles, claims or statements to explore a broad array of research questions. This approach has been used for example to measure exposure to or engagement with false and misleading content, to evaluate the extent to which people can identify false claims, and to explore circulation dynamics. Many claim-based surveys use web-based survey methods to measure susceptibility to false and misleading content. In such studies, survey respondents are presented with a series of headlines, news articles or statements and are prompted to indicate if the claims presented are true or false, real or fabricated, or they are asked to rate the veracity, reliability, or manipulativeness of the claims (Roozenbeek et al., 2022[12]).

Participants may also be prompted to indicate if they have previously seen, shared or otherwise circulated certain headlines or articles (Allcott and Gentzkow, 2017[6]). Researchers often source the claims included in their surveys from real headlines or articles that are available online and that have been fact-checked by independent fact-checkers or third parties, although AI tools have also been used. Studies relying on this methodology have explored topics such as susceptibility to COVID-19-related false and misleading content in five countries (Roozenbeek et al., 2020[13]), differences in how such content spread in different geographies (Zeng and Chan, 2021[14]), and the psychological attributes of individuals who are susceptible to "fake news" (Pennycook and Rand, 2020[15]).

The origin of the content (human- or AI-generated)

Recent advances in AI, and in particular generative AI,[1] have raised questions in numerous contexts, and the false and misleading content space is no different. Concerns about the ability of AI, and in particular large language models, to produce false and misleading content at scale are particularly widespread. Recent research suggests that such concerns may not be misplaced. Indeed, research estimates a 55% increase in synthetic content on mainstream websites (and a 457% increase on websites labelled by researchers as generally spreading false, misleading, or unreliable information) between January 2022 and May 2023 (Hanley and Durumeric, 2023[16]).

Other research by Kreps, McCain and Brundage (2022[17]) find that generative AI is capable of producing content that individuals find to be as credible, if not more credible, than human-written content. This research also suggests that people struggle or may be unable to distinguish between AI- and human-generated content (Kreps, McCain and Brundage, 2022[17]; Solaiman et al., 2019[18]; Goldstein et al., 2023[19]). Zhang and Gosline (2023[20]) find a positive bias toward human-generated content, though not necessarily an aversion to AI, when people know the content's source. However, when the source of content is not disclosed, study participants found AI-generated content more compelling than human-generated content. Additional research suggests that humans are less trusting of content when it is disclosed that it was produced by AI, and that people perceive AI-generated news headlines as less accurate compared to human-generated ones (Longoni et al., 2022[21]).

The behaviour of people in relation to false and misleading content online

Studies have also sought to understand the behaviour of people as they interact with false and misleading content online, and in particular in the context of social media. Research in this area includes empirical studies that rely on datasets that track how individuals behave online, including on social media. Such data include but are not limited to web browsing histories, social media posts and social media interactions.

For instance, in a 2019 study, researchers used microdata on sharing activity on Facebook to determine how often false news stories were disseminated on Facebook and the characteristics of individuals most likely to share such content during the 2016 presidential election in the United States (Guess, Nagler and Tucker, 2019[5]). In another study, researchers used the application programming interface (API) from Twitter (now called X) to create a dataset of tweets associated with newsworthy events that they then analysed to determine how Twitter users engaged with rumours that were later proven to be true or false (Zubiaga et al., 2016[22]).

Data on behaviour online, including on social media, can also be used to perform social network analyses. For example, in a 2020 study researchers conducted a social network and content analysis of Twitter data to study the spread of a COVID-19 conspiracy theory and strategies for mitigating the spread of such false information (Ahmed et al., 2020[23]).

The perceptions of people in relation to false and misleading content online

Perception surveys are used to collect or consolidate data on how target populations view, perceive or experience different dimensions of false and misleading content online. These surveys often aim to gather representative samples in one or more geographic areas Perception surveys allow researchers to gather data on a wide range of topics, including but not limited to perceived exposure to false and misleading content in the news or on social media, concerns about the spread and impacts of false and misleading content online, perceived ability to identify false information correctly, and general attitudes and concerns toward such content.

Perceptions are the most common aspect of false and misleading content online that are measured, in part because it is easiest to do so. Examples include the survey by United Nations Educational, Scientific

and Cultural Organization (UNESCO) on people's perceptions about the impact of hate speech and false content online related to politics (UNESCO, 2023[24]); Eurostat's questions in its survey on Information and Communication Technology Usage in Households and by Individuals on perceptions of information integrity, the propensity to fact-check, and personal encounters with false and misleading content online (Eurostat, 2021[25]); and the Reuters Institute which surveys respondents about their concerns about and encounters with false and misleading content (Newman et al., 2023[26]). However, it should be stressed that perception surveys by definition measure people's opinions, and given that research indicates the prevalence of false and misleading content is not correlated with levels of concern about it (Knuutila, Neudert and Howard, 2022[27]), perception surveys do not shed light on other aspects of false and misleading content online, including the rate at which such content circulates.

Modalities for administering surveys are varied, but two are key for measuring false and misleading content online

Surveys are administered to a target population, either in their entirety or through experiments that assess how different interventions on a control and target population influence behavioural or psychological outcomes. There are a range of ways in which surveys are administered: over the phone, in person, by email, and via text messages and apps, among others. The two modalities that are particularly relevant for the study of false and misleading content are gamification and web-based surveys.

Games help increase engagement, but researchers must be cognisant of potential biases

Gamified surveys[2] allow researchers to collect diverse data in the non-traditional context of a game. In such studies, survey questions may be embedded within an Internet-based game or game design elements such as avatars, stories and points may be incorporated into the survey design. It is a method that has been deployed in numerous domains including business, computer science, education, health, marketing and social networking, among others (Seaborn and Fels, 2015[28]). Motivational affordances incorporated in games, including rewards, challenges and leaderboards, can produce positive psychological and behavioural outcomes (Hamari, Koivisto and Sarsa, 2014[29]; Harms et al., 2015[30]). Gamification is often used to enhance respondent engagement and improve the quality and accuracy of survey results.

Several studies have assessed the extent to which gamification introduces or remediates biases, measurement errors or variance that may impact the validity or accuracy of survey data (Keusch and Zhang, 2016[31]). In the context of research focused on false and misleading content, games have been developed as an educational tool to improve people's ability to identify false and misleading content and to reduce their susceptibility to it (Basol et al., 2021[32]; Basol, Roozenbeek and van der Linden, 2020[33]). For example, free games online such as Go Viral![3] and Bad News[4] test the extent to which pre-emptive debunking (or "prebunking")[5] can improve an individual's ability to identify such content, reduce their susceptibility to it and limit its spread.

Web-based surveys reach a wide and diverse population, but are only representative of those connected to the Internet

Web-based surveys have been used as an alternative to traditional forms of data collection in a range of areas (Keusch and Zhang, 2016[31]). Among other benefits, such surveys allow researchers to obtain relatively large sample sizes and facilitate reach across jurisdictions. They can also be administered in diverse formats, tailored to different demographic groups and completed with relative ease (Evans and Mathur, 2005[34]). Depending on the technology, web-based surveys enable greater anonymity for respondents, and thus enhanced privacy, which is especially important for surveys on sensitive topics.

Greater anonymity may also lead to a decrease in social-desirability bias which can enhance data quality, as respondents are more likely to answer completely truthfully when their identity is protected and not connected to their responses (Heerwegh, 2009[35]).

Representativeness is often cited as a weakness of web-based surveys, as only people with basic digital literacy skills and access to the Internet can participate. While Internet uptake across OECD countries is relatively high on average (92% in 2023) (OECD, 2024[36]), it is lower in some geographic areas and for some groups (i.e., people who are older, living in rural areas and the less educated). However, for topics such as false and misleading content online, the target population is those who already use and have access to the Internet, thus making web-based surveys a logical and practical approach.

3 The Truth Quest methodology

Truth Quest combines aspects of all four of the most commonly measured phenomena. It seeks to understand whether some types of content are more easily distinguishable as false and misleading than others and whether the theme of the content plays any role in its detection. It aims to understand if the origin of content (human- versus AI-generated) influences people's ability to tell facts from falsehoods. It gathers information on people's behaviour as they interact with false and misleading content and their perceptions about their ability to recognise it. It likewise uses both modalities for administering surveys highlighted above – gamification and a web-based survey experience. This section describes in detail the methodology used to design and administer Truth Quest.

Target population, geographical coverage, representativeness and survey modalities

Truth Quest measures the ability of people[6] to identify false and misleading content online in a real-life setting across 21 countries (Table 1). It was designed by the OECD and administered by an external polling company to ensure a representative sample in each country. Truth Quest was translated into the primary languages of each country and the languages were localised by country. The survey was administered in January and February 2024.

Table 1. Truth Quest country coverage and languages

Country	Language(s)
Australia	English
Belgium	French, Flemish
Brazil	Portuguese
Canada	English, French
Colombia	Spanish
Finland	Finnish
France	French
Germany	German
Ireland	English
Italy	Italian
Japan	Japanese
Luxembourg	French, German, Luxembourgish
Mexico	Spanish
Netherlands	Dutch
Norway	Norwegian
Poland	Polish
Portugal	Portuguese
Spain	Spanish
Switzerland	French, German, Italian
United Kingdom	English
United States	English

Source: OECD Truth Quest Survey, 2024.

Representativeness is an essential part of any well-designed survey. Truth Quest was administered to approximately 2 000 people in each country that are representative of the population based on demographic variables including age, gender, sub-national region, educational attainment, and income level using country-specific quotas (see Table A B.1). Quotas were calculated based on the data from national statistical offices and related institutes. To ensure comparable data across countries, the quotas on demographics were as similar as possible. Using the quotas, respondents were recruited from opt-in panels and, when necessary, partner panels, undergoing thorough verification process (for example, IP addresses were cross-referenced with declared country of origin). The respondents who completed the survey in less than 3 minutes (so-called speeders) were excluded. Post-stratification weights were calculated to ensure nationally representative samples. In total, 40 765 people completed Truth Quest across five continents.

A gamified survey is particularly well-suited to measure people's ability to identify false and misleading content online because it mimics real conditions in which individuals interact with such information. In Truth Quest, participants interact with both true and false content on an interface that resembles a "real life" social media site. Affordances such as avatars and scores aim to encourage engagement.

While probabilistic samples are generally preferred to web-based panels because the latter only cover the share of the population that uses the Internet (Hargittai, 2002[37]), this shortcoming of web-based surveys does not apply to Truth Quest. Indeed, the aim of the survey is to measure people's ability to identify false and misleading content online. As a result, a web-based approach helps ensure that the survey reaches its target population.

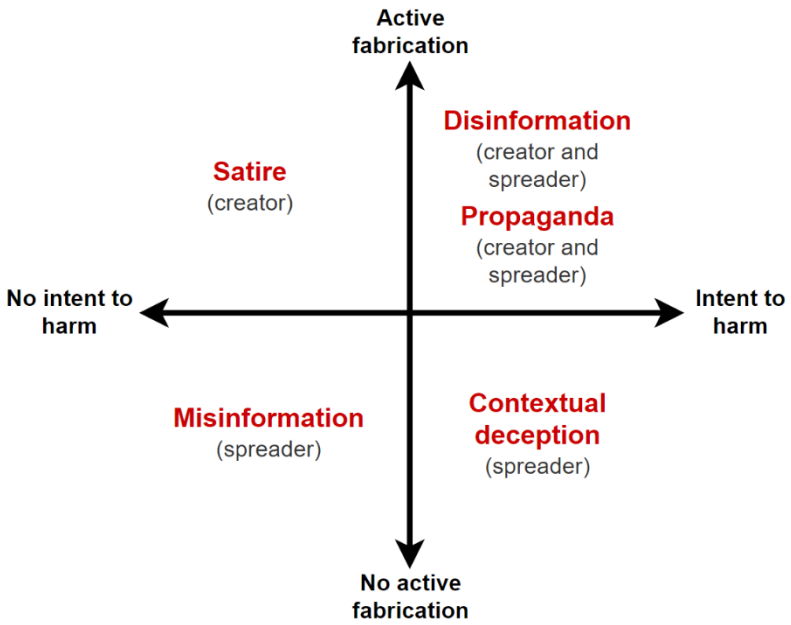
Research questions

The range of possible questions about the scale and scope of false and misleading content online is vast, including the quantity of false and misleading content online, if and in which circumstances such content is shared (and with whom), and whether it is broadly distributed or concentrated among trending topics in the news cycle, among many others. While all such questions are valid from a research perspective, it is important to clearly define the phenomena to be measured. The primary aim of Truth Quest is to gather cross-country comparable evidence on media literacy skills by type, theme and origin of the content (human- versus AI-generated). Other aspects are also of interest, including the effect of AI labels on people's performance, media consumption habits, if people trust information from various news sources, and their views about the importance of several topics related to information quality, AI and privacy.

Can people distinguish some types of false and misleading content better than others?

One of the primary questions the survey aims to answer is whether people are able to distinguish some types of false and misleading content better than others. This line of research sheds light on the factors that influence susceptibility to false and misleading content online. False, inaccurate and misleading information often assumes different forms based on the context, source, intent and purpose. The claims included in Truth Quest are categorised according to the OECD taxonomy of false and misleading content online (Figure 1).

