WAN IFRA | REPORT

News Automation

The rewards, risks and realities of 'machine journalism'

MARCH 2019

IMPRINT

NEWS AUTOMATION – A WAN-IFRA GUIDE TO THE FIELD

PUBLISHED BY:
WAN-IFRA
Rotfeder-Ring 11
60327 Frankurt, Germany

CEO:
Vincent Peyrègne

COO:
Thomas Jacob

DIRECTOR OF INSIGHTS:
Dean Roper

EDITORS:
Lindén, Carl-Gustav
Tuulonen, Hanna

CO-AUTHORS:
Bäck, Asta
Diakopoulos, Nicholas
Granroth-Wilding, Mark
Haapanen, Lauri
Leppänen, Leo
Melin, Magnus
Moring, Tom
Munezero, Myriam
Sirén-Heikel, Stefanie
Södergård, Caj
Toivonen, Hannu

COPY EDITING:
Simone Flückiger, Steve Shipside

DESIGN/LAYOUT:
Ivan Cosic & Snezana Vukmirovic,
Plain&Hill

CONTENTS

About the report

Dean Roper
Director of Insights
WAN-IFRA

Journalism created by machines. You can count me in the camp that was wary of this development years ago when whispers of algorithms possibly cranking out content started to build some volume.

But thanks to advances in technology and SOME adoption in newsrooms, today those whispers have turned into open, practical conversations about the real application of news automation, powered by machine learning and AI.

Much of the rationale for automated news revolves around the idea of letting machines focus on producing "commodity news" so journalists are freed up to focus on more enterprise journalism. But even for the latter, machine learning can prove helpful. Therefore the message from the top tends to be: "why have our journalists waste their precious time when machines can do much of the nitty gritty work for them?" But don't worry, nobody, at least not in this report, is (yet) seriously talking about replacing the journalistic instincts of we humans.

Our conversation about this report began in the fall of 2017 when researcher Carl-Gustav Lindén told us about the Immersive Automation research project he had led in Finland: It was supported (and funded) by a number of research organisations and participating media companies, as well as Business Finland.

One of the main goals of the project was to study the feasibility of automating news and editorial processes, ultimately, creating a solution for producing engaging, data-driven content. The project started in early 2017 and continued until the end of May 2018. To learn more about the difficulties of automation and which parts of editorial processes are in fact worth automating, the Immersive Automation research team developed its first prototype, the multilingual election bot Valtteri, which we cover on page 22.

Having learned a lot from that project, Carl-Gustav suggested that we cooperate on a larger report, researching and highlighting where the industry stands on news automation. A big thanks to Carl-Gustav and his team, especially Hanna Tuulonen who did the heavy lifting of reporting and writing here.

This report focuses purely on news creation and automation; not so much on how, for example, algorithms are being used for personalization, something we are addressing in a number of our reports. For now, it's all about automated news. Enjoy.

Executive summary

- This report focuses on a specific part of news automation: **the automated generation of news texts based on structured data**. This is not about crystal ball gazing. News automation is already making itself felt in the daily life of newsrooms and the examples presented in this report show how automation can aid journalism as well as the implications, and the ethics involved.

- Media outlets face ever-growing commercial pressure to extract higher margins from dwindling resources and that is a key driver for news automation. Right now, **one of the main goals of automated content is to save journalistic effort, especially on repetitive tasks, while increasing output volume**. Automated production is foremost a tool, aiding and creating added content.

- One of the characteristics of what is labelled "automated news" is that its focus is on writing stories that journalists cannot or do not necessarily have the time to write. The good news is that so far, **news automation has not replaced humans, and looks set to work alongside humans in the newsroom**.

- For all the hype about "robot journalism" we are more or less in the same spot as three years ago. **AI has a hype problem and we need to put aside our Hollywood-inspired ideas about super-advanced AI and instead see the automation process as a logical extension of the Industrial Revolution.** The future of automation lies in decomposition, or deconstruction, of the fundamental principles of journalism. That means breaking down journalistic work into the actual information artefacts and micro processes to analyse what can be automated and what are inherently human tasks.

- In this report, **we present five examples of how news automation has been implemented in newsrooms around the world:** Mitt-Media and United Robots (Sweden), Radar (UK), Washington Post (US), Valtteri (Finland), and Xinhua and Caixin (China).

- Publishers considering implementing news automation systems have a lot of judgement calls to make. **The biggest decision is whether the system should be bought from a service provider or created and modified in-house.** In addition, the approach to implementation of news automation, ethical considerations and transparency should be considered.

- Automated journalism transforms structured data into news articles, and **the quality of the output is highly dependent on the quality of the data that is fed into it.** The quality of data is often described as the five V's: volume, velocity, variety, value and veracity. Volume, variety and velocity are largely relevant from a business perspective, satisfying content-hungry customers and driving revenue streams. Veracity, on the other hand, matters more from an ethical and journalistic viewpoint.

- The process of translating digitally encoded data into human language is called Natural Language Generation (NLG). **There's been a lot of research into NLG, but it remains little exploited in the context of algorithmic journalism**. One reason for this is the complexity of the natural language used in journalistic settings: journalists are extremely skilled at avoiding repetition, and easy-to-implement NLG approaches only really work where the range of possible news stories is relatively limited.

- **Because templates used in news automation are designed by humans, there is a risk that the automation might reflect what the designers consider important.** The looser the template, the greater the chance for discrepancies, and the higher the risk that readers won't buy into it. Beyond fact-belief discrepancies, NLG systems can also produce fact-claim discrepancies, which can also be called "incorrect statements" or more simply "lies".

- In all use cases, good user perceptions are crucial. This report looks at the impact of saying when stories are created automatically, as well as what happens when users are asked to compare content created by machines with that created by journalists.

10 takeaways about news automation

Traditional journalistic virtues still apply to automated news. Local content, fast publishing and a large number of texts are considered strengths in all news articles: algorithm and human written ones.	1
Think about the notion of news value. Instead of finding an angle that everyone finds interesting, the same news story can have several angles depending on who is reading the article.	2
Texts and algorithms should be written well enough so that they can be published directly without human editing.	3
Will your distribution systems and platforms handle the vast amount of texts produced when automating? The ability to produce very large amounts of texts in a short time is one of automation's great strengths.	4
Automation can potentially produce new content categories that attract a particular group of advertisers and content that would convert visitors to paying readers.	5
Automation prompts a rethink in personnel schedules. If texts are automated in the morning you may not need so many staff then. Automation may have uses in adding value not just completing routine reporting.	6
Automated systems can provide sophisticated alerts for journalists and editors when something interesting has happened or is about to happen. Algorithms are much better equipped to find hidden relationships and outliers than people.	7
When implementing and working with news automation, the editorial management team must engage in the project.	8
Transparency is key. Being able to explain how the stories are created is relevant both in-house and towards audiences.	9
Be clear about how the data has been collected and handled, and how the data has affected the outcome of the article. Think about the article's by-line, information and data verifications processes, how to avoid problems concerning personalisation, and corrections policy.	10

Chapter 1

1. What is news automation:

The case of automated text generation

News automation is making its way to newsrooms all over the world. Developed from other disciplines such as computer science, statistics, and engineering, artificial intelligence (AI) tools are now helping journalists tell new kinds of stories that were previously too resource-impractical or technically out of reach (Hansen et al., 2017).

This report focuses on a specific part of news automation: automated generation of news texts based on structured data. News automation is not a radical break with the past but a gradual continuation of the use of software in the newsroom. Which is why we avoid the term 'news robot' and prefer 'news automation' instead.

The examples presented in this report show how automation can aid journalistic production in several ways. Instead of viewing automation as robots replacing humans the potential and benefits for newsrooms should be understood and taught. Automated production is foremost a tool, completing and creating added content depending on how the systems are designed, and ultimately controlled by the newsrooms.

In this report we will elaborate on technical solutions offered by news automation and data, shine a spotlight on the state-of-the-art, and suggest ways news automation will evolve in the future.

1.1. Why automate?

Right now, the main goals with automated content are to save journalistic effort on repetitive tasks, and to increase the output volume, although more sophisticated uses are possible.

Automatic news can take many shapes. For example, in simple automation systems journalists and developers create the lion's share of the text and a machine just fills in the blanks in the relevant template.

In the past 10 years, news automation has increasingly been making journalistic work more productive. Different systems have been deployed in many newsrooms in Europe, the USA, China and the Middle East, making it hard to obtain an overall picture. One indication comes from Europe's largest media market, Germany, where seven per cent of newspaper publishers have experimented with news automation according to a 2017 survey carried out by the Association of German Newspaper Publishers BDZV (2017). Another 20 percent planned to try news automation.