Figure 1. OECD taxonomy of false and misleading content online



Note: Fabricated content refers to information that is manufactured or doctored. This includes modified photos or videos, manipulated text, or statements that lack a factual basis. Intent to harm refers to the act of knowingly or purposely intending to deceive, manipulate or inflict harm on a person, social group, organisation or country.
Source: OECD (Lesher, Pawelec and Desai, 2022[1])

The taxonomy distinguishes between five different types of false and misleading content online: disinformation, misinformation, contextual deception, propaganda and satire (see Annex A). The definitions in the taxonomy can be broadly adopted to characterise the different types of false and misleading content that circulates online. However, in practice, it can be challenging to classify some types of claims,

particularly those related to propaganda and disinformation. To address this challenge, a decision tree framework was used to accurately and consistently categorise the different types of false and misleading content in the Truth Quest database (see Annex A).

Does the theme of false and misleading content influence people's ability to discern fact from fiction?

Another seminal question in Truth Quest is whether the theme or topic of content influences people's ability to discern its veracity. Truth Quest includes a balanced set of three main themes: the environment, health and international affairs. The three themes were chosen because there is a significant amount of true and false and misleading content on these themes. Moreover, people across countries and cultures can identify with and understand them. This is an important criterion because of the cross-country nature of Truth Quest and the need to tap into a large number of fact-checked claims to construct the database.

While the survey does not shed light on why the ability to identify false and misleading content may vary by theme, knowing if there are differences can help identify focus areas for media literacy policies and programmes. Health-related information, for example, affects people directly and on a personal level, while international affairs is more removed from people's daily lives. This difference between themes that are highly personal and others that are more abstract and distant can shed light on the mechanisms that underlie the detection of false and misleading content online.

The environment is another topic that is often the subject of the creators and disseminators of false and misleading content online. As with vaccine hesitancy, climate change denial and doubts are related to a growing phenomenon of science scepticism (Rutjens and van der Lee, 2020[38]). Strong and polarised beliefs influence the perception of false and misleading content, but exposure to fact-checked information has the potential to counter these views (Hameleers and Van der Meer, 2020[39]).

Does the origin of content (human- versus AI-generated) affect people's ability to detect false and misleading content?

Although there are many beneficial uses for generative AI, it has also raised concerns about its ability to exacerbate the generation and dissemination of false and misleading information. Truth Quest tests people's ability to identify false and misleading content generated using natural language processing[7] and compares it to their ability to identify false and misleading content generated by humans for content of the same type and theme. The purpose of this line of research is to test the hypothesis that there is a difference in people's ability to identify false and misleading content based on the origin of the content (human- versus AI-generated).

The research into origin can also help overcome some of the challenges associated with identifying the intent to deceive and the role of the actor as a creator or spreader of content that are identified in the OECD taxonomy. As such, respondents were given either a true claim (fact-checked after it was generated by GPT-4), or a claim classified as disinformation under the OECD taxonomy. In total, two AI-disinformation and two AI-true claims were generated for each of the environment, health and international affairs themes.[8] Respondents were then randomly shown three AI-generated claims in each of the three themes. These three claims were either a true claim or disinformation defined by the OECD taxonomy. A matrix of all claims can be found in Table A B.2.

Do AI labels influence people's identification of content as true or false?

AI labelling – sometimes referred to as "watermarking" – is a mechanism that has been promoted in some policy circles as a way to mitigate the risks of generative AI. In this paper, the term "AI label" means a visible label or warning that informs the public that the content is generated by AI.[9] AI labels can serve two

different goals. First, they can communicate the provenance of a piece of content or information regarding the process by which it was produced. Second, AI labels can be used to signal that a piece of content is true or misleading (Wittenberg et al., 2024[40]). A discussion of AI labelling can be found in Box 1.

Box 1. The psychology associated with labelling content online

Policy makers around the globe have considered or promoted AI labelling as a mechanism for combating the spread of false and misleading content online. For example, in July 2023 the United States announced that several leading AI companies had voluntarily committed to developing "robust technical mechanisms to ensure that users know when content is AI generated, such as a watermarking system" (The White House, 2023[41]). Moreover, in October 2023 the United States issued an Executive Order where it announced its intent to "help develop effective [AI] labelling and content provenance mechanisms" (The White House, 2023[42]). In addition, the People's Republic of China enacted AI regulation requiring "deep synthesis service providers" to visibly label synthetically generated content that may mislead or confuse the public (Sheehan, 2023[43]) (Latham & Watkins Privacy & Cyber Practice, 2023[44]).

Several researchers have conducted experimental studies to evaluate the efficacy of labelling interventions, the ways in which individuals perceive AI labels, and how these labels may alter behaviour. For example, Epstein et al. (2023[45]) examined how individuals understand AI-related terms and perceive content affixed with these labels. Participants in the study tended to associate terms such as "AI Generated", "Generated with an AI tool" and "AI Manipulated" with content that was created using AI, regardless of whether it was misleading. In contrast, terms such as "Deepfake" and "Manipulated" were associated with content that was misleading, regardless of whether it was generated by AI. Such findings suggest that certain labels may be more effective in mitigating the harms associated with content produced by generative AI.

However, research conducted by Wittenberg et al. (2024[40]) and Saltz, Leibowicz and Wardle (2021[46]) suggests that labelling interventions may be an imperfect mechanism for curbing the harms of false and misleading content online. First, it can be challenging to identify AI-generated content post hoc and labelling interventions may rely on voluntary disclosures or imperfect and/or inadequate computational approaches for identifying AI-generated content. Second, labelling interventions may not always elicit the desired reaction from the public. For example, one study found that labelling interventions evoke different reactions from individuals. While some participants in this study felt positively about labelling interventions, others found them to be punitive, patronising, paternalistic, inappropriately applied or a form of censorship (Saltz, Leibowicz and Wardle, 2021[46]). Third, there could be negative indirect effects associated with this form of intervention. For example, research on the impact of fact-check warnings on false or misleading news headlines has found that there is an "implied truth effect" associated with certain false and misleading content labelling interventions (Pennycook et al., 2020[47]).

Truth Quest introduces an experiment to test the impact of AI labelling on people's ability to identify false and misleading by comparing whether the presence of no label, a "Text generated by artificial intelligence" label in light grey (less obvious) and a "Text generated by artificial intelligence" label in medium blue (more obvious) influences people's ability to identify false and misleading content (see an example of the grey label in Figure 3). It is important to note that not all AI-generated content is factually inaccurate or developed with the intent to mislead. Thus, AI labels may not be intended to connote or imply that content is false.

Questions related to respondents' perceptions and behaviour

At the outset of the survey, respondents were asked to self-report their level of confidence in recognising false and misleading content online (see Table A A.1 or the exact wording of this question and the response options). This question was included prior to respondents seeing the first claim to not bias their answers. This question was carefully worded to be comparable with the same question that has been asked in the Eurobarometer survey on fake news and disinformation online to enable a time series (European Commission, 2018[42]), and it is similar to questions asked on other popular perception surveys about false and misleading content online. This question was included to test the validity of people's perception of their ability to identify false and misleading content and to compare it to their actual ability as measured by Truth Quest. This line of research sheds light on the validity of perception surveys for measuring people's contact with false and misleading content.

At the end of the game, but prior to being shown their score, respondents were also asked about their media consumption habits, if they trusted information from various news sources, and their views about related issues, including the state of democracy in their country and information quality. Additional questions regarding views about AI, as well as about privacy-related behaviour and perceptions, completed the survey. The exact wording, the response options and the position of these questions can be found in Table A A.1. While interesting in and of themselves, these questions also help provide useful context when interpreting the Truth Quest results.

The content and interface of the gamified survey

At its core, Truth Quest is comprised of a database of fact-checked claims and an online interface that mimics a social media environment (desktop and mobile). Respondents interact with the claims (e.g., by "liking" and "flagging" claims) and they answer questions about the veracity of the claims. They likewise answer other questions related to their contact with and perceptions about false and misleading content online, among other issues. This section discusses the construction of the Truth Quest database and describes the interface that was designed to administer the survey.

The Truth Quest database of claims

The Truth Quest database consists of true and false claims that were extracted from databases of fact-checked content and fact-checking websites. Beginning with a set of keywords related to the three thematic areas of interest (the environment, health and international affairs), over 25 000 fact-checked claims were identified.

To reduce the number of claims to a manageable subset for initial testing, filters were implemented to remove the following types of claims:

- duplicates;
- those that could be categorised into more than one thematic area or for which the categorisation according to the OECD taxonomy was not clear cut;
- claims about a country in which the survey was administered or with whom one of the surveyed countries has a close tie;
- claims about topics that are widely known or which might be controversial (e.g., COVID-19 and Russia's war of aggression against Ukraine);
- those that were based on content that was not text-based; and
- claims for which there was no content beyond a headline.

Once the database of claims was cleaned, claims were categorised according to the OECD taxonomy of false and misleading content (see Annex A) and 170 claims were selected to generate a balanced set of claims in the three themes (the environment, health and international affairs) and the five categories in the taxonomy (disinformation, misinformation, contextual deception, propaganda and satire).[10] These 170 claims were assessed for emotional valence,[11], veracity (i.e. true or false)[12] and level of comprehension of each claim in a pre-test using Amazon Mechanical Turk (MTurk).[13] The pre-test was conducted in English in the United States in various periods throughout August, September and October 2023 with 538 respondents in total. Using the data collected from the pre-test, the subset of claims was further reduced to 54.

Using GPT-4, 12 claims were then generated in the three thematic areas (see endnote 8). These claims were either true or disinformation according to the OECD taxonomy. The 12 claims were taken as given from GPT-4 apart from cosmetic edits to ensure consistency with other claims in the database (i.e., words in all capital letters were put into lower case). This was done to simulate the mass production of false and misleading content at scale, where human intervention is unfeasible. These claims were fact-checked by the Authors. Together with the 54 claims that were analysed in the pre-test, a total of 66 claims are included in the Truth Quest database.[14]

Professional images were then chosen to accompany each of the claims. Efforts were made to choose neutral images that did not include people's faces, and only human-generated images were selected.[15] The claims were then fielded in a pilot study in the United States in which the full survey was tested to ensure optimal respondent experience. Pilot data from 200 respondents in the United States were collected in October 2023. Following the pilot, minor changes were made to improve clarity and respondent experience.

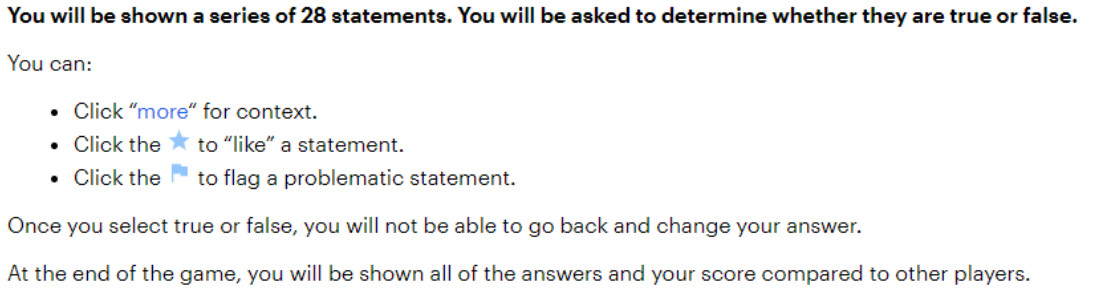
The Truth Quest "game" interface

Truth Quest begins with a series of demographic questions regarding age, gender, educational attainment, income level, country of residence and sub-national region. Other demographic questions (e.g., regarding political affiliation) were also asked, but not used directly as quotas. The demographic questions function as quotas to ensure a representative sample in each country as well as variables that allow the results to be disaggregated by demographics.