One of the field's pioneers, Kristian Hammond – a computer science professor and co-founder of Narrative Science – explains that news automation systems have a journalistic approach and process of looking at data: they figure out what the facts are, find some overall characterisations of those facts, and analyse the information with regard to what is important and interesting.

Automated news generation has been around for a long time but the first steps towards a new era were taken in 2009 when Narrative Science, then known as Stats Monkey, emerged from Northwestern University in the United States. Ten years later news automation covers sports, weather, financial and election news and basically anything that can be automated including videos, graphics and photo editing.

Drivers for automation include the commercial pressures and higher profit expectations that media outlets face. Another is that the tsunami of structured and unstructured data makes it possible to uncover information and write more articles than ever before, on totally new subjects. But as the amount of data is estimated to double every 40 months (Latar 2015), the quantity of information is too great for journalists to deal with, which makes news automation intriguing.

"Take all of the data and the implications and transform them into an explanation that is readable for someone who wants the information that is most interesting to them right now. That is the story, the natural language narrative."

– Kristian Hammond, Chief Scientist at Narrative Science and a Professor of Computer Science at Northwestern University (Interview Linden, 2015)

A third reason to automate news production is the idea put forward by Harvard professor Shoshana Zuboff that anything that can be automated, will be automated (1988). Among journalists, this has, of course, raised a heated discussion and concern over whether automation will take away jobs. Recent studies and practical experience show, however, that this fear is unfounded as news automation concentrates on writing stories that journalists cannot or do not necessarily have the time to write because the underlying data is too big or the audience too small – such as all local soccer games. So far, news automation has not replaced humans, and it appears that the human and automatic news-generation components in newsroom production tend to remain complementary.

"In our impression the very good writers – the journalists that write very well, that have really something special – they will have absolutely no problem."

- Project Manager Philippe Sordet, Project Manager at Swiss Tamedia (Immersive Automation, 2018)

Artificial Intelligence and Intelligence Augmentation

When talking about automated news, a crucial difference we need to understand is the distinction between artificial intelligence (AI) and intelligence augmentation (IA) – a divide that was described by co-founder and VP of Marketing and Strategy Harmon.ie's David Lavenda.

AI can be described as an autonomous system that can be taught to imitate or replace human cognitive functions. However, while AI will clearly play a larger role in our daily lives, it is by itself not a cure to all problems. This is because AI based solutions work best in structured environments where all relevant information can be considered and where the goals of the system are clearly defined. (Lavenda, 2016.)

This is where IA steps in, and advanced news automation is a perfect example of this. According to Lavenda there are many business processes where the human will remain in the driver's seat for years to come. Thus, AI and IA are not in a war but rather supporting one another, and they both have a key role to play in our future. (Lavenda, 2016.)

1.2. News automation in practice

To be able to perform assigned tasks, the news automation system needs a set of rules. These decision-making rules are called algorithms, and they are designed to carry out a specific task.

When talking about automation, algorithms, data and news, we need to remind ourselves that data and information are not the same thing. As Celeste Lecompte (2015), vice president of business development at ProPublica, puts it "algorithmic content creation is not just about turning a spreadsheet of numbers into a string of descriptive sentences; it is about summarising that data for a particular purpose".

For an algorithm to work and write news articles it needs accurate data about a specific topic and guiding principles on how to write. For example, to be able to write a financial story, the system needs up-to-date and accurate data about a company's quarterly earnings, as well as narrative know-how – how to tell an earnings story. (Lecompte, 2015.)

In order to transform even the simplest data set into a meaningful and interesting news story, the developers and journalists must convert the loose and roughly-followed guidelines of human reporters into strict rules for the computer. Alternatively, machine learning can mean the system itself forms a hypothesis of how it is expected to behave based on training examples along the lines of "if you get this input, you are expected to produce this story". Automatically generated models can easily make subtle but significant mistakes, however, so for right now there is a lot of human input involved in shaping algorithms.

A human with statistical analysis tools could do the same as an automated system but a job a reporter might need days for what would probably take a matter of seconds for an automated system.

Importing, analysing and evaluating the data is only a part of the process, however. After that, the system has to produce a readable, understandable and newsworthy article based on the information.

Translating the factual content of the story to natural language text is called Natural Language Generation (NLG) and NLG systems work in different ways. At its simplest, a program that adds a name of a student to a diploma could be seen as an NLG system: it contains a template of the body of the diploma and knows to add the name of the student, a variable, in a certain position. In practice, such trivial systems are only useful in highly limited settings and most NLG systems are much more complex.

Even though a news automation system can produce more articles than a human ever could, it can also make mistakes due to faults in the algorithm or inaccurate data. Again, the biggest difference compared to a human reporter is that once a mistake is noticed, it can be corrected so that it never happens again. There might also be context that the system cannot understand, for instance, the referee accepting an offside goal that changes the nature of a soccer game.

1.3. Use and potential of news automation

Suppliers tell us that media companies are way behind in the race towards augmented intelligence and, instead of investing they are holding back.

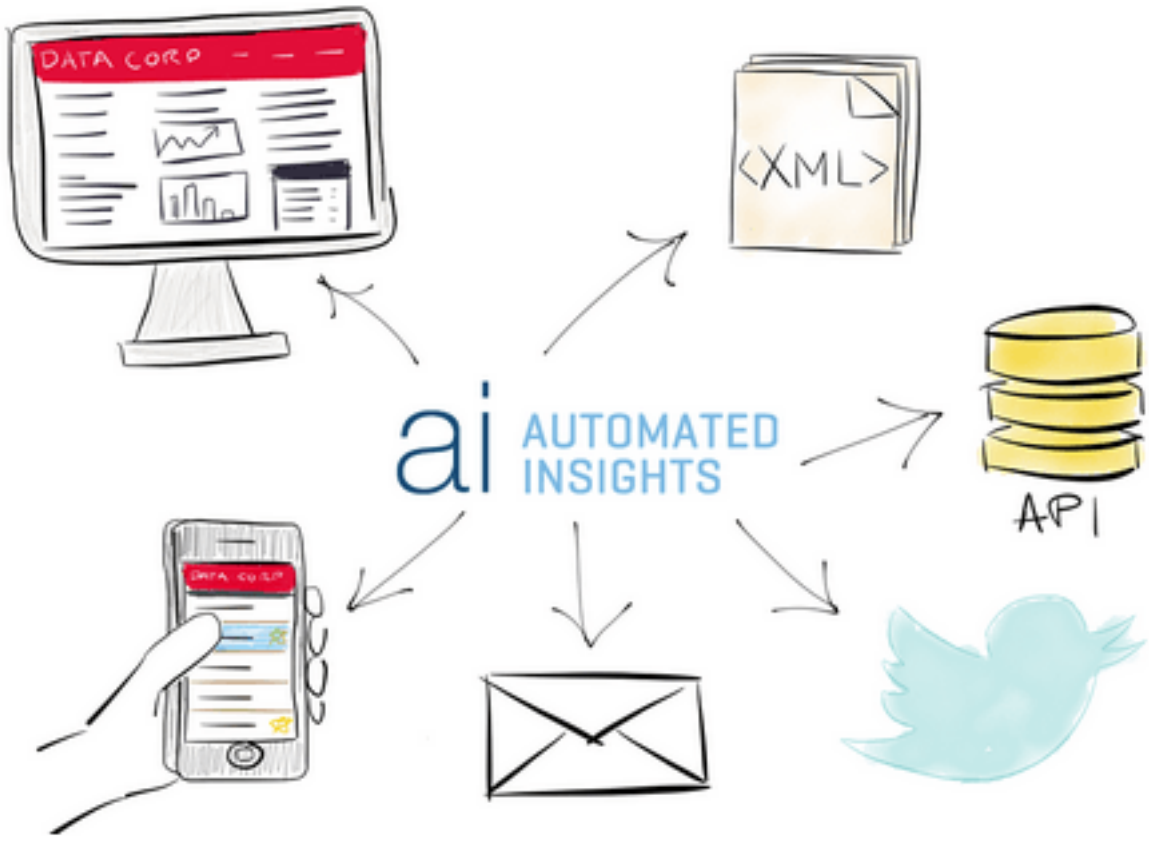
Media outlets who have never worked with news automation seem to be more than happy to dismiss its potential (Lindén, 2017). This scepticism is strange, considering the number of functions that have already been digitised in any given newsroom, beginning with word processing and photo editing.