Respondents then saw an opening screen of instructions for the game (Figure 2, Panel A). Interactive functionalities like clicking "more" to see more content about a claim, the ability to "like" a claim and to "flag" it as problematic were also explained. Truth Quest captured data about how often respondents used these interactive features. Respondents were then asked to select an avatar (Figure 2, Panel B). The avatars were chosen to reinforce the game-like experience and to add a sense of fun to foster engagement. Neutral avatars were chosen (i.e., no gender, age, nationality or other such characteristics that could potentially affect a respondent's personal experience and opinion of the survey) and to fit within the overall style of the game.[16]

Figure 2. Truth Quest interface: Instructions and avatars

Panel A. Truth Quest instructions



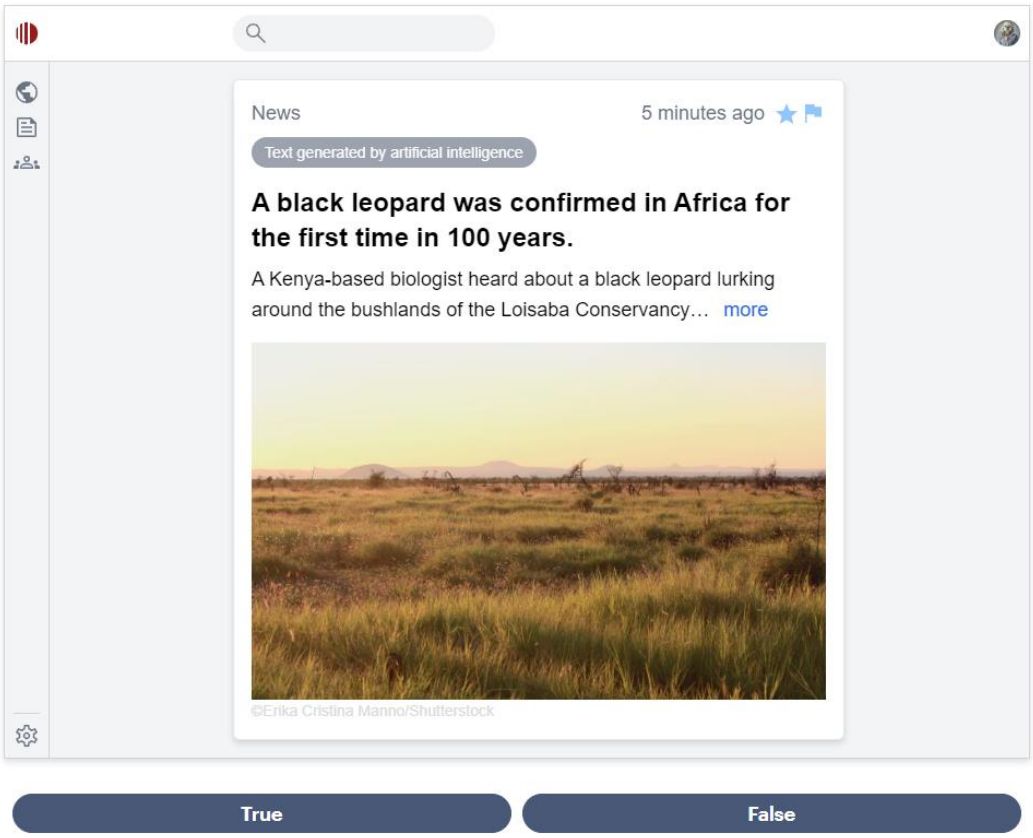
Panel A. Truth Quest avatars

Source: OECD Truth Quest Survey, 2024. Image credit: © Cranach/Shutterstock.com.

Upon selecting an avatar, respondents were asked the question "Do you think that the following statements are true or false?" after which they were shown a series of 28 claims from the Truth Quest database.[17] The claims were randomised, and each respondent saw a balanced set with respect to theme, type and origin. Each of the 28 claims were displayed to the respondent in a "frame" that mimics a real-life social media post (Figure 3). Below the frame, respondents then indicated whether they thought that the claim is true or false. The labels "News" and "5 minutes ago" were included in all frames to resemble a real-life social media experience and did not vary throughout the survey.

Figure 3. An example of the Truth Quest frame

Claim in the international affairs theme



Source: OECD Truth Quest Survey, 2024. Image credit: © Erika Cristina Manno/Shutterstock.com.

For the last claim, respondents saw a human-generated true claim in one of the three themes that wasn't shown previously. Half of the respondents saw a grey label "Text generated by artificial intelligence" (see Figure 3) and half of respondents saw the same label in a medium blue colour. The grey colour is less obvious and more in line with similar labels that have been proposed by industry (Saltz, Leibowicz and Wardle, 2021[41]). The blue label is more obvious and was added after the pilot because it was unclear if respondents noticed the grey label. Since the human-generated true claims were also shown randomly in the earlier part of the survey to other respondents, it is possible to assess the influence of the label (grey and blue) on respondent behaviour.

At the end of the game, respondents were asked about their media consumption habits, if they trusted information from various news sources, and their views about the importance of several topics related to information quality (see Table A B.3). Respondents were presented their score in comparison to the average score (computed based on the pilot) and then a table with the answers they selected alongside the correct answer (true or false) for each claim to which they were exposed. The purpose of showing the correct answers is to ensure respondents are aware of the false and misleading information they saw.

4 Truth Quest findings

The findings presented in this section are based on data from 40 765 respondents in 21 countries (see Table 1). The exact wording for the behavioural and perception questions is included in Table A B.3, as is the position these questions were asked in the actual survey. The overall findings are presented first, including a comparison between people's perceptions about their ability to identify false and misleading content online and their actual ability as measured by Truth Quest. This is followed by findings related to whether the type, theme and origin of the content (human- versus AI-generated) plays any role in its detection. Insights into the effects of AI labelling is then discussed based on evidence from the survey, followed by findings about people's media consumption behaviour and their perceptions about the trustworthiness of various media sources.

It should be noted that interpreting survey results always requires caution because an observation about an entire population is extrapolated from a sample. As a result, the associated uncertainty about observations is expressed in the margin of error. In this survey, for a country with a sample of 2 000 respondents and a significance level set to 5%, the margin of error for one question is approximately 2%. This means, for example, that if 60% of respondents answered correctly for a particular claim, the "true" result is most likely between 58% and 62% of the online adult population of a given country. This range is called a "confidence interval", and the difference between two scores is statistically significant when their confidence intervals do not overlap.

Perceived ability to identify false and misleading content online is uncorrelated with measured ability

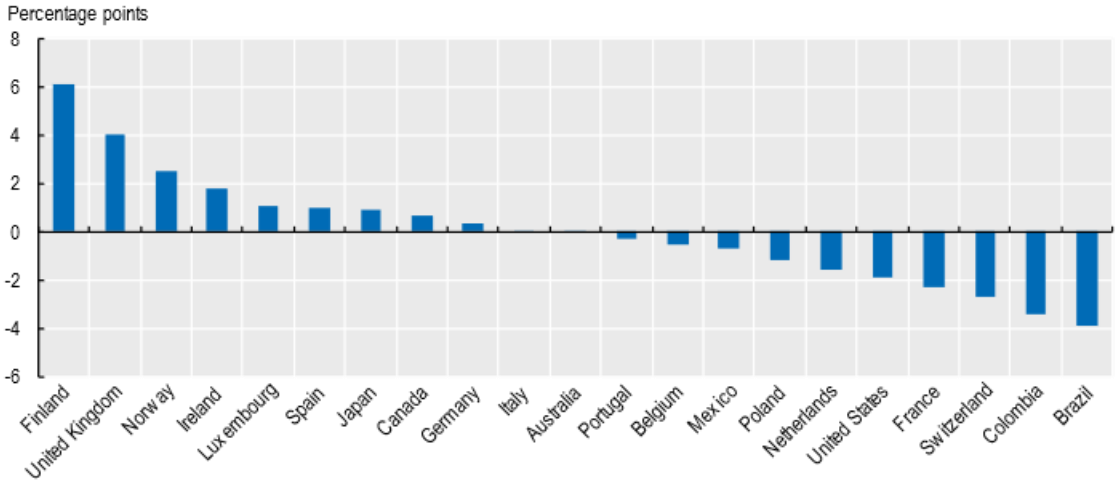
While Truth Quest aims to address a range of specific research questions, at its core the survey provides evidence on people's ability (or skills) to determine if a claim online is true or false. To calculate the overall score, the total number of correct responses is divided by the total number of claims seen. A country score is thus an average of all respondents' results. Post-stratification weights were applied to increase the representativeness of the sample collected to each country's overall population.

On average, respondents correctly identified the veracity of content 60% of the time which indicates that the survey is measuring a phenomenon that can be distinguished from chance (50%). The Truth Quest score is slightly higher than previous research by Bond and DePaulo (2006[3]) who found a 54% success rate overall. True claims were on average more difficult to detect by Truth Quest respondents, who identified true content correctly on average 56% of the time, 5 percentage points less than in the Bond and DePaulo study. The scores for true claims were lower than the scores for false claims in all countries except for Brazil, Colombia and the United States.

Country scores ranged from 66% in Finland to 54% in Brazil, a difference of 12 percentage points (Figure 4). From a regional perspective, the Nordic countries performed best (Finland and Norway) and the Latin American countries have the most opportunity to catch up, particularly Brazil and Colombia. While differences among the countries in the middle of the distribution may not be large, the differences among countries at both ends of the distribution are sizeable and understanding why they exist is important for designing effective media literacy strategies, programmes and policies.

Figure 4. Ability of adults to identify the veracity of online news

Distance from average performance, 2024



Note: The score is calculated as the total number of correct responses divided by the total number of claims seen. A country score is thus an average of all respondents' results and expressed in percentages. Distances are calculated as the difference between the average overall Truth Quest score for all countries and the overall country-specific Truth Quest score. Country scores above zero reflect above average ability and country scores below zero reflect below average ability.
Source: Authors' calculations based on the OECD Truth Quest Survey, 2024. The data underlying this figure are available at: OECD (2024[43]), "Ability of adults to identify the veracity of online news", OECD Going Digital Toolkit, https://goingdigital.oecd.org/indicator/80.

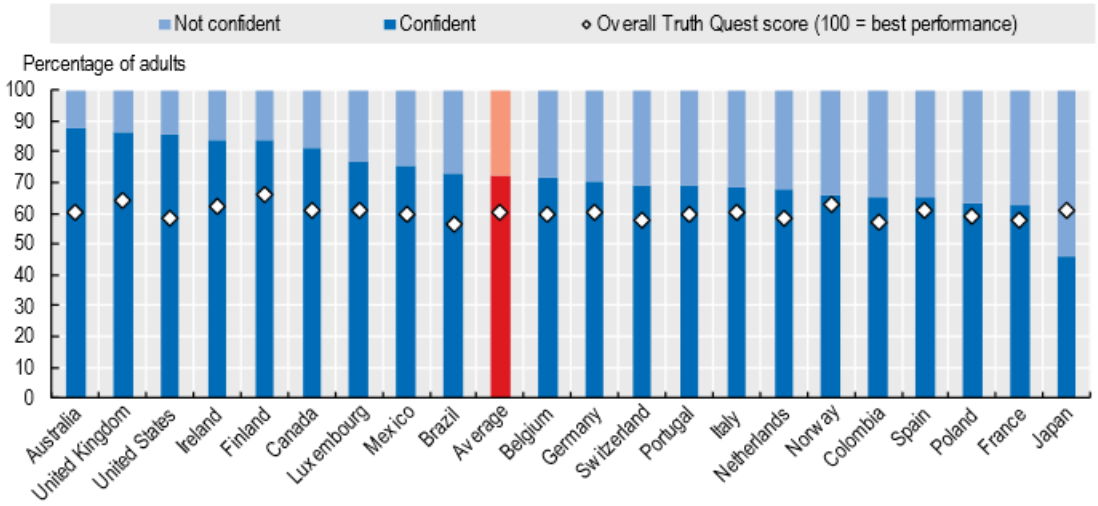
Demographics seem to be uncorrelated to the overall average Truth Quest scores, but across countries some interesting patterns emerge at the country level. For instance, despite a very marginal difference in scores based on income overall, respondents in the lowest income bracket of most countries consistently performed the worst, while respondents in the highest income bracket performed the best. In some countries, including France, Ireland, Italy, the Netherlands and the United States, the difference between the bottom 20% of income earners (by household) and the top 20% is in the range of 5-6%. On the other end of the spectrum, in Poland and Japan the difference in scores between the highest and lowest income earners is less than 1%. A similar pattern is observed for education, where respondents with the lowest level of education performed the worst and those with the highest level of education performed best.[18]

Across age and gender there is little difference in the overall scores at the country level. On average, Truth Quest scores increase slightly with age. However, among different age groups in the United States this difference is more striking. Among US respondents, those aged 18-24 and 25-34 scored the lowest of any age group of any country (53%), while US respondents over 45 scored 61% on average. This suggests that in the United States, media literacy efforts may be best targeted at younger generations.

Assessing skills is one core element of Truth Quest, but it is also important to understand differences between measured and perceived ability. The first question respondents answered in Truth Quest was about confidence in their ability to identify false and misleading content online. In nearly all countries, over half of respondents indicated that they are very or somewhat confident in their ability to identify false and misleading content online, with the notable exception of Japan, where just under half of respondents (47%) are very or somewhat confident in their ability to recognise such content (Figure 5).

Figure 5. People's perception of their ability to recognise false and misleading content online

2024



Note: The "Confident" category groups the "Very confident" and "Somewhat confident" sub-components, and the "Not confident" category groups the "Not at all confident" and "Not very confident" sub-components. The score is calculated as the total number of correct responses divided by the total number of claims seen. A country score is thus an average of all respondents' results and expressed in percentages. The average score is calculated as a simple unweighted average of the 21 country scores covered by Truth Quest. Adults are defined as people aged 18 and older.
Source: Authors' calculations based on the OECD Truth Quest Survey, 2024.

Using four sub-components,[19] the share of respondents who are on average not very confident in their ability (24%) exceeds those who are very confident (16%) and this trend holds for more than half of the countries surveyed. In Japan, respondents were 10 times more likely to feel not very confident compared to those feeling very confident in their ability.

Confidence tends to increase with education and income level. With respect to gender, across all countries men are also more confident than women (+7 pp) and in four countries the gap is above 10 pp: Japan (15 pp), Germany (14 pp), Switzerland (14 pp) and Poland (10 pp). Age differences are particularly striking. On average, the confidence gap is more than 18 pp between those aged 18-24 and those 65 and older. In other words, younger people tend to be more confident than older people in their ability to recognise false and misleading content online. However, important differences across countries also emerge. In Ireland, for example, older people are only 3 pp less confident than the youngest age cohort while in Portugal the gap is 36 pp.