Right now there are no signs of automation directly leading to journalist job cuts. The Wall Street Journal's Francesco Marconi, who previously worked at the AP, states something that should be obvious to us all: not all journalistic work can be automated. Yet, for the media industry it makes sense to explore new opportunities offered by NLG.

Neither automation nor computer anxiety are new in journalism (Akst, 2013). There are potential pitfalls, but they should be kept separate from the debate concerning lost jobs which is where the metaphor of "robot journalism" is so damaging. In chapter 5 we will discuss in detail what newsrooms should take into consideration before experimenting with news automation, and why it is not always a meaningful solution.

"Artificial intelligence can enable journalists to analyse data; identify patterns, trends and actionable insights from multiple sources; see things that the naked eye cannot see; turn data and spoken words into text; text into audio and video; understand sentiment; analyse scenes for objects, faces, text or colours – and more."

- Emilie Kodjo, Director of Communications and Public Affairs at The Global Editors Network


DATA CORP
<XML>
ai
AUTOMATED
INSIGHTS
API

Chapter 2

2. News automation around the world

In this part of the report we present five examples of how news automation has been implemented around the world. We also include the system we built in our own R&D project.

The Associated Press news agency in New York, along with the software company Automated Insights, based in Durham, North Carolina, have created a system that automatically generates earnings reports for AP's customers. It is a simple system, relieving financial reporters from the tedious work of digging through financial reports, and getting totally bogged down by the "earnings season" that lasts for several weeks four times a year. The output has gone up from around 300 stories from each earnings season to 4,400 at the latest count. Based on the positive experiences Francesco Marconi, while at AP, wrote a guide to automation in 2017 for those who want to follow progress in this area.

However, extensive research interviews reveal that the progress in the field is still slow and while other agencies are experimenting with automation, AP remains one of a few isolated success stories. In the final chapter we examine the implications of this hesitance.

2.1. Sweden: MittMedia and United Robots

United Robots is a Swedish service provider partly owned by the media company MittMedia, which publishes almost 30 newspapers around the country. United Robots has developed its own NLG system for news generation, Rosalinda (named after Pippi Longstocking's parrot) and has for several years been producing news reports on most team sports played in Sweden.

For sports data, MittMedia and United Robots rely on Everysport Media Group, a closely connected company that provides the input for news. Everysport employees actually call team leaders and referees by phone after each game or collect the match facts online if possible in order to get scores from all matches played in Sweden on all levels. This is a manual, laborious and tedious way of creating your own data, but since most lower division sports organisations lack a common reporting system, this is the only feasible solution.

Recently, the companies have introduced a chat bot that collects usable quotes from team leaders. MittMedia and United Robots are constantly scanning open data in search of new opportunities. In early 2018 United Robots and MittMedia launched a system that writes news articles about bankrupted companies. In 2017, they also started to produce automated news about property deals, a service that is popular and one of the main news items converting free readers to subscribers (Govik, 2018).

"[People's] needs have to be the guidance," said Robin Govik, CDO at MittMedia. "You start in the wrong end if you look for data that is easy to manage [...] We saw a need around sports, a need around weather and traffic. There is great interest. What we see now is a need around home sales."

According to Sören Karlsson, CEO of United Robots, there is an ongoing challenge for such an effort due to the lack of structured data. Even counties or municipalities do not use the same IT-systems, which makes it difficult to achieve a centralised, scalable solution.

MittMedia has a data strategy manager and its own data management platform, Soldr, a digital ecosystem that enables product development. Soldr collects, compiles and combines three different types of data – user data, event data and content data (Sundgren, 2017).

"We have 480 football teams in 59 leagues playing every week. And that is just soccer. Now we cover them all. We publish 3000 automated texts a month."

- Robin Govik, Chief Digital Officer at MittMedia (interview with Tuulonen, 2018)

8 tips from United Robots

Sören Karlsson, CEO at the Swedish NLG service provider United Robots, shared eight takeaways from working on a third-party system. This is a summary of his thoughts (Lindén & Karlsson, 2018).

1

Support from the top

The editorial management team must engage in the project, show the editors that this is important, and that they believe in the project. In the Swedish newsrooms where implementation has been most successful, managers have been perfectly clear that this should be done, and argued for the project. This point applies to all change projects in an organisation.

"I have also seen the opposite, where top executives keep away and the poor news or sports manager has to take the responsibility, and the newsroom is able to sink the project."

2

Engaging ad and marketing departments

Introducing automated content is an excellent opportunity for the newsroom to do business and product development together with other departments. Questions to ask are: Can we produce a brand new vertical site? A superb local offer? A new content category that attracts a particular group of advertisers? Content that would convert visitors to paying readers, such as the automated real estate texts produced by MittMedia?.

3

Traditional journalistic virtues apply

You get local content, faster publishication and a large number of texts. These qualities are considered strengths even when journalists write the news. In other words, automated content is good local content.

4

Use automatic texts as news tips

Data analysis is an important part of automated processes, and algorithms are much better equipped to find hidden relationships, outliers and so on, than people. Ensure that journalists and editors are alerted by automated systems when something interesting has happened, for example, when the most expensive house on the market is sold, or when a nobody scores a hat-trick in the 6th division.

Regard the texts as ready for publishing

If you want to add human creativity to a text, you should be able to do so. When the texts are written well enough to be published directly, it should be done to maximise their potential.

5

Increase the volume and take advantage of the speed

The ability to produce very large amounts of texts in a short time is one of the great strengths of automation. Think about how your distribution systems and platforms can handle the vast amount of texts produced when automating, for example, all real estate sales or soccer games. Special sites, personalisation, local apps and push notifications are examples of different solutions.

6

Review the organisation

You do not necessarily need to reduce staff, but maybe review schedules and tasks. Questions to ask are: Do we need as many staff in the morning if the texts about all sports games are automatically written? Do we need as many, or the same type, of freelancers as we do now? Can we do anything other than routine reporting, which would add more value?

7

Think about the notion of news value

The news value of a print product that fits the audience, published once a day, can not be compared to the news value of a local digital flow that produces push notifications around the clock. Over the past centuries, the structure of news and the work processes behind them haven't changed much, and the news values and ways of presenting a news event have followed similar criteria. When the approach to writing and publishing changes, these news criteria are put to the test. Instead of finding an angle that everyone finds interesting, the same news story can have several angles depending on who is reading the article.

8

2.2. UK: RADAR

RADAR (Reporters and Data and Robots) is a local news agency formed by a partnership between Urbs Media and the British news agency Press Association. The joint venture company is using human-authored journalism and automation to produce a daily diet of data-driven local news stories for publishers across the UK. The development of this new service has been funded by Google through the Digital News Initiative.

The NLG system is based on an NLG software tool from the company Arria, and the input is derived from public open data, such as the London Datastore. Journalists at Urbs Media write text templates and the system chooses the data for the location of the news outlet, which means that each template can be used for hundreds of different stories.

Thanks to the use of a market-available product, RADAR started production of stories for a pilot group of around 40 newspaper titles that turned into a full, UK-wide trial. In early 2019, RADAR became a subscription-based local news agency mainly serving large- and medium-size local news publishers.

In addition, RADAR has developed two further technical aspects.

The first part is a way of storing a copy of the original data so that the source can be traced back from the story. After first relying on third-party suppliers, RADAR has developed its own management system for the storage and distribution of open data. The system has a UI which allows it to initiate a process of sending a dataset to an NLG template.

The second part is a distribution system, which can match each version of a story that comes out of the NLG to the correct end user. This needs to be done for most common geographical granularities. The basic unit is a UK local authority, a local government area. Content is distributed via 391 local channels covering the UK. Customers can download the content from the RADAR portal or distribute it through their own

"We wanted to build a collection of tools, a journalist workbench, to enable our reporters. This includes an archive storage area to retain a local record of all original source data, NLG software as a writing tool, and a smart distribution system that can deliver each local version of the stories we produce to the right news publisher."

- Gary Rogers, Co-founder of Urbs Media, Editor-in Chief of RADAR AI (interview with Lindén, 2018)

"We looked at how many data sets were available in mainly open data sources, how many of them were interesting and how many of them were localised, and from that we saw that we could probably get to five stories a day."

- Gary Rogers, Co-founder of Urbs Media, Editor-in Chief of RADAR AI (interview with Lindén, 2018)

systems via API. A user will choose the areas they are interested in and will be able to access all stories relating to it. Gary Rogers, Co-founder of Urbs Media, Editor-in Chief of RADAR AI, explains:

"The idea is to eliminate any unnecessary work and so free up time for the end client, the local newspaper, while at the same time providing good, localised stories. This means more comprehensive local coverage for the paper while releasing reporters to work on their own leads."

The ambition is to be able to create a handful of stories a day that can then be "duplicated" for 200-400 localised versions each.

For initial experiments, RADAR used automation on a couple of datasets from the London Datastore. Now RADAR uses national datasets for delivering a UK-wide service. The biggest source of data is the country's national health service, the NHS. The representative says that it puts out an enormous amount of data about everything, from individual hospital consultations to staff members.

2.3. USA: The Washington Post

For the 2016 Rio Olympics the Washington Post developed Heliograf, an automated storytelling technology, in-house. It is an NLG system that automatically generated short multi-sentence updates for readers. These updates appeared in The Post's live blog, on Twitter (@WPOlympicsbot), and were also accessible via The Post's Olympics on Alexa-enabled devices and The Post's bot for Messenger (The Washington Post, 2016).

After that, The Post broadened its service to areas with a lot of data, such as election results, crime, real estate or earnings announcements. Output remains modest in its first year, The Post has produced around 850 articles using Heliograf. Shailesh Prakash, Chief Product and Chief Information Officer, sees other limitations as well. "My goal is not to replace journalists. ... I do not think that engineering has advanced that far to write an opinion piece or to write a deep analysis of what is going on," he said.

Heliograf is a flexible system and publishes to any channel, including Twitter and Alexa on Amazon Echo.

Prakash sees Heliograf as transforming the newsroom, greatly expanding the breadth of coverage and allowing journalists to focus more on in-depth reporting. In his view, journalists can concentrate on writing about the big stories while the NLG system handles daily news coverage.

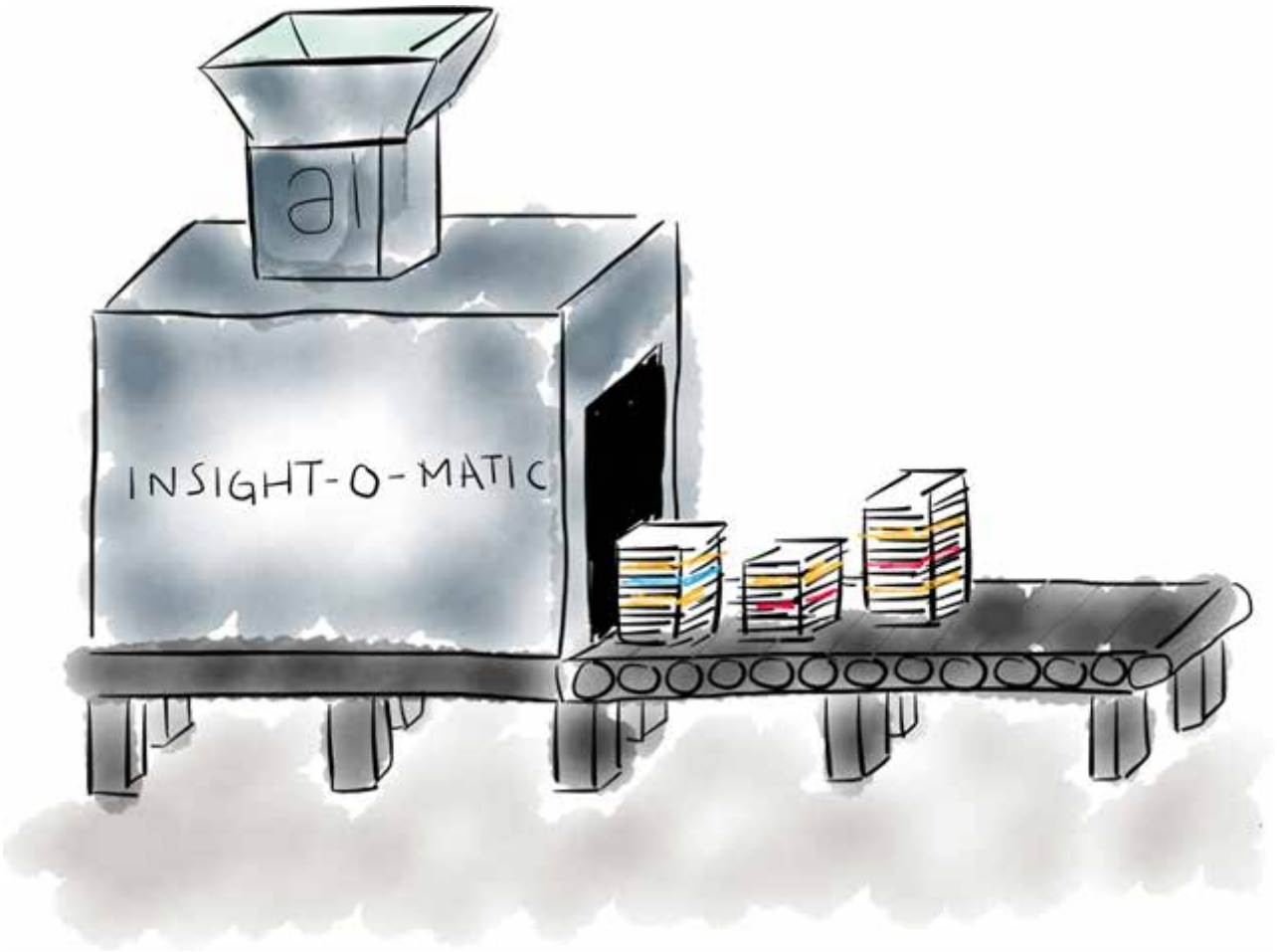
For the Washington Post news automation is a new part of the business model. By October, 2017 the newspaper was approaching 100 million unique visitors a month of US-only audience, in other words, almost a third of the population go to the Post to get news and information. With international visitors, the total number was about 150 million a month.

This long tail of content also drives usage, and ultimately advertising and subscriber revenue.

Recently, the Washington Post has taken on the goal of putting out results of every high school football game in the D.C. region. Each game story will draw from scoring plays, individual player statistics and quarterly score changes, along with The Post's own weekly Top 20 regional rankings. (The Washington Post, 2017.)

"In today's world, while content is still king, it is equally important to get your product and technology right. That is why the marriage of technology and journalism is so important."

- Shailesh Prakash, - Shailesh Prakash, Chief Product and Chief Information Officer at The Washington Post (Prakash, 2017)


ai
INSIGHT-O-MATIC

2.4. Finland: Valtteri

In April 2017, the Immersive Automation (www.immersiveautomation.com) team launched a tri-lingual news bot that automatically generated news articles about the municipal elections of Finland in three languages – Finnish, Swedish and English. The election news bot Valtteri (www.vaalibotti.fi) based its news generation on traditional journalistic news values, and an open data set produced by the Finnish Ministry of Justice. Unusually the bot autonomously decides what to report and how. Its selections were data-driven, with no predefined story structures.

Not only that, the bot can be personalised by users. Readers used it to search and find individualised news in terms of geographical area, party and candidate of interest; its value for "translating" vast numerical data to formats understandable for the public; and its general capability of being transparent in its decision-making rather than opaque.

Valtteri produced more than two million news stories across all three languages. A practical test showed a seasoned journalist took one hour to write a single comparable story. Meaning that it would need a thousand 40 hour working weeks for a human to produce the same volume.

This comparison is not entirely fair, though, since the cost of the first article is significantly higher for a machine than for a human. Also, the cost of the first system is significantly higher than the cost of the second system which can reuse a huge chunk of the engineering effort made for the original system. Estimating programmer effort and time is a classical unsolved problem in programming.

The election bot provided evidence of the potential of news automation by (1) converting vast data sets into user-friendly formats, (2) freeing up journalists' time for more creative tasks, and (3) serving readers' specific interests and needs. At the same time, the independent "ecosystem" it creates – consisting of data, software, stories, journalists and consumers – also necessitates discussion on issues such as transparency and accountability that reach far beyond the walls of the newsroom. On the technical side, Valtteri provided evidence towards the feasibility of repurposable and multilingual systems for automated journalism.

After the initial focus on election news, Valtteri was repurposed to work with data on crime statistics available at Statistics of Finland. We explain the system in more detail in subheading 3.2.

About the same time, both the Finnish division of the Nordic media group Sanoma and the public service broadcasting company Yle trialled their own news automation systems. While the publicly available details on Sanoma's system are scarce, YLE eventually released their system as an open source project, "avoin-voitto," on GitHub (Yle, 2018).