Interestingly, respondents who didn't feel confident in their ability to recognise false and misleading content online were as good at identifying such content as those who were confident (Table 2). This finding holds across all 21 countries surveyed by Truth Quest. This suggests that confidence is not associated with ability to identify false and misleading content online, calling into question the use of perception surveys in this area. Furthermore, perception surveys inquiring whether people have seen false and misleading content are likely to be unreliable given that people can encounter false content without being aware of it.

Table 2. Truth Quest score by confidence level

2024

Confidence level	Overall Truth Quest Score (100 = best performance)
Confident	61%
Not confident	60%

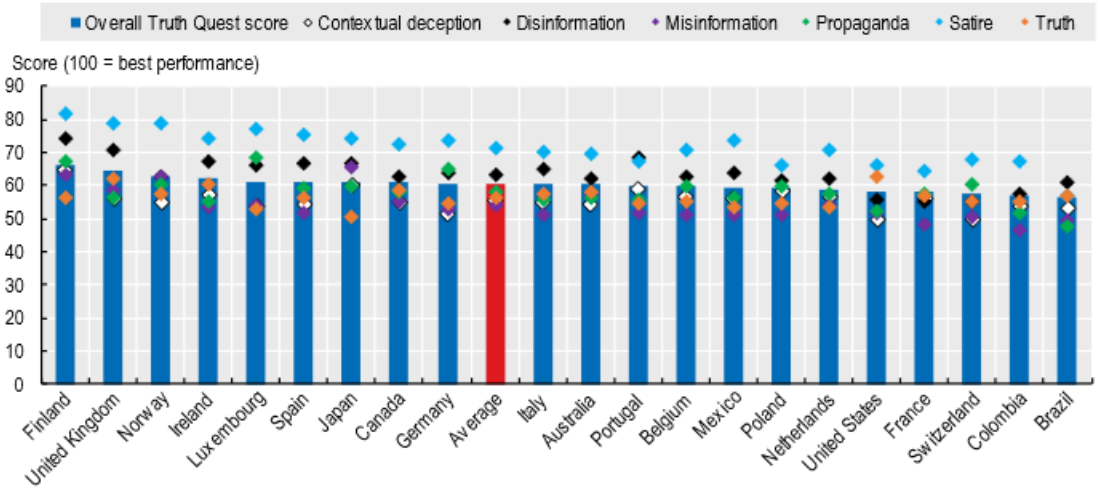
Note: The "Confident" category groups the "Very confident" and "Somewhat confident" sub-components, and the "Not confident" category groups the "Not at all confident" and "Not very confident" sub-components. The overall Truth Quest score is on a scale of 0-100 (100 = best performance). The difference between the average scores for Confident and Not confident (1%) is not statistically significant. The average is calculated as a simple unweighted average across the 21 country scores covered by Truth Quest and expressed in percentages.
Source: Authors' calculations based on the OECD Truth Quest Survey, 2024.

The type of false and misleading content drives more differences within than across countries

The OECD taxonomy distinguishes between five different types of false and misleading content online: disinformation, misinformation, contextual deception, propaganda and satire (see Annex A). Truth Quest seeks to assess whether people are able to distinguish some types of false and misleading content better than others. Across all countries surveyed, humorous content (satire) is the easiest type of content to correctly identify (Figure 6). However, it is also the type in which the largest differences across countries are observed: 82% of respondents in Finland correctly recognised satire as false compared to 57% of respondents in Brazil.

Figure 6. Average Truth Quest scores by type

2024



Note: The score is calculated as the total number of correct responses divided by the total number of claims seen. A country score is thus an average of all respondents' results and expressed in percentages. The average is calculated as a simple unweighted average across the 21 country scores covered by Truth Quest.
Source: OECD (2024[43]), "Ability of adults to identify the veracity of online news", OECD Going Digital Toolkit, based on the OECD Truth Quest Survey, 2024, https://goingdigital.oecd.org/indicator/80.

Content created and disseminated with the intent to deceive (disinformation) is also generally easier to identify as false. France, Switzerland and the United States are exceptions as the disinformation score is slightly lower compared to the overall score. Conversely, Finland, Portugal and the United Kingdom had a disinformation score 7-8 pp higher than their overall score. Large differences are also observed between countries, ranging from 55% in France and Switzerland to 74% in Finland.

In contrast, misinformation, contextual deception and true content are more difficult types of content for respondents to detect on average. Misinformation in Colombia (47%) was the lowest score among all types and across all countries. Together with France and Spain, these three countries had a misinformation score 10 pp lower than their overall score.

The scores for contextual deception were lower than the overall score across all countries and ranged between 50% in the United States and Switzerland and 65% in Finland. Japan is the only country in which respondents recognised disinformation more easily as false than the average claim (+5 pp). The scores for propaganda oscillated around the overall score of each country, and it was the most difficult content type to identify as false in Brazil, with 48% of Brazilians correctly identifying propaganda as false. Conversely, in Luxembourg respondents recognised propaganda as false more easily (+7 pp above the overall score in Luxembourg).

True content had the lowest variability across countries, ranging between 51% in Japan to 63% in the United States. The United States and Brazil are the only countries in which the scores for true content were higher than each country's respective overall score, indicating that in these two countries respondents more easily identify true content than false content on average. Conversely, the scores for identifying true content in Japan (11%) and Finland (10%) were lower than each country's respective overall score.

While on average the differences between most types are relatively small, there are notable differences within countries. In Finland, which has the largest gap between different types of content, the difference between the score for true content (56%) and satire (82%) is 26 pp. In contrast, in Brazil, which has the smallest gap between different types of content, the difference between the scores for disinformation and propaganda is only 13 pp on average.

In terms of demographics, overall there were no sizeable gender differences between the content types, although women identified contextual deception as false at a higher rate in the Netherlands and the United States, and men identified misinformation as false at a higher rate in the Netherlands and Brazil. The difference between respondents with tertiary and low or no education is most visible with disinformation and satire, with gaps between respondents in these two cohorts of 9 pp and 7 pp, respectively. Similarly, gaps between respondents in the highest and lowest income cohorts were 7 pp for disinformation and 8 pp for satire. With respect to age, on average people aged 65 and older were better at recognising satire as false (+10 pp) than those in the youngest age cohort, but they performed worse than those in the youngest age cohort with respect to identifying true content (-5 pp).

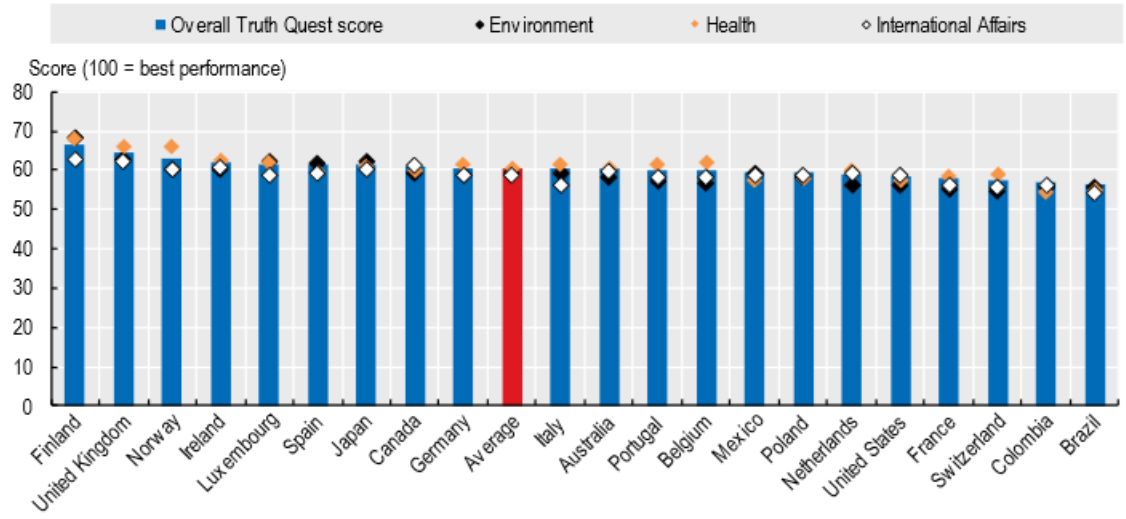
The theme of content online does not affect people's ability to identify its veracity on average

A key question in the literature is whether the theme (or topic) of content influences people's ability to discern its veracity. Truth Quest tests this hypothesis by measuring people's ability to identify false and misleading content in three main themes: the environment, health and international affairs.

Overall, there are no major differences in people's ability to correctly identify true or false and misleading content based on thematic area. For all three of the topics, respondents correctly identified the veracity of claims in line with the overall average score of 60%. This trend is generally observed across all countries (Figure 7).

Figure 7. Truth Quest scores by theme

2024



Note: The score is calculated as the total number of correct responses divided by the total number of claims seen. A country score is thus an average of all respondents' results and expressed in percentages. The average is calculated as a simple unweighted average across the 21 country scores covered by Truth Quest.
Source: OECD (2024[43]), "Ability of adults to identify the veracity of online news", OECD Going Digital Toolkit, based on the OECD Truth Quest Survey, 2024, https://goingdigital.oecd.org/indicator/80.

On average, content related to health (61%) was slightly easier to identify than content related to the environment (59%) and international affairs (59%). Within each theme, the scores ranged from 55% to 68% for content related to health and the environment, and from 54% to 63% for international affairs, which had the lowest variation across countries.

Cross country comparisons mask some significant differences within countries. In Norway, for example, the health score was 6 pp higher than the scores for the two other themes. In Belgium, an important gap is observed in the scores for content related to the environment (56%) and health (62%). Finland has the highest score for each theme, but it is noteworthy that the score for international affairs (63%) is lower than for health (68%) and the environment (68%).

Demographic differences between themes are very small. On average, there are no notable gaps between genders, income and educational background, and the average difference between age cohorts is below 3%. One exception relates to the United States. As indicated above, significant differences are observed between the youngest and the oldest age cohorts in the United States, but those differences are smaller for content related to the environment (6%), health (11%) and international affairs (12%).

AI-generated content is easier to identify than human-generated content on average

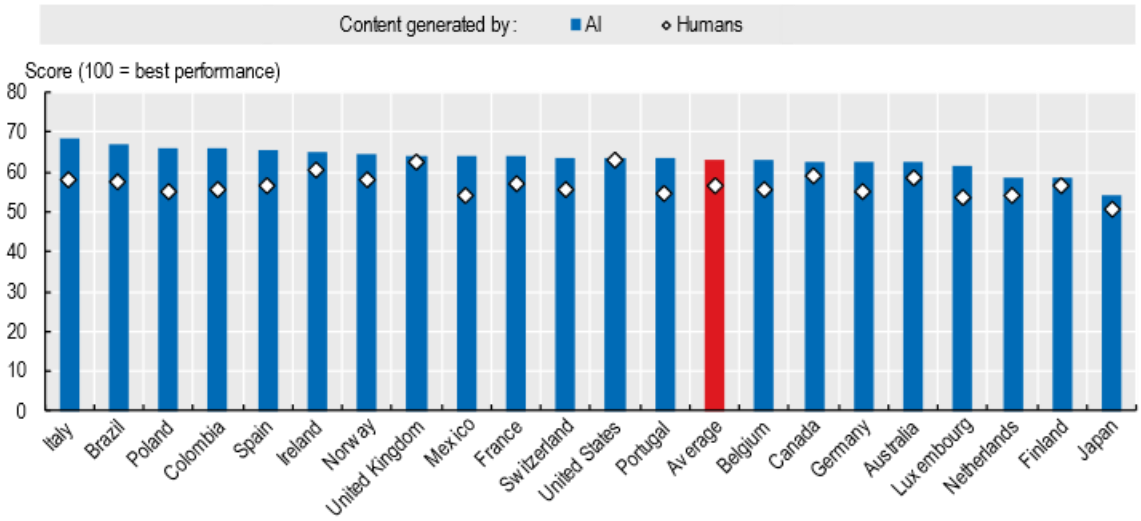
Truth Quest tests the hypothesis that there is a difference in people's ability to identify false and misleading content based on its origin (human- versus AI-generated) for content of the same type and theme. AI-generated claims were generated using GPT4 in the main themes used in the survey: health, environment and international affairs. The AI-generated claims were fact-checked by the Authors and identified as disinformation under the OECD taxonomy or as true.

On average, AI-generated claims were correctly recognised as true or false 68% of time. As with the scores for human-generated content, true claims were more difficult to correctly identify (63%) than disinformation

(74%). Across all surveyed countries, AI-generated true claims were on average 7 pp easier to correctly identify as true compared to human-generated claims (Figure 8). However, differences are evident across countries. For example, the gap is marginal in Finland, the United Kingdom and the United States, but the gap is over 10 pp in Italy, Mexico and Poland. It is noteworthy that countries with a low overall Truth Quest score such as Brazil and Colombia are among the countries with the highest score for identifying AI-generated true claims. Conversely, Finland, the country with the highest overall score, has one of the lowest scores for identifying AI-generated true claims.

Figure 8. Truth Quest scores for AI- and human-generated true claims

2024

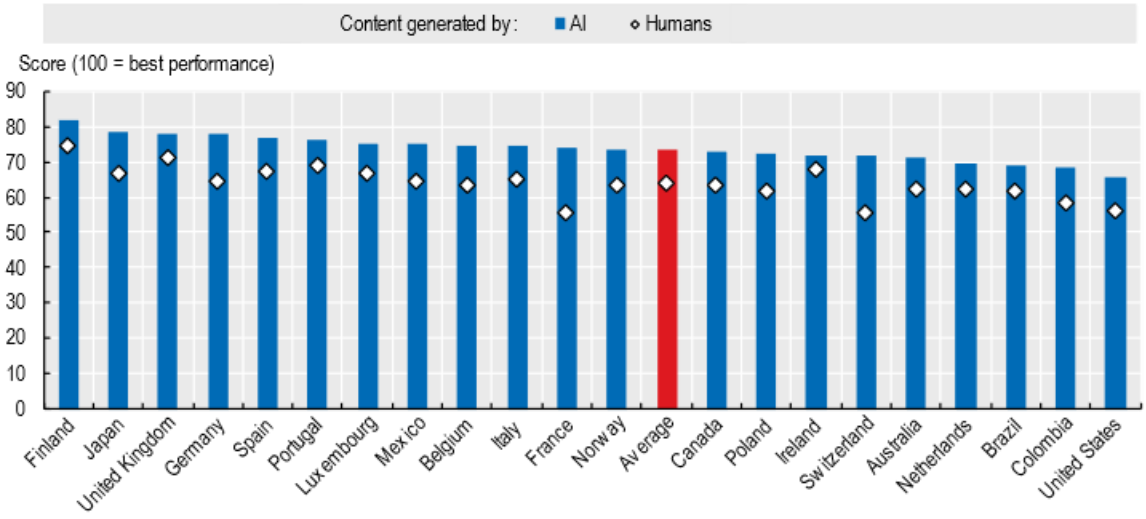


Note: The score is calculated as the total number of correct responses divided by the total number of claims seen. A country score is thus an average of all respondents' results and expressed in percentages. The average is calculated as a simple unweighted average across the 21 country scores covered by Truth Quest.
Source: OECD (2024[44]), "Ability of adults to identify online disinformation created by generative AI", OECD Going Digital Toolkit, based on the OECD Truth Quest Survey, 2024, https://goingdigital.oecd.org/indicator/81.

Although there are many beneficial uses of generative AI, it has also raised concerns about its ability to exacerbate the generation and dissemination of false and misleading information. However, on average AI-generated disinformation was 10 pp easier to correctly identify as false compared to human-generated disinformation (Figure 9). The gap between Truth Quest scores for AI- and human-generated disinformation ranges from 4 pp in Ireland to 18 pp in France.

Figure 9. Truth Quest scores for AI- and human-generated disinformation

2024

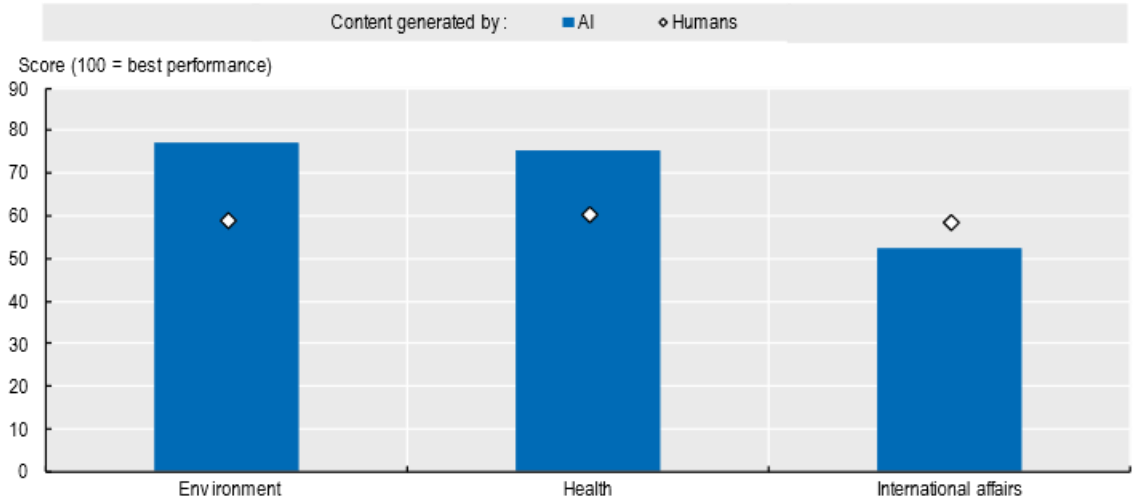


Note: The score is calculated as the total number of correct responses divided by the total number of claims seen. A country score is thus an average of all respondents' results and expressed in percentages. The average is calculated as a simple unweighted average across the 21 country scores covered by Truth Quest.
Source: OECD (2024[44]), "Ability of adults to identify online disinformation created by generative AI", OECD Going Digital Toolkit, based on the OECD Truth Quest Survey, 2024, https://goingdigital.oecd.org/indicator/81.

While on average the theme of content generated by humans does not tend to affect people's ability to identify its veracity, greater differences are observed for AI-generated content. AI-generated content about the environment (77%) and health (75%) was easier to identify as true or false than AI-generated content about international affairs (53%) (Figure 10). At the same time, the gap between Truth Quest scores for AI- and human-generated content are relatively larger for the environment (18 pp) and health (15 pp) than they are for international affairs (5 pp).

Figure 10. AI- and human-generated content by theme

2024



Note: The score is calculated as the total number of correct responses divided by the total number of claims seen. A country score is thus an average of all respondents' results and expressed in percentages. This figure shows the average, which is calculated as a simple unweighted average across the 21 country scores covered by Truth Quest.
Source: Authors' calculations based on the OECD Truth Quest Survey, 2024.

While these results may be unexpected given the concerns raised about generative AI in particular to create and spread disinformation quickly and at scale, it could be driven by the fact that machine learning models replicate common patterns, structures and themes which may lead to the easier identification of those claims as false. Moreover, ChatGPT has encoded security mechanisms that prevent users from generating false and misleading content, and while these mechanisms can be overcome, they may result in the generation of comparatively "easier" claims to identify, although more research in this area is needed.

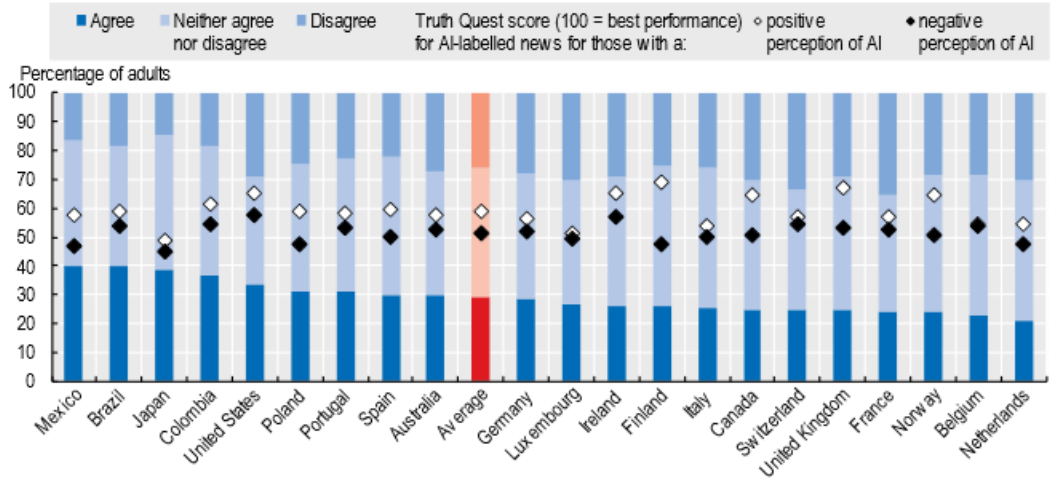
Perceptions about AI affect people's ability to identify the veracity of content online

Truth Quest introduces an experiment to test the impact of AI labelling on people's perception of true claims. To not bias respondents' answers to other parts of the survey, the experiment is conducted on the last claim seen by each respondent. A true, human-generated claim not seen previously by the respondent is shown with either a grey or blue label. At first blush, the AI label does not appear to have a large effect. The overall Truth Quest score for true claims (56%) is very close to the Truth Quest score for AI-labelled claims (55%) regardless of whether respondents saw a grey or a blue label.

However, it appears that people's perception of AI as a positive or negative force affects their ability to identify the claim as true. Indeed, people who agree AI will have a positive impact on their life are more likely to correctly identify the veracity of the claim the AI label (Figure 11). Given the claim is always true, people with a positive opinion about AI are more likely to say the claim is true and people with a negative opinion are more likely to say it is false. On average, 59% of respondents with a positive perception of AI correctly responded "true" to the AI-labelled question and 52% of those who don't agree with the positive impact of AI on their life. The gap between the two groups jumps to 12 pp when only respondents who strongly agree and strongly disagree are considered.

Figure 11. Perceptions of AI and Truth Quest score for AI-labelled content

Share of respondents who feel AI will have a positive impact on their life and Truth Quest scores for AI-labelled content, 2024



Note: The "Agree" category groups the "Strongly agree" and "Agree" sub-components, and the "Disagree" category groups the "Strongly disagree" and "Disagree" sub-components. The score is calculated as the total number of correct responses divided by the total number of claims seen. A country score is thus an average of all respondents' results and expressed in percentages. The average is calculated as a simple unweighted average across the 21 country scores covered by Truth Quest. Adults are defined as people aged 18 and older.
Source: OECD (2024[45]), "Share of adults who feel AI will have a positive impact on their life", OECD Going Digital Toolkit, based on the OECD Truth Quest Survey, 2024, https://goingdigital.oecd.org/indicator/82.

Important cross-country differences emerge for those who have a negative view of AI. In Finland, only 48% of respondents who do not agree that AI will have a positive impact on their life responded correctly to the AI labelled claim while 69% did so in the group of respondents who have a positive opinion of AI, a gap of 21pp. Conversely, in Belgium, Luxembourg and Switzerland, the gap between the two groups is below 3pp.

Likewise, the share of respondents who agree that AI will have a positive impact on their life also differs across countries. Latin American respondents show the most positive perception of AI, while European respondents are generally more sceptical. In Mexico and Brazil, 40% of respondents have a positive opinion of AI's impact on their life, compared to less than 25% of respondents in Belgium, France, the Netherlands and Norway. On average, 29% of respondents have a positive perception of AI and 26% do not. Nearly half of respondents neither agree nor disagree.

Not surprisingly, younger people tend to have a more positive perception of the impact of AI on their life. On average, 39% of respondents aged 18-34 hold a positive view of AI compared to only 22% of those aged 55 and older. In Latin America and Japan, the age gap is below 10 pp whereas in the United States it is over 40 pp. Across all countries, men have a more positive perception of AI than women, with an average gap between the two genders of 10 pp. The biggest gender gap can be found in the United States (17 pp).

In terms of income, a positive perception of AI increases with the level of income except in France. In Finland and Poland, the gap between the highest and lowest income groups is over 20 pp. Across countries those with tertiary education are on average 5 pp more likely to have a positive opinion about AI compared to those with low or no education with the biggest gap in Finland (17 pp). In Norway and Poland, respondents with tertiary education are 11 pp and 9 pp respectively less likely to view AI as a positive force than those with low or no education.

Importantly, there is no difference in the overall Truth Quest scores of respondents with positive and negative perceptions of the impact of AI on their life (Table 3). This indicates that it is not the ability to identify false and misleading content online that is driving the difference between these two groups with respect to the claim, but rather the presence of the AI label.

Table 3. Perceptions of AI and overall Truth Quest score

2024

Positive perception of the impact of AI	Overall Truth Quest Score
Agree	59%
Disagree	60%
Neither agree nor disagree	61%

Note: The "Agree" category groups the "Strongly agree" and "Agree" sub-components, and the "Disagree" category groups the "Strongly disagree" and "Disagree" sub-components. The overall Truth Quest score is on a scale of 0-100 (100 = best performance) and expressed in percentages. The score is calculated as the total number of correct responses divided by the total number of claims seen per country. The average is calculated as a simple unweighted average of the 21 country scores covered by Truth Quest.
Source: OECD (2024[45]), "Share of adults who feel AI will have a positive impact on their life", OECD Going Digital Toolkit, based on the OECD Truth Quest Survey, 2024, https://goingdigital.oecd.org/indicator/82.