2.5. China: Xinhua and Caixin

Little of this is known outside China, but many large Chinese media organisations are already involved in news automation: Tencent has a news writer called Dreamwriter, Alibaba's news automation system is named Writing Master, Toutiao and Xinhua News Agency produce automated news. All these systems are mainly used for financial, weather and sports news with given algorithms.

At the Chinese state media organisation, Xinhua collects everyday information from official websites such as data and weather, and also pays for data from providers such as the IOC.

Zhongxuan Dai – a doctoral student at the Hong Kong Baptist University who has done research on Chinese news automation – explains that at Xinhua, the in-house developed system starts with the journalist who sets the template and a programmer translates it into an algorithm. When a news article is finished the system does not publish directly but instead waits for a journalist to proofread (Tuulonen & Dai, 2018.).

Another Chinese media outlet, Caixin, opts for a third-party solution.

Caixin gets data from intelligence service and data provider Caixin Insight which receives it directly from the markets. Although mostly focussing on stock markets, it can report on sports and the aim is to extend coverage to debt and other markets.

"Our automation application is based on HanLP models and algorithms, which specialise in processing Chinese. With them we also expand our own text corpus and dictionaries."

- Jiapeng Wang, Senior Operation Director at Caixin (interview with Tuulonen, 2018)

Chapter 3

3. Technical solution: Natural Language Generation

NLG is a subfield of AI in which computers use data to produce text, often called narratives. NLG – from data to text – is the opposite to natural language understanding – from text or audio to data.

A brief history

NLG has been used in news for decades primarily for well-understood domains such as weather reports. Only recently have we seen new applications in the fields of finance (Nesterenko, 2016; Mendez-Nunez and Trivino, 2010; Andersen et al., 1992), sports (Bouayad-Agha et al., 2012; Theune et al., 2001; United Robots, 2017), earthquake reporting (Los Angeles Times, 2007) and elections (Leppänen et al., 2017a, 2017b). In addition, NLG has also been applied outside of journalism to produce, for example, medical summaries (Portet et al., 2009), public service announcements (Reiter et al., 2003), product descriptions and advertisements.

Whilst a large amount of research has resulted in the development of architectures and models for tackling the tasks involved and there are indeed proven applications – such as weather reporting – the technologies remain poorly exploited in the context of journalism.

One reason for this is the complexity of the natural language used in journalism. Easy-to-implement NLG approaches are only applicable in settings where the range of possible news stories is well understood and relatively limited, for example sports results.

Applying NLG to more complex settings is expensive, which leaves journalism in a difficult position with simple methods not being up to the job and more complex bespoke approaches proving prohibitively expensive. Currently there are few ready-made tools publicly available to bridge that gap.

3.1. One data set, hundreds of thousands of articles

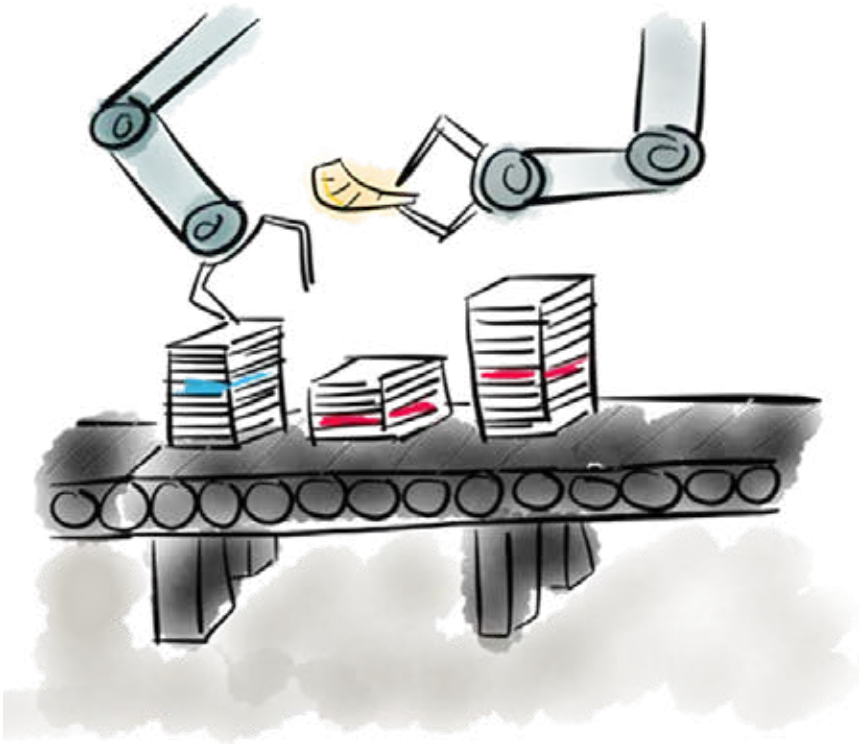
Immersive Automation's R&D projects have attempted to develop NLG systems that take advantage of these breakthroughs.

We have chosen to focus on applications where a considerable amount of quantitative data exists, allowing us to generate a large number of articles from a single dataset. This allows highly flexible content and structure in response to a user's interests. This contrasts with using fixed article structures where most of the content is crafted by a human journalist and small details vary according to the input data.

Whilst this alternative approach results in a system that reliably produces readable articles with high quality output, it greatly limits the scope for exploring the potential application of NLG techniques like content selection and ordering.

Unlike systems that use a fixed, handwritten template for a whole document, our system builds a document up from a set of facts that are automatically determined to be relevant and interesting based on user queries. Having selected these facts, it formulates each one as a sentence, and organises them into a coherent document structure with multiple paragraphs.

The system notes the subject of the article which helps avoid time and location errors as well as ensuring correct use of pronouns. It's an architecture, drawn from common NLG standards (Reiter & Dale, 2000).



3.2. One system can do one thing

Our system was designed to cover elections. More specifically, it produced reports of the 2017 Finnish municipal elections. Finnish municipalities have a significant degree of autonomy and are responsible for many decisions that have a direct effect on daily lives so it is a significant media event. The results of the municipal elections also act as a gauge of the national political landscape.

The system we designed is modular, allowing problems to be addressed and enhancements to be added at the appropriate point in the process. The modular structure also allowed us to make the system multilingual, building each of the relevant components of the pipeline architecture so as to take into account the language of the text being produced, while sharing much of the process between the different languages: Finnish, Swedish and English.

To produce sentences or fragments of sentences, and to express individual facts, the system draws on a collection of short hand-written templates such as "[party/candidate] got [number] votes in [place]", where the bracketed sections are replaced by other words or numbers when the template is realized into natural language.

The system was able to produce a large number of articles based on the data with relatively little manual effort. Once the system was producing Finnish language articles it took an experienced person (one familiar with the system) less than half a working day to add basic support for English-language articles. Languages then weren't a problem. But in order to adapt the system to produce articles about another type of data, new templates need to be written each time.

An alternative, still only attempted in non-production research prototypes, would be to extract templates from existing, human-authored news articles covering the domain in question where the underlying data is known or can be inferred. Although these templates are likely to include incorrect or incoherent examples they could be used as drafts or suggestions for the template authors to speed up new template creation.

It should be noted that while the system as a whole is tied to the single context – in this case, municipal elections – most of the individual components can be reused with relative ease. Through careful division of labour between the components, it is possible to ensure that most components are specific to either only the domain, the language or neither. This allows for the transfer of the software to another domain with minimal re-engineering.

3.3. The problem of discrepancies

Templates are designed by humans and reflect human decision-making. The looser the template, the greater the chance for discrepancies. For example, if you use Automated Insights' Wordsmith tool to fill in a ready-made story with data, it will not diverge much from what you want it to do, but it will also need a variety of templates to avoid it being boring.

Among the more significant problems associated with applying flexible automated journalism systems is that the produced articles can end up producing fact-belief discrepancies in the readers. That is, even if the content of the generated articles is technically objectively true, the human reader may draw conclusions leading ultimately to an objectively false impression. Such cases should be seen as failures of the system, rather than the reader, similar to the way we view ambiguous language written by human authors as something the author needs to address.

For example, consider the claim "Party X got the most seats in the parliament" in a context where party X was in a tied first position. While the claim is objectively true, it is also highly misleading. For the reader the natural inference is that party X is unique in this respect, which is not true. Ensuring that no fact-claim discrepancies arise becomes a significant problem as the complexity of the generated expressions increases.