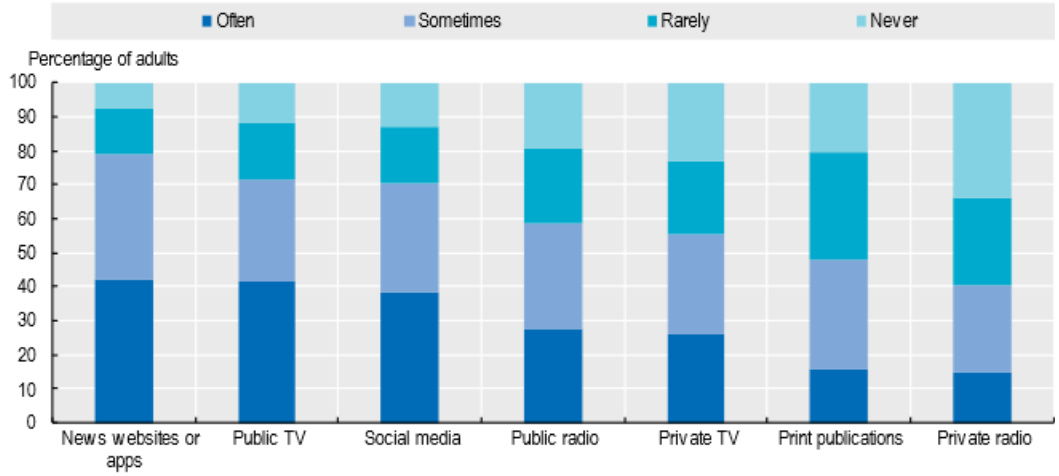
AI labelling is a mechanism that has been promoted in some policy circles as a way to mitigate the risks of generative AI. The Truth Quest results suggest labelling may not have a neutral effect on people's perceptions about such content, with those who have a positive perception of AI generally more likely to correctly identify the true claim with the AI label. Additional research, including on false content, could shed further light on the effects of AI labelling.

Social media is a popular news source, and those with relatively lower Truth Quest scores trust social media the most

Truth Quest posed a series of questions related to media consumption and perceived trust in various forms of media (see Table A B.3). On average, news websites and apps together with public television (TV) are the most frequent sources of news, followed closely by social media (Figure 12). For these three categories of media, at least 70% of Truth Quest respondents sometimes or often get their news from these sources. Other media sources include public radio and private TV, followed by print publications and private radio.

Figure 12. Media consumption patterns

News source by frequency, 2024

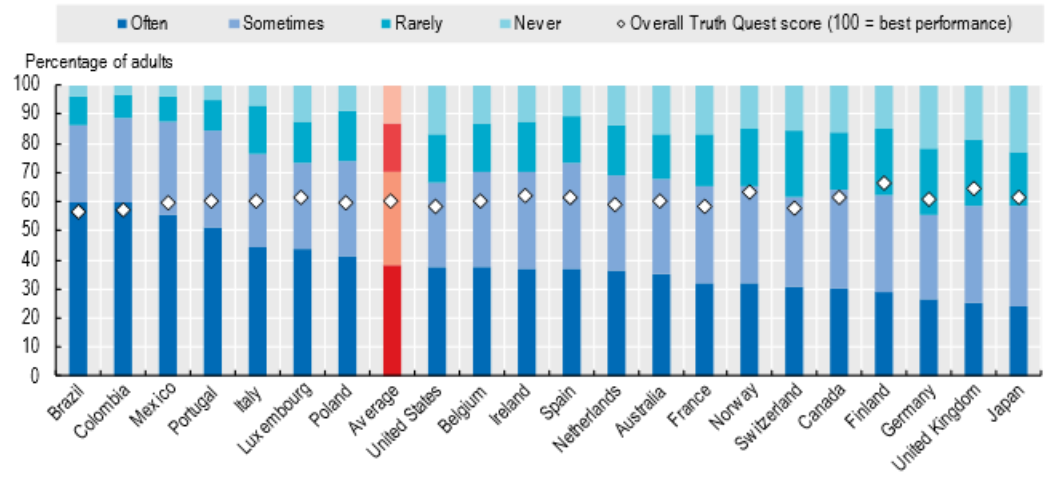


Note: Adults are defined as people aged 18 and older.
Source: Authors' calculations based on the OECD Truth Quest Survey, 2024.

These overall patterns mask significant cross-country variation (Figure 13). For example, in the Latin American countries covered (Colombia, Mexico and Brazil), over 85% of people get their news often or sometimes from social media. Together with Portugal, over half of respondents often get their news from social media in these four countries. On the other side of the spectrum, less than 60% of people get news often or sometimes from social media in Germany, Japan and the United Kingdom. Understanding the source of news is important for targeting media literacy strategies, programmes and policies at the country level.

Figure 13. Consumption of news on social media

By frequency, 2024

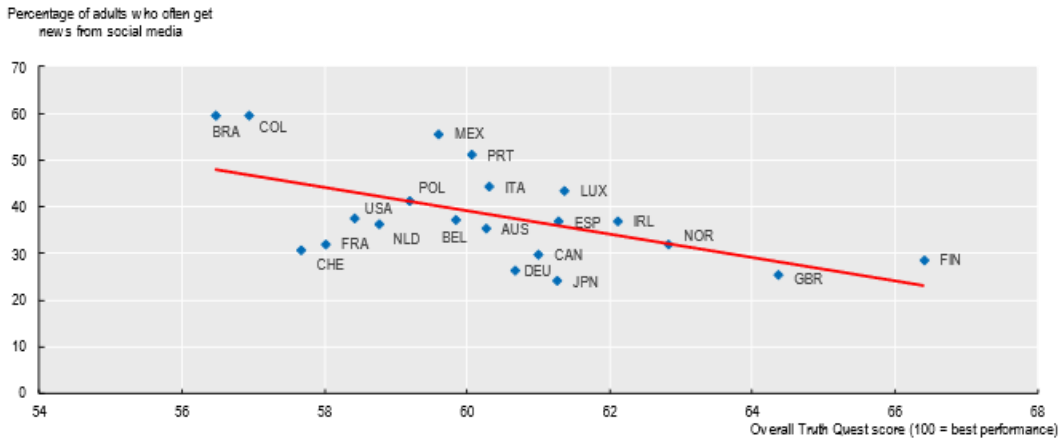


Note: The score is calculated as the total number of correct responses divided by the total number of claims seen. A country score is thus an average of all respondents' results and expressed in percentages. The average is calculated as a simple average of the 21 country scores covered by Truth Quest. Note: Adults are defined as people aged 18 and older.
Source: Authors' calculations based on the OECD Truth Quest Survey, 2024.

Countries with the highest shares of respondents that source their news from social media have lower overall Truth Quest scores (Figure 14). Inversely, countries with the highest Truth Quest scores also have the lowest shares of people sourcing their news from social media. This is a counterintuitive finding given the Truth Quest interface mimics a social media news post. *A priori*, one would have expected that people who more often see news on social media would be more skilled at detecting false and misleading content in the same environment.

Figure 14. Truth Quest score and percentage of adults who often get news from social media

2024

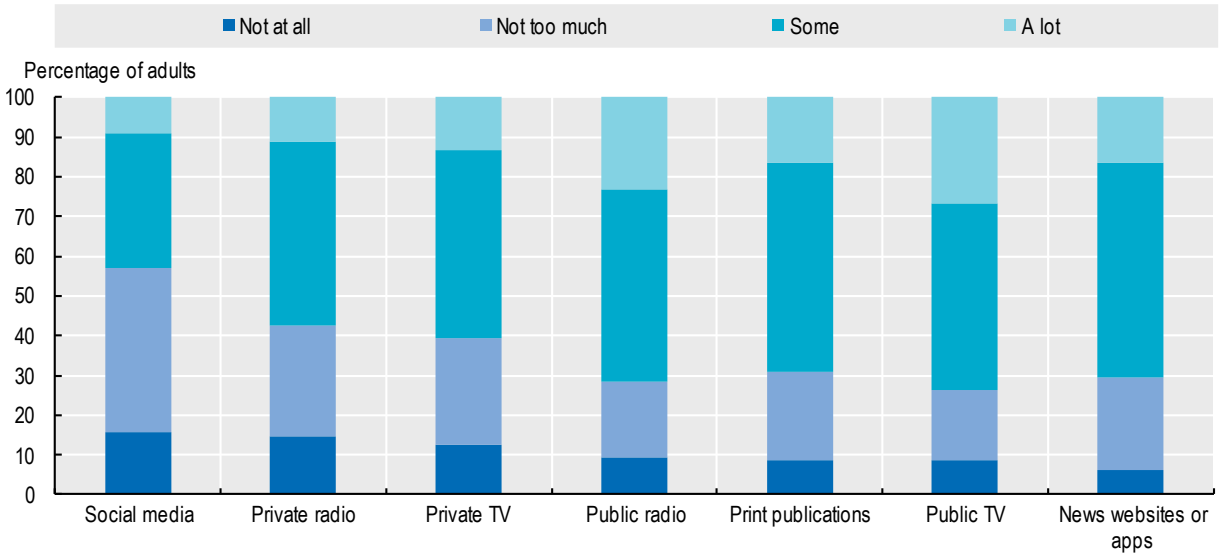


Note: The score is calculated as the total number of correct responses divided by the total number of claims seen. A country score is thus an average of all respondents' results and expressed in percentages. Adults are defined as people aged 18 and older.
Source: Authors' calculations based on the OECD Truth Quest Survey, 2024.

While people often source their news from social media, it is also the least trusted source of news (Figure 15) with 57% of people on average not trusting it too much or at all as a reliable source of news. In contrast, only 9% trust news on social media a lot. Public news sources (TV and radio) are the most trusted source of news across all countries overall, followed by news websites and apps and print publications. Private news sources (TV and radio) have relatively lower perceptions of trust.

Figure 15. Trust in news sources

2024

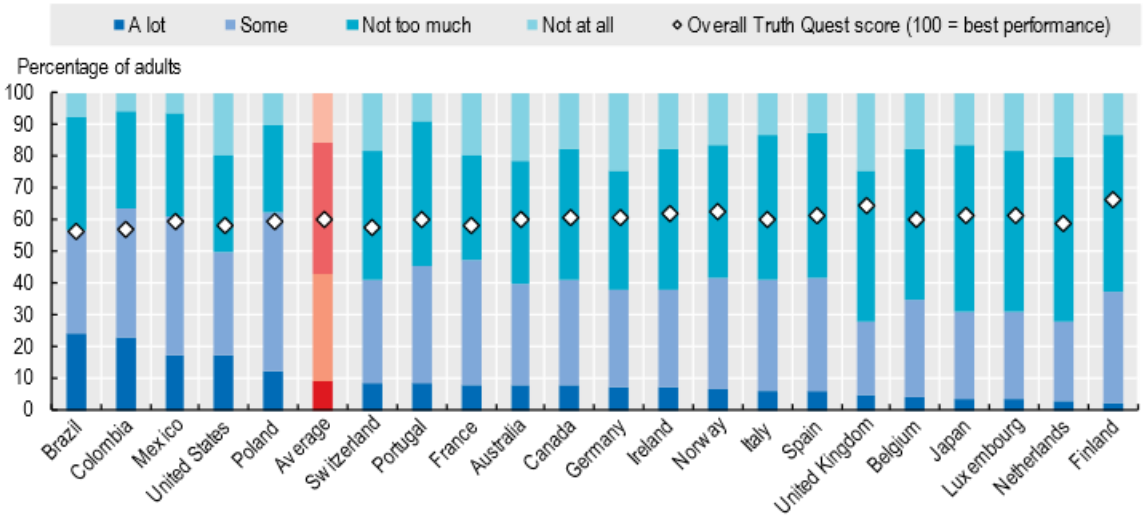


Note: Adults are defined as people aged 18 and older.
Source: Authors' calculations based on the OECD Truth Quest Survey, 2024.

While social media is the least trusted news source overall, there are significant differences across countries (Figure 16). For example, 11% of respondents trust public media a lot in Poland compared to 50% in Finland. Respondents from the United Kingdom have the lowest trust in news from social media, with about a quarter of people trusting social media news some or a lot. In contrast, nearly two-thirds of respondents in Colombia trust the news from social media to some extent and 23% trust it a lot. Again, regional differences are present, with Latin American countries trusting news on social media more than others.

Figure 16. Trust in news from social media

2024



Note: The score is calculated as the total number of correct responses divided by the total number of claims seen. A country score is thus an average of all respondents' results and expressed in percentages. The average is calculated as a simple average of the 21 country scores covered by Truth Quest. Adults are defined as people aged 18 and older.
Source: OECD (2024[46]), "Share of adults who trust news from social media sites or apps", OECD Going Digital Toolkit, based on the OECD Truth Quest Survey, 2024, https://goingdigital.oecd.org/indicator/83.

Taken together, the survey results show that respondents who trust news from social media have lower Truth Quest scores (Table 4). On average across countries, those who trust social media a lot had a relatively lower Truth Quest score (54%) compared to those who trust news on social media somewhat (59%) and not much or not at all (62%). In the United Kingdom, the difference between those who trust news on social media a lot and not at all is 14 pp. However, in countries where trust in social media is high, such as Brazil or Colombia, the difference is relatively small (3 pp).

Table 4. Truth Quest score and trust in news from social media

2024

Trust in news from social media	Overall Truth Quest Score
A lot	54%
Some	59%
Not too much	62%
Not at all	62%

Note: The overall Truth Quest score is on a scale of 0-100 (100 = best performance). The score is calculated as the total number of correct responses divided by the total number of claims seen per country. The average is calculated as a simple unweighted average of the 21 country scores covered by Truth Quest and expressed in percentages.
Source: Authors' calculations based on the OECD Truth Quest Survey, 2024.