Beyond fact-belief discrepancies, NLG systems can also produce fact-claim discrepancies. These can also be called "incorrect statements" or simply "lies". The underlying problem is that it is possible to take multiple true statements and accidentally produce language that is objectively false in the sense in which most people read it.

As a concrete example, our system produced a sentence "Party X got 152 votes and 3 seats in the previous elections". The natural reading of this sentence is that both the number of votes and the number of seats pertain to the previous election. However, the system was trying to describe a situation where the votes are from this election, and the number of seats are from the previous. In other words, the sentence should have read "Party X got 152 votes in these elections and got 3 seats in the previous elections." Ensuring that the output is fluent while avoiding confusion is difficult.

Simpler systems with less fluent and flexible NLG often end up avoiding these issues by increasing the amount of control the journalists and programmers have, but this is a trade-off between fluency, correctness and development effort. In cases where there may be an overwhelming need for absolute and verifiable accuracy, such systems can however be paired with editors for better human oversight (Diakopoulos, 2019).

Chapter 4

4. User acceptance and experiences from media outlets

Automated articles pose a variety of opportunities and challenges. This chapter presents findings from stakeholders in the Immersive Automation project including economic perspectives and how audiences and media outlets perceived automated content.

4.1. User perceptions

Over the last five years there has been a lot of research into how people perceive automatically generated news. Besides general quality and credibility assessments, one of the key issues has been what the impact is on readers when they are told the story was written automatically or by a human.

As we mentioned already, news articles that rely heavily on numbers and statistics and relate to known recurring events, lend themselves well to automation. The repetitiveness of the events supports ordering facts and building templates accordingly. Reports from sport events and finance are therefore classic cases for automation and are the most frequently evaluated as a result.

There are many ways of testing user perception, but Sundar (1999) has served as inspiration for many test designs. Sundar extracted 21 measures and grouped

the related measures using factor analysis into four common factors. These factors are credibility, liking, quality and representativeness – each semantically associated with a set of key words as follows:

- Credibility: biased, fair, objective
- Liking: not boring, enjoyable, interesting, lively, pleasing
- Quality: clear, coherent, comprehensive, concise, well-written
- Representativeness: important, relevant, timely

Figure 1 gives an example of an evaluation tool developed and used in the Immersive Automation project.

Articles were evaluated according to the four factors listed above. Additionally, test users could give free text comments as to what was good and bad in the article and indicate whether they thought it was written by a journalist or created automatically.

When looking at the results, males aged 55 or more liked the automatic articles best and females aged 34 or less liked them the least. Evaluators mistook 21 percent of the computer-written articles as written by humans and 10 percent of the human-written articles as computer-written. The share of users making these mistakes grew with their age. More information about the tools and results obtained can be seen in Melin et al. (2018).

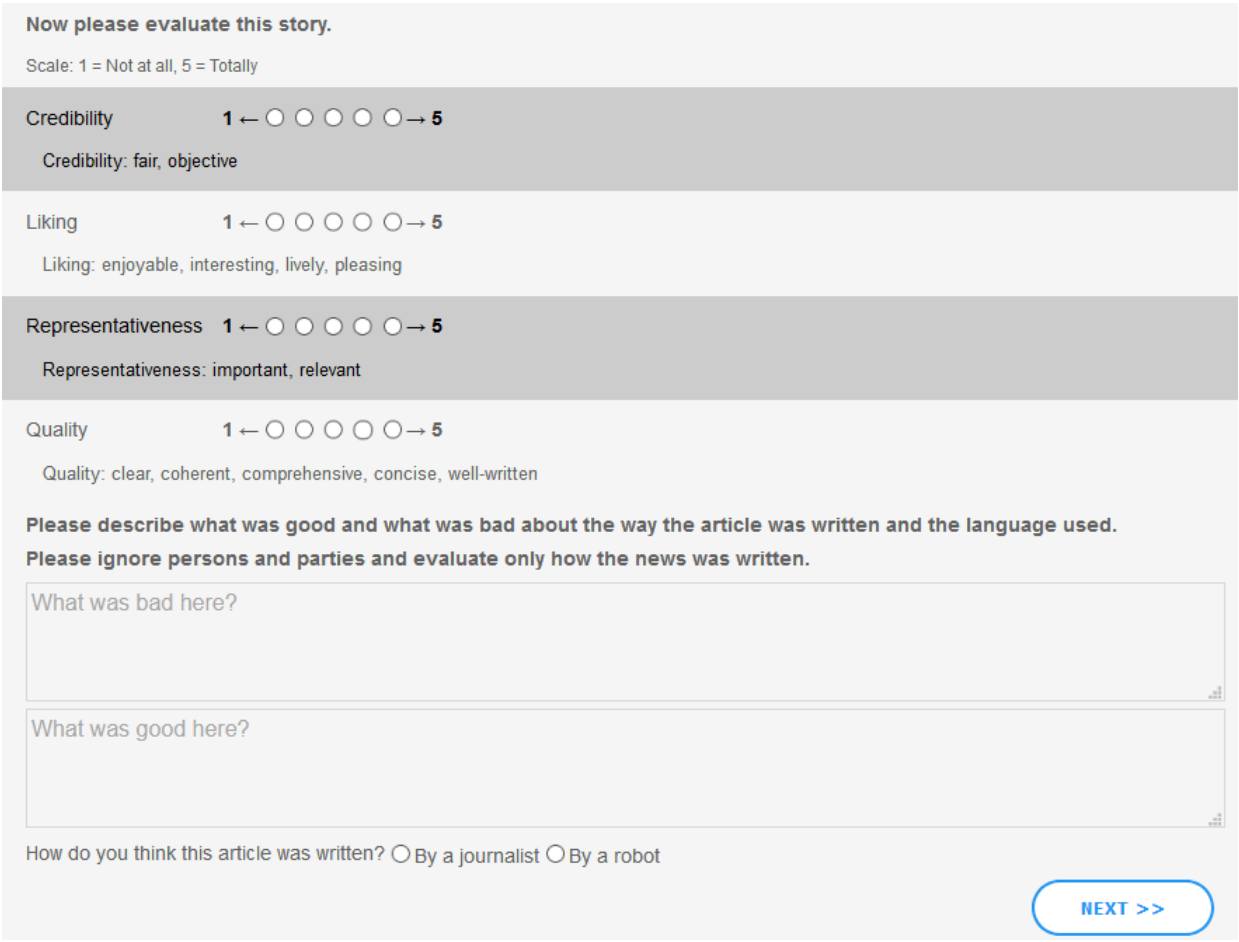


Figure 1. Screenshot from the Immersive Automation evaluation tool, allowing user perceptions of election news written by a computer to be compared.

What we see is that automatically generated articles compete well in credibility with human-written content. The by-line information does not seem to have any significant impact on the perceived credibility (van der Kaa & Krahmer, 2014), at least not in the European context. In some cases (Graefe et al., 2016; Wölker & Powell, 2018), the automated content even got somewhat higher credibility ratings than the articles written by humans, especially in sports news (Wölker & Powell, 2018) – but, in general, the credibility of automated news can be considered to be on par with human-written news. Of course, we have not yet seen detectable "outbreaks" of automated fake-news, which could severely affect the credibility of this type of content if the automation is detectable.

Regarding the readability (Graefe et al., 2016) and pleasant-to-read metrics (Clerwall, 2014) computer-written stories were always rated lower than the human-written stories.

For media companies, the credibility of their stories is essential. However, credibility is far from the only factor to consider. Stories must be exciting enough for consumers to choose to read them. The importance of credibility varies between topics, and for some areas credibility and timeliness alone may be all that count. Therefore, in a newspaper context, user acceptance testing should be made in a real application context, because how the news is delivered to the end user plays an important role in the total experience in addition to the quality of the language of the actual news stories.

Looking at automated news from another perspective, it may be that fact and angle selections are more important traits than producing grammatically perfect output. You can often transform dry sounding numbers into invigorating stories if you can only find the relevant connection to the reader. That connection may be places or people of interest, statistical trends and abnormalities, or relation to popular trivia – to name but a few.

4.2. The cultural issue: Case study of South Korea

While automated systems may seem creepy (Mori, 1970), people's reactions to robots can be influenced. This is called "the Hollywood robot syndrome" (Sundar, Waddell, & Jung, 2016).

For instance, South Korea has massively invested in robotics and automation during the last two decades and put a positive public spin on it, which might affect the way people perceive machines and feel trust towards automated systems. Automation bias is a well-researched cognitive tendency that leads to systematic deviations from a standard of rationality and good judgement. There is a related cognitive tendency, pro-innovation bias, that might as well have effects on the outcome.

Research on the reception of news written by machines shows (Jung et al, 2017) that the cultural context is crucial. In South Korea, tests indicate that people show more faith in news produced automatically than generated by human journalists, with some

results indicating that the reason is that journalists and news media are linked to corruption. Trust in news media is exceptionally low in South Korea (Reuters, 2017); the low credibility of journalists may be one reason people preferred texts written by machines (Jung et al, 2017). There seems to be cause for self-reflection among Korean journalists themselves if automation is seen as the key to improving credibility in news. In fact, journalists themselves gave higher scores to texts written by computers as well.

Cultural differences between countries can play a role in the perception and trust of automated systems. South Korea is a country with a collectivist cultural orientation, and so it will be necessary and valuable to compare differences in reception with other countries that are more individualistically inclined. Whether low trust in news and a higher trust in automated content might also be correlated in other countries is a topic for further research.

4.3. Media outlets' views

When talking about journalism and its change, technology is often seen as an autonomous agent which develops according to its own inner logic causing this or that in journalism. But as news automation has already made its way into newsrooms, its effects are starting to show in journalism and journalists' everyday work. For research conducted with Finnish and Swedish journalists who have been working with news automation, the consequences are two-fold.

- Firstly, **the journalists say news automation increases the overall value of** media outlets', newsrooms' and journalists' work since it provides the audience with completely new angles and information. This means, for example, that it can help journalists in finding stories in data that would otherwise have gone unnoticed.
- Secondly, **news automation can also produce articles that media outlets previously bought from outside** – articles such as weather reports or sports game stories from freelancers. It also does a lot more work than has been previously possible with existing resources. So we can see how an algorithm can add value to what journalists and media outlets are already doing. However, it may not help significantly to maintaining and improving the brand of the media outlet. (Tuulonen, 2017.)

4.4. Observations from the Steering Group of the Immersive Automation project

Members of the Steering Group (13 persons who work as product managers, development managers and news editors) of the Immersive Automation project and three key persons in the Finnish Broadcasting Company (Yle) were interviewed with the aim of charting their understandings of news automation developments in a broader societal perspective: social participation, developments in the ecosystem, and in the practices of newsrooms.

- On the positive side, **respondents generally find that automated news production increases the fact density of news, reduces ambiguous news, and boosts the availability of targeted (niche) news that interests small audiences**. Detection of trends that appear in big data, including interesting irregularities, were also mentioned.

- **Frequently mentioned risks relate to tacit power embedded in data-driven routines, the division of customers/audiences into minute groups, commoditised content lacking a personal touch; and negative effects on language.** To counter these risks, transparency of the processes (including the algorithms) is required.

- **Several actors on the commercial media side saw benefits in B2B applications**. Differences in focus could be observed between the four Finnish media organisations (Alma Media, Sanoma, Keskipohjanmaa and Yle) represented. The main priority areas varied from production process development (text and graphics, including narrative-structure solutions) to the development of solutions based on audio and interactive-home solutions, and local media content.

- The public broadcaster sees its role partly as a competitor, but also as a co-developer in co-operation with commercial media although there is a question about whether that's how it's seen by its smaller competitors. **Representatives of the public broadcaster also looked at automation in light of how public services can be offered more generally** (e.g., national 'passwords', automated participation). The broader societal perspective of a 'digitalised society' rose to the fore in interviews with researchers on the academic side, whereas representatives of the industry tended to look at the development from the point of view of how their competitiveness can be enhanced.

- A noteworthy observation was that **automated information processes as a whole may become a "fifth estate" that guide and guard the performances of the fourth (media)**.

- A general view arising from the interviews is that **the media will lose out to other actors if this field is not rapidly developed and implemented**. Short-term changes (five years) will be relatively small: mainly the substitution of routines and provision of material for tedious journalistic tasks. In the short term, automation will not endanger jobs; however, skill requirements will change.

- **When the respondents were asked to define qualifications for a new hire for their team, multitasking was given priority**; either in the form of journalist/data competences in the same person or in complementary hiring where skills in the team were lacking. Few wanted to recruit 'more of the same'. A risk factor mentioned in this context was whether those responsible for hiring have the nerve to employ "those better than themselves".

- Bigger changes were foreseen when machine learning comes into broader use. Visions for this development were still only vague, but the shared insight is that it is only a matter of time, and that this will mould the field in dramatic ways. **The time span for this to happen is, however, estimated to be more than five years.**

- A main concern mentioned was "missing the train". The interviews only partially confirmed earlier observations of newsrooms being conservative and reluctant in their attitudes to automated processes in journalism; **such attitudes have now been substituted by an awareness of the "competence gap"**.

4.5. Proactive or reactive business models

Although there is a general awareness of how the media business has developed from a mass media market towards first a niche market and then further towards a personalised market (Picard, 2010), this has still made little impact on actual business solutions in the legacy media sector. The mass media sector is characterised by conservative investment strategies. In general, legacy media have developed reactive, not proactive strategies to defending their business rather than expanding it.

While print was the dominant marketing platform for advertisers in the Nordic countries 20 years ago, this picture is now reversed, with electronic media taking the lion's share of advertising revenues. Legacy media have lost considerable market share to digital newcomers.

For commercial legacy media the business vision remains quite bleak. In the aforementioned interviews made with representatives of the media in the Steering Group of the Immersive Automation project, there appears to be a lack of broader vision on how digitalisation changes the demarcation line between media and other functions in people's digital life. Media people still tend to see their business as separate from an overarching digital convergence that is changing the entire society. Thus, main focuses were given to features that maintain, rather than change the media sector. These included the protection and strengthening of brands; credibility; interface with target customers/audiences; content production capacity and originality. Other factors being focussed on were advertising strategies, the careful introduction of automation so as not to endanger the traditional product, the development of stable customer/audience relations, and co-operation with data providers. Other concerns mentioned included more active strategies aimed at better presence in the mobile sphere, development of (audio) services for home-based use, and archive re-use. A need for new forms of digital business in an environment characterised by information verflow was clearly felt, but remains poorly defined.

Chapter 5

5. The data issue

The five V's of data quality: volume, velocity, variety, value and veracity (Suthaharan, 2014; Wamba et al., 2015) will be elaborated on in this chapter.

The quality of output of an automated journalistic system is highly dependent on the availability and quality of the data that is fed into it. The common motto "garbage in, garbage out" applies here: no matter how great the automated journalistic system and its algorithms are, it cannot produce high quality news from bad data.

Consider, for example, an automated data source describing a city's public transport. News organisations already use systems that automatically generate news alerts about major public transport problems. But if the automated data source is updated only rarely, many problems are already solved, or irrelevant by the time the news gets out. Similarly, if a significant number of problems are not reported by the automated system, or are reported with incorrect details, customers are unlikely to trust the news alerts.

5.1. Volume and velocity

Volume refers to how much of the data is available in total and how quickly it is becoming available. Availability is affected by velocity which means how fast the data is processed. For an automated journalistic system to be a meaningful investment, data must have either high volume, high velocity or both.

An example of high-volume low-velocity data would be presidential election results: large amounts of data are made available on rare occasions. On the other hand, low-volume high-velocity data would be, for example, the temperature readings of a city. At any single time, very few temperature readings are made. But at the same time, these readings can be done extremely often, producing a steady stream of data. An example of high-volume high-velocity data would then be something like detailed meteorological information produced by all the weather stations in a country. Another good example of high-volume, high-velocity data is video streaming; a 4K video might require 25 Mbit/s

decoded by fast hardware in a computer or TV set.

From a business perspective, it is only meaningful to automate the production of a news type when it produces positive value. That is, when the cost of creating the system is less than the value of the news items created over the lifespan of the system. Examples of this are automated news on stock prices, which German business paper Handelsblatt produces purely with Google search in mind. Discovery of these small news items direct readers to the Handelsblatt home page. Another example is the property deals news bot at MittMedia that motivates free readers to pay for news.

In high-volume low-velocity situations, the automated system can quickly produce large amounts of news stories from each batch of incoming data. In the case of election results, an automated system can be used to near-instantly produce a tailored news article of the results for each candidate or geographical location.

In low-volume high-velocity situations, automated systems can either produce a constant stream of news reports or, alternatively, keep a small number of stories constantly updated. For example, a weather report can be kept up-to-date with real-time data coming in from a nearby weather station.

At the same time, if data is both low-volume and low-velocity, an automated journalistic system is only able to produce short stories with long intervals in between. Such systems are unlikely to be beneficial for a news organisation, unless the cost of setting them up is minimal.