5 Conclusion

False and misleading content online poses significant risks to the well-being of people and society, but a lack of cross-country comparable evidence persists. This paper contributes to the statistical literature in this area by presenting the methodology and results from the OECD Truth Quest Survey. The cross-country comparable data from the survey will help policy makers better understand the mechanisms underlying the diffusion of false and misleading content online. Country-specific findings will enable the more targeted design of media literacy strategies, programmes and related policies to address the negative effects of such content.

The overall Truth Quest scores are based on true and false content in three main themes: the environment, health and international affairs. Differences in people's ability to identify the five types of false and misleading content in the OECD taxonomy of false and misleading content was also assessed. This core part of the survey was complemented by analysis of AI-generated true content and disinformation, and by studying the impact of AI labels on people's ability to correctly identify the veracity of content online.

While this paper presents some of the rich data obtained in Truth Quest, further analysis will be possible in other areas, including multivariate analysis to study interactions between demographic variables as well as those related to respondents' behaviour during their Truth Quest experience. For example, is the time spent on completing the survey or is clicking "more" to read additional context correlated with overall Truth Quest scores? How do perceptions about privacy on social media, democracy in the countries surveyed, and media manipulation affect, if at all, respondents' scores?

Looking ahead, in addition to covering more countries and potentially additional themes, research into sharing behaviour could be undertaken in a future edition of Truth Quest. Partnerships with national statistical organisations could also be explored with a view to implementing Truth Quest on a periodic basis. In addition, further exploration of questions related to AI could be undertaken. In particular, several large language models could be used to verify the results obtained in Truth Quest that AI-generated content is easier to identify than human-generated content. Moreover, the impact of AI labelling could be further explored by testing the impact of labelling on false content as well. Skills related to identifying AI-generated images could be explored in a future edition of the survey.

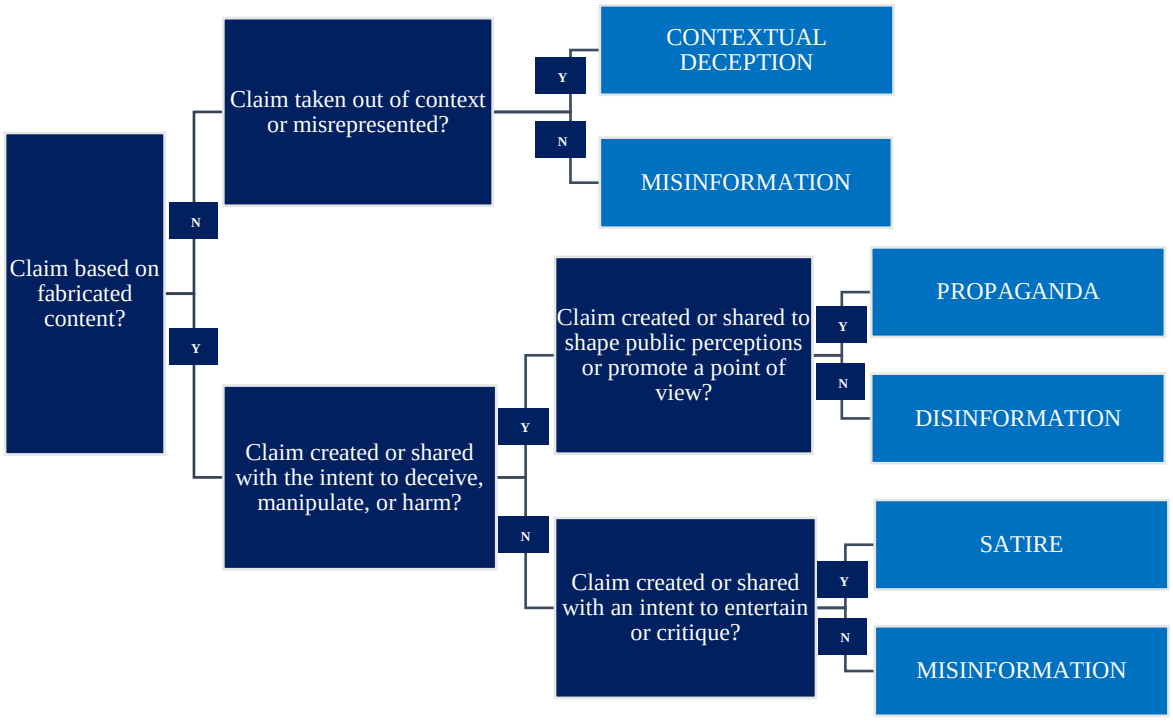
Annex A. The OECD taxonomy of false and misleading content online and its implementation in Truth Quest

The OECD taxonomy of false and misleading content includes five distinct types (Lesher, Pawelec and Desai, 2022[1]):

- **Disinformation**: Disinformation refers to verifiably false or misleading information that is knowingly and intentionally created and shared for economic gain or to deliberately deceive, manipulate or inflict harm on a person, social group, organisation or country. Fake news, synthetic media, including deepfakes, and hoaxes are forms of disinformation, among others.
- **Misinformation**: Misinformation refers to false or misleading information that is shared unknowingly and is not intended to deliberately deceive, manipulate or inflict harm on a person, social group, organisation or country. Importantly, the spreader does not create or fabricate the initial misinformation content.
- **Contextual Deception**: Contextual deception refers to the use of true but not necessarily related information to frame an event, issue or individual (e.g., a headline that does not match the corresponding article), or the misrepresentation of facts to support one's narrative (e.g., to deliberately delete information that is essential context to understanding the original meaning). While the facts used are true (unlike disinformation) and unfabricated (unlike misinformation), the way in which they are used is disingenuous and with the intent to manipulate people or cause harm.
- **Propaganda:** Refers to the activity or content adopted and propagated by governments, private firms, non-profits and individuals to manage collective attitudes, values, narratives and opinions. While propaganda can contain both true and untrue elements, it is often used to appeal to an individual's or social group's sentiments and emotions rather than being informative.
- **Satire:** Satire is defined as language, film or other works of art that use humour and exaggeration to critique people or ideas, often as a form of social or political commentary. Satire is an important form of social and political criticism, using humour and wit to draw attention to issues in society, and when satire is first published, the viewer often recognises the content as satire in part because of where and how they view it (e.g., directly from a satirical newspaper). However, as the content is shared and re-shared, this connection is sometimes lost intentionally (or not) by the spreader, leading new viewers to misunderstand the original meaning.

Classifying the claims in the Truth Quest database was undertaken with a decision tree (Figure A A.1). The decision tree leverages the definitions outlined in the taxonomy of false and misleading content online, identifies key elements that differentiate unique forms of such content based on these definitions, and proposes a logical flow of considerations to guide and structure the classification process.

Figure A A.1. Decision tree to categorise the Truth Quest claims



Note: Y = Yes and N = No.
Source: Authors' elaboration.

The decision tree provides a practical and process-oriented approach for implementing the OECD taxonomy. However, challenges remain in practically implementing the taxonomy. For example, a piece of false and misleading content circulated online can have a different classification based on the context in which it was shared and its source. An untrue claim intentionally fabricated to mislead may be classified as disinformation when shared by the original creator or fabricator but as misinformation when later shared by an individual who believes the claim to be true. Moreover, intent plays a central role in differentiating between different forms of false and misleading content, but it can be challenging to discern an actor's intention in circulating false and misleading content online, particularly when content is viewed out of context.

Nonetheless, the vast databases of claims that were used as a basis to construct the Truth Quest database were helpful insofar that questionable claims could be discarded. The research into origin can also help overcome some of these issues given that only true and disinformation claims were generated, and as such intent (to deceive) and role (as creator) is clear.

Annex B. Statistical tables

Table A B.1. Population and quotas used for targeting

Country	Total number of respondents	Age	Gender	Sub-national region	Income	Education
Australia	2 021	Y	Y	Y	Y	Y
Belgium	2 014	Y	Y	Y	Y	Y
Brazil	2 013	Y	Y	Y	Y	Y
Canada	2 039	Y	Y	Y	Y	Y
Colombia	2 041	Y	Y	Y	N*	Y
Germany	2 006	Y	Y	Y	Y	Y
Finland	1 902	Y	Y	Y	Y	Y
France	2 002	Y	Y	Y	Y	Y
Ireland	2 020	Y	Y	Y	N*	N*
Italy	2 021	Y	Y	Y	N*	Y
Japan	2 012	Y	Y	Y	N*	N*
Luxembourg	1 503	Y	Y	Y	N*	N*
Mexico	2 005	Y	Y	Y	Y	Y
Netherlands	2 053	Y	Y	Y	N*	Y
Norway	1 811	Y	Y	Y	N*	Y
Poland	2 011	Y	Y	Y	N*	Y
Portugal	1 710	Y	Y	Y	N*	Y
Spain	2 007	Y	Y	Y	Y	Y
Switzerland	1 531	Y	Y	Y	Y	Y
United Kingdom	2 021	Y	Y	Y	Y	Y
United States	2 022	Y	Y	Y	Y	Y

Note: * Quotas used not for targeting, but for post-stratification weighting.
Source: OECD Truth Quest Survey, 2024.

Table A B.2. Matrix of claims in the Truth Quest database

Theme	Type	Origin	Veracity	Number of claims
Environment	Disinformation	Human	False	2
Environment	Disinformation	AI	False	2
Environment	Misinformation	Human	False	2
Environment	Contextual deception	Human	False	2
Environment	Propaganda	Human	False	2
Environment	Satire	Human	False	2
Environment	Truth	Human	True	4
Environment	Truth	AI	True	2
Health	Disinformation	Human	False	2
Health	Disinformation	AI	False	2
Health	Misinformation	Human	False	2
Health	Contextual deception	Human	False	2
Health	Propaganda	Human	False	2
Health	Satire	Human	False	2
Health	Truth	Human	True	4
Health	Truth	AI	True	2
International affairs	Disinformation	Human	False	2
International affairs	Disinformation	AI	False	2
International affairs	Misinformation	Human	False	2
International affairs	Contextual deception	Human	False	2
International affairs	Propaganda	Human	False	2
International affairs	Satire	Human	False	2
International affairs	Truth	Human	True	4
International affairs	Truth	AI	True	2

Source: OECD Truth Quest Survey, 2024.

Table A B.3. Key behavioural and perception-related questions in Truth Quest

Question	Response options	Position of the question
How confident do you feel that you can recognise false and misleading online content when you encounter it (e.g., disinformation, fake news)?	• Very confident • Somewhat confident • Not very confident • Not at all confident	First question prior to seeing the first claim
How often do you get news from… • Public television • Private television • Public radio • Private radio • Print publications • News websites or apps • Social media sites or apps (e.g., Facebook, Twitter/X, Instagram, WhatsApp, Telegram, etc.)	• Often • Sometimes • Rarely • Never	First question after finishing the last claim
How much, if at all, do you trust the information you get from… • Public television • Private television • Public radio • Private radio • Print publications • News websites or apps • Social media sites or apps (e.g., Facebook, Twitter/X, Instagram, WhatsApp, Telegram, etc.)	• A lot • Some • Not too much • Not at all	Second question after finishing the last claim
How much do you agree or disagree with the following statements? • All in all, democracy works well in [country name] today. • It is important to tolerate opinions that you disagree with. • Politicians and the media collude to manipulate public opinion. • A democracy can function without independent journalism. • Artificial intelligence will have a positive impact on my life. • I avoid using certain websites, apps, or social media due to privacy concerns. • I feel I have control over my personal information when using websites, apps, or social media.	• Strongly agree. • Agree • Neither agree nor disagree • Disagree • Strongly disagree	Third question after finishing the last claim

Source: OECD Truth Quest Survey, 2024.

References

Ahmed, W. et al. (2020), "COVID-19 and the 5G conspiracy theory: Social network analysis of twitter data", *Journal of Medical Internet Research*, Vol. 22/5, https://doi.org/10.2196/19458. [23]

Allcott, H. and M. Gentzkow (2017), "Social media and fake news in the 2016 election", *Journal of Economic Perspectives*, Vol. 31/2, pp. 211-236, https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211. [6]

Allen, J. et al. (2020), "Evaluating the fake news problem at the scale of the information ecosystem", *Science Advances*, Vol. 6/14, https://www.science.org/doi/10.1126/sciadv.aay3539. [8]

American Psychological Association (2024), *APA Dictionary of Psychology*, https://dictionary.apa.org/emotional-valence (accessed on 14 March 2024). [50]

Basol, M. et al. (2021), "Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation", *Big Data & Society*, Vol. 8/1, https://doi.org/10.1177/20539517211013868. [32]

Basol, M., J. Roozenbeek and S. van der Linden (2020), "Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news", *Journal of Cognition*, Vol. 3/1, http://doi.org/10.5334/joc.91. [33]

Bond, C. and B. DePaulo (2006), "Accuracy of deception judgments", *Personality and Social Psychology Review*, Vol. 10/3, pp. 214-234, https://doi.org/10.1207/s15327957pspr1003_2. [3]

Cravath, Swaine & Moore LLP (2023), *Tech Explainers, Watermarking of AI-Generated Text*, https://www.cravath.com/news/cravath-publishes-tech-explainer-on-the-watermarking-of-aigenerated-text.html. [49]

Deterding, S. et al. (2011), "From game design elements to gamefulness: Defining 'gamification'", *Proceedings of the 15th International Academic MindTrek Conference*, pp. 9-15, https://doi.org/10.1145/2181037.2181040. [48]

European Commission (2018), *Fake news and disinformation online*, Flash Eurobarometer 464, https://europa.eu/eurobarometer/surveys/detail/2183. [42]

Eurostat (2021), *How many people verified online information in 2021?*, https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20211216-3 (accessed on 22 February 2024). [25]

Evans, J. and A. Mathur (2005), "The value of online surveys", *Internet Research*, Vol. 15/2, pp. 195-219, https://doi.org/10.1108/10662240510590360. [34]

Fletcher, R. et al. (2018), *Measuring the reach of "fake news" and online disinformation in Europe*, Reuters Institute, https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/Measuring%20the%20reach%20of%20fake%20news%20and%20online%20distribution%20in%20Europe%20CORRECT%20FLAG.pdf. [7]

Goldstein, J. et al. (2023), "Generative language models and automated influence operations: Emerging threats and potential mitigations", *arXiv*, No. 2301.04246, https://arxiv.org/abs/2301.04246. [19]

Guess, A., J. Nagler and J. Tucker (2019), "Less than you think: Prevalence and predictors of fake news dissemination on Facebook", *Science Advances*, Vol. 5/1, https://doi.org/10.1126/sciadv.aau4586. [5]

Hamari, J., J. Koivisto and H. Sarsa (2014), "Does gamification work? A literature review of empirical studies on gamification", *2014 47th Hawaii International Conference on System Sciences*, pp. 3025-3034, https://doi.org/10.1109/HICSS.2014.377. [29]

Hameleers, M. and T. Van der Meer (2020), "Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers?", *Communication Research*, Vol. 47/2, pp. 227-250, https://doi.org/10.1177/0093650218819671. [39]

Hanley, H. and Z. Durumeric (2023), "Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites", *arXiv*, No. 2305.09820, https://doi.org/10.48550/arXiv.2305.09820. [16]

Hargittai, E. (2002), "Second-level digital divide: Mapping differences in people's online skills", *First Monday*, Vol. 7/4, https://doi.org/10.5210/fm.v7i4.942. [37]

Harms, J. et al. (2015), "Gamification of online surveys: Design process, case study, and evaluation", *15th Human-Computer Interaction (INTERACT)*, Bamberg, Germany, pp. 219-236, https://doi.org/10.1007/978-3-319-22701-6_16. [30]

Heerwegh, D. (2009), "Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects", *International Journal of Public Opinion Research*, Vol. 21/1, pp. 111-121, https://doi.org/10.1093/ijpor/edn054. [35]

Keusch, F. and C. Zhang (2016), "A review of issues in gamified surveys", *Social Science Computer Review*, Vol. 35/2, pp. 147-166, https://doi.org/10.1177/0894439315608451. [31]

Knuutila, A., L. Neudert and P. Howard (2022), "Who is afraid of fake news? Modeling risk perceptions of misinformation in 142 countries", *Harvard Kennedy School (HKS) Misinformation Review*, Vol. 3/3, https://doi.org/10.37016/mr-2020-97. [27]

Kreps, S., R. McCain and M. Brundage (2022), "All the news that's fit to fabricate: AI-generated text as a tool of media misinformation", *Journal of Experimental Political Science*, Vol. 9/1, pp. 104-117, https://doi.org/10.1017/XPS.2020.37. [17]

Lesher, M., H. Pawelec and A. Desai (2022), "Disentangling untruths online: Creators, spreaders and how to stop them", *OECD Going Digital Toolkit Notes*, No. 23, OECD Publishing, Paris, https://doi.org/10.1787/84b62df1-en. [1]

Longoni, C. et al. (2022), "News from generative artificial intelligence is believed less", *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*, Seoul, Korea, pp. 97-106, https://doi.org/10.1145/3531146.3533077. [21]

Lorenz, P., K. Perset and J. Berryhill (2023), "Initial policy considerations for generative artificial intelligence", *OECD Artificial Intelligence Papers*, No. 1, OECD Publishing, Paris, https://doi.org/10.1787/fae2d1e6-en. [47]

Luke, T. (2019), "Lessons from Pinocchio: Cues to deception may be highly exaggerated", *Perspectives on Psychological Science*, Vol. 14/4, pp. 646-671, https://doi.org/10.1177/1745691619838258. [4]

Newman, N. et al. (2023), *Digital News Report 2023*, Reuters Institute for the Study of Journalism, https://doi.org/10.60625/risj-p6es-hb13. [26]

OECD (2024), "Ability of adults to identify online disinformation created by generative AI", *OECD Going Digital Toolkit*, https://goingdigital.oecd.org/indicator/81 (accessed on 28 June 2024). [44]

OECD (2024), "Ability of adults to identify the veracity of online news", *OECD Going Digital Toolkit*, https://goingdigital.oecd.org/indicator/80 (accessed on 28 June 2024). [43]

OECD (2024), "Internet users as a share of individuals", *OECD Going Digital Toolkit*, https://goingdigital.oecd.org/indicator/20 (accessed on 12 March 2024). [36]

OECD (2024), "Share of adults who feel AI will have a positive impact on their life", *OECD Going Digital Toolkit*, https://goingdigital.oecd.org/indicator/82 (accessed on 28 June 2024). [45]

OECD (2024), "Share of adults who trust news from social media sites or apps", *OECD Going Digital Toolkit*, https://goingdigital.oecd.org/indicator/83 (accessed on 28 June 2024). [46]

OECD (2022), *Programme for International Student Assessment (database)*, https://www.oecd.org/pisa/data/2022database/ (accessed on 12 March 2024). [51]

Pennycook, G. and D. Rand (2020), "Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking", *Journal of Personality*, Vol. 88/2, pp. 185-200, https://doi.org/10.1111/jopy.12476. [15]

Pennycook, G. and D. Rand (2019), "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning", *Cognition*, Vol. 188, pp. 39-50, https://doi.org/10.1016/j.cognition.2018.06.011. [10]

Roozenbeek, J. et al. (2022), "Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking", *Judgment and Decision Making*, Vol. 17/3, pp. 547-573, https://doi.org/10.1017/s1930297500003570. [12]

Roozenbeek, J. et al. (2020), "Susceptibility to misinformation about COVID-19 around the world", *Royal Society Open Science*, Vol. 7/10, https://doi.org/10.1098/rsos.201199. [13]

Rutjens, B. and R. van der Lee (2020), "Spiritual skepticism? Heterogeneous science skepticism in the Netherlands", *Public Understanding of Science*, Vol. 29/3, pp. 335-352, https://doi.org/10.1177/0963662520908534. [38]

Saltz, E., C. Leibowicz and C. Wardle (2021), "Encounters with visual misinformation and labels across platforms: An interview and diary study to inform ecosystem approaches to misinformation interventions", *CHI '21: CHI Conference on Human Factors in Computing Systems*, No. 340, Yokohama, Japan, pp. 1-6, https://doi.org/10.1145/3411763.3451807. [41]

Seaborn, K. and D. Fels (2015), "Gamification in theory and action: A survey", *International Journal of Human-Computer Studies*, Vol. 74, pp. 14-31, https://doi.org/10.1016/j.ijhcs.2014.09.006. [28]

Solaiman, I. et al. (2019), "Release strategies and the social impacts of language models", *arXiv*, No. 1908.09203, https://arxiv.org/abs/1908.09203. [18]

Southwell, B. et al. (2022), "Defining and measuring scientific misinformation", *The ANNALS of the American Academy of Political and Social Science*, Vol. 700/1, pp. 98-111, https://doi.org/10.1177/00027162221084709. [9]

UNESCO (2023), *Online disinformation : UNESCO unveils action plan to regulate social media platforms*, https://www.unesco.org/en/articles/online-disinformation-unesco-unveils-action-plan-regulate-social-media-platforms (accessed on 22 February 2024). [24]

Wall Street Journal (2023), *The Facebook Files*, (accessed on 8 March 2023), https://www.wsj.com/articles/the-facebook-files-11631713039. [2]

Watts, D., D. Rothschild and M. Mobius (2021), "Measuring the news and its impact on democracy", *The Proceedings of the National Academy of Sciences*, Vol. 118/15, https://doi.org/10.1073/pnas.1912443118. [11]

Wittenberg, C. et al. (2024), "Labeling AI-generated content: Promises, perils, and future directions", *An MIT Exploration of Generative AI*, https://doi.org/10.21428/e4baedd9.0319e3a6. [40]

Zeng, J. and C. Chan (2021), "A cross-national diagnosis of infodemics: comparing the topical and temporal features of misinformation around COVID-19 in China, India, the US, Germany and France", *Online Information Review*, Vol. 45/4, pp. 709-728, https://doi.org/10.1108/OIR-09-2020-0417. [14]

Zhang, Y. and R. Gosline (2023), "Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human-GAI collaboration in persuasive content generation", *Judgment and Decision Making*, Vol. 18, https://doi.org/10.1017/jdm.2023.37. [20]

Zubiaga, A. et al. (2016), "Analysing how people orient to and spread rumours in social media by looking at conversational threads", *PLOS ONE*, Vol. 11/3, https://doi.org/10.1371/journal.pone.0150989. [22]

Endnotes

[1] Generative AI refers to algorithms that are capable of producing new content such as texts, images, videos, audio or code (Lorenz, Perset and Berryhill, 2023[47]). Large language models, for example, are a form of generative AI that can be used to generate texts and serve as the foundation of technology such as ChatGPT.

[2] Gamification has been defined as "the use of game design elements in non-game contexts" (Deterding et al., 2011[48]) or the "process of enhancing services with (motivational) affordances in order to invoke game-like experiences and further behavioural outcomes" (Hamari, Koivisto and Sarsa, 2014[29]).

[3] Go Viral! is a free game online developed to help protect people against false and misleading content online related to COVID-19 by teaching game participants about common techniques used to disseminate such content. The game was developed by researchers at the University of Cambridge in collaboration with the UK Cabinet Office and other stakeholders.

[4] Bad News is a free game online developed by researchers at the University of Cambridge, in collaboration with other stakeholders, that was created to help individuals build "cognitive resistance against common forms of manipulation [they] may encounter online."

[5] Such activities involve exposure to "weaker" forms of false and misleading content and the techniques used to produce it.

[6] The OECD Truth Quest Survey was administered to adults which are defined as people aged 18 and older.

[7] GPT-4 was used to create false and misleading content for the Truth Quest database. GPT-4 is the latest version of Generative Pre-trained Transformers, a type of deep learning model used for natural language processing and text generation.

[8] The prompt used to generate the AI claims is: "According to the OECD definition, disinformation refers to verifiably false or misleading information that is knowingly and intentionally created and shared for economic gain or to deliberately deceive, manipulate or inflict harm on a person, social group, organisation or country (EC, 2019). Fake news, synthetic media, including deepfakes, and hoaxes are forms of disinformation, among others. Please generate one surprising short claim that is true and one that falls under this definition. It should be in the [health/environment/international affairs] theme."

[9] In the context of generative AI, the term “watermark” may also refer to information embedded within AI-generated content, in a technical manner, that shares details regarding the provenance of the information (Cravath, Swaine & Moore LLP, 2023[49]; Wittenberg et al., 2024[40]).

[10] Other thematic claims were included in the Truth Quest database and administered to respondents, but they were experimental and do not represent the core of the survey.

[11] Emotional valence is defined as “the value associated with a stimulus as expressed on a continuum from pleasant to unpleasant or from attractive to aversive. In factor analysis and multidimensional scaling studies, emotional valence is one of two axes (or dimensions) on which an emotion can be located, the other axis being arousal (expressed as a continuum from high to low). For example, happiness is typically characterized by pleasant valence and relatively high arousal, whereas sadness or depression is typically characterized by unpleasant valence and relatively low arousal” (American Psychological Association, 2024[50]).

[12] 54 claims were chosen from professionally fact-checked sources and the veracity of the claims was based on this fact-checked assessment. The 12 AI-related claims were fact-checked by the Authors.

[13] MTurk is a crowdsourcing platform used to engage people to perform on-demand tasks, such as answering survey questions.

[14] While 66 claims are included in the Truth Quest database, this paper describes the results for 54 claims only (see Table A B.2). The other claims were experimental and do not form the basis of the findings in this paper.

[15] All images except for two were selected from Shutterstock’s main image library, which does not include AI-generated content. Two images were selected from Dreamstime’s stock of images that are not identified as being generated by AI. While none of the photos were identified as AI-generated, it isn’t possible to ascertain if editing tools using AI were applied to the photos.

[16] On average across countries, the avatar chosen did not impact Truth Quest scores. Gamification may also have improved the respondent experience, as minimal respondent fatigue was observed.

[17] Respondents expected to see both true and false content in the survey, but they were unaware of the proportion of true and false claims they would see, thus mitigating the potential bias of resulting from greater vigilance.

[18] Truth Quest scores were also correlated with OECD Programme for International Student Assessment (PISA) scores for 2022, but no significant correlation was observed (OECD, 2022[51]).

[19] See notes to Figure 5.