5.2. Variety and value

In addition to volume and velocity, the variety of data must be reasonable. Variety refers to the complexity and the number of types of data.

At its simplest, data is structured and numerical or categorical, meaning that it consists of numbers or other known values, for each of which it is known exactly what each value refers to. Traditional databases are often examples of such data.

Examples of more complex types of data are natural language text documents, social media postings, sound files and video recordings. As the complexity of the data increases it becomes more difficult to make sense of it automatically. It is easy to produce an automated report from the stream of temperature values produced by a thermometer at the local weather station, but it is extremely difficult to automatically produce quality reports from audio-visual recordings of parliamentary debates. At the same time, more interesting things can potentially be found from complex data sources.

There must also be value in the data. In terms of journalistic efforts, the data must contain something worth writing about. Furthermore, the financial value must be greater than the cost associated with obtaining, cleaning and processing that data.

5.3. Veracity

All the above considerations are, however, meaningless if the data is not veracious. Veracity refers to the trustworthiness and accuracy of the data. Data produced by a broken weather sensor is worthless. While non veracious election results would have large news value from the perspective of reporting on election fraud, they are worthless for reporting on the actual results.

Automated reporting from unveracious data has the same consequences as humans writing news that are found to be untrue. As such, the veracity of data should be considered both as a business concern as well as an ethical concern of utmost concern to the editorial integrity of automated content.

5.4. Dig into data

All told, for a journalistic system to be useful, it must receive sufficient amounts of data – either in volume, velocity or both – that is of a reasonable variety and that is veracious. Which raises the question of how to get hold of such data?

One option in sensor data. Sensors are used to collect data from their environment. The data collected from the sensor is typically used either to detect trends, or to react to certain given events. The use case varies from temperature detection to determination of the best fish routes by locating fish shoals. In industrial usage connected sensors can produce vast amounts of raw data.

Another option is open data. In recent years, we have seen a significant increase in the amount of open data produced by public institutions and governments, including the U.S., the UK, France, Finland and Sweden. Open data is defined by its freeness, in both senses of the word. It is data that can be freely used, re-used and redistributed by anyone – subject only to the requirement to attribute and share. However, it is important to note that there is a lot of information available that can be freely accessed but cannot be freely reused such as figures and data from private business.

While open data is not standardised across different sources, the technical formats very rarely pose practical problems unless the data is either in some proprietary format – that is, a format that is owned by some commercial entity and the internals of which are not public – or the data is completely unstructured.

Unstructured data is data in a format that is difficult for a computer to process and interpret semantically, such as images, video and free text. This is contrary to structured data that refers to any data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets.

Making any sort of analysis on unstructured data is much more difficult than on structured data, and thus structured data should be sought whenever possible.

Chapter 6

6. Pragmatic considerations in implementing and deploying news automation

What should publishers consider when implementing news automation systems? This chapter gives general guidance on the implementation of news automation.

The biggest decision is probably whether the system should be bought "off the shelf", which leaves the media company at the mercy of the provider, or created and modified in-house. This question is related to the size of the media company and the resources at hand.

In addition to the implementation of news automation, ethical considerations and transparency must be taken into account. It is nothing unique to news automation, more a common issue when creating and designing software.

"The interesting thing is that no point really is about technology."

- Sören Karlsson, CEO of United Robots (interview with Lindén, 2018)

6.1. Implementation

For news media organisations, certain norms, routines and values are key features in maintaining both a journalistic and business identity. When faced with potentially radical changes, such as accessible data and news automation, a successful strategy incorporates a holistic view of how these new tools or products reflect – or deflect – the identity of the organisation. What data is used, and how the algorithms are designed, remain editorial decisions.

Questions to ask are how the norms and values of the news medium will be transferred to the automated systems, if the data will be allowed to "speak for itself", or if the process of automation is shaped according to predetermined needs, such as AP's automated financial reports.

For managers, automation heightens a demand for connecting and communicating between departments and skill-sets – or negotiating with external actors to ensure that services correspond to in-house values and processes. If changes in production are communicated poorly, or if the solutions proposed are based on assumptions, the implementation can run aground due to mistrust and misunderstandings. Software integration and adaptation of new technology is not a new challenge in newsrooms, but news automation can be perceived as an existential threat to jobs and professional identity. That needs to be addressed with carefully crafted communication strategies that reinforce the complementary nature of the technology.

New technologies always pose a risk for target blindness: Is the product you are planning born out of a data-first or audience-first approach? How will you utilise and distribute the material? Does your audience trust and understand the product or service you are offering? What are the financial or other benefits of automation for your organisation?

Automated processes require technological skills and competences that are often not found in news media organisations, and even if they are, implementing automated strategies requires the re-allocation of resources. In-house solutions can offer easier modification, independence and access to the systems whereas solutions offered from service providers can mean quicker "plug-and-play" implementation and support.

How AP did it

Tom Kent, President and CEO of Radio Free Europe/Radio Liberty, was one of the editors in charge of AP's automation project in 2014–2015. In a blog post from 2015, he shares some of the things editors should consider when testing and using automatic news writing (Kent, 2015):

- ## Data accuracy and rights to the data

  Is the data reliable? What does the underlying data consist of? Has it been properly transmitted or moulded by the provider? Does the provider have the legal right to send it to you? Do you have the further right to process and publish it, and if so, on what platforms? What happens if the background material needed by the algorithm changes or the data source suddenly becomes less reliable?

- ## Images and video

  Can you assure that the system is accessing images that you have a legal right to use? How can you avoid satirical or hateful images that are not in line with your standards, and how can you make sure the images and videos pertain to the actual event?

- ## What and how to automate

  What information will you compare? What data will the algorithm highlight? How will you match spellings, general writing style and capitalization with the rest of your content?

- ## Testing and maintenance

  How to test for errors? Do human editors check every story before publishing? Who maintains the data and reviews the choices that algorithms make? Who is watching the machine, how often and for how long?

- ## Disclosing how the automation has been done

  Will you tell readers that a story was automatically produced? How will you document the automation so that you are able to explain how every story came to be? Can you defend how the story was "written"? Will you be willing to reveal how your software works? Are you willing to share the source code or do you consider it proprietary information?

6.2. Ethical considerations and transparency

Working with big datasets required for automation means handling three dynamics: the technological, that is how to maximise the output from the data efficiently; the analytical, which refers to building a system that locates correct data and makes claims about it; and the mythological, in other words the belief that data in itself is inherently more neutral, objective or accurate (Lewis & Westlund, 2015; Boyd & Crawford, 2012).

The idea that data in itself is neutral has consequences for considerations around accountability and transparency. Currently, there are no specific standards for how automated content should be acknowledged. Some media organisations that use automation have opted to inform their readers about how the material has been produced, whilst others have chosen not to, or are relying on the trustworthiness of their brand. However, media organisations can benefit from transparency and explainability when competing for audiences. It can also clarify responsibility in potential cases of liability or error.

Transparency towards audiences supports trust, showcases technological innovation and limits potential future issues. Since the algorithmic systems producing the stories are often trade secrets and competitive assets, complete transparency of the systems is often not an option. There is also a risk in revealing too much, as outsiders could affect the stories produced by, for example, manipulating the source data (Myles, 2018). Nonetheless, as exemplified by Stuart Myles from the AP, there are levels of transparency, ranging from simple disclosure to what Myles calls justification, explainability and full reproduction of stories based on the data used. The question is to what extent the audience actually cares about how content is produced. Media companies that are providing explanations find that not many people click on those links, but this does not mean a carte blanche.

In most western societies it is a commonly shared value that editorial decisions should be made in newsrooms. Newsrooms should carefully consider how to mark a story completely created by automation as there are a number of options. Has it been automatically edited? Is the story based on an automated source text? If the text is automated and lacks a byline who is responsible for the content? According to Myles, it is relevant to consider that there are several stakeholders involved in how editorial transparency is communicated. These include the technical creators of the systems, the journalists working within the newsrooms, and the audiences consuming the automatically produced stories.

Understanding of how the systems operate is relevant for publishers, as the decisions are ultimately editorial. Being able to explain how the stories are created is relevant both in-house and towards audiences (Hansen et al., 2017). This transparency should include:

Data (e.g. text and photos)

Information on how the data is collected and by whom. How often it is updated and how is its accuracy verified and validated. Does the data show what we think it shows, is it limited, or does it have gaps? As an example: police statistics will only show you the cases that have been reported and logged, not the actual number of crimes.

When using data, publishers should consider copyright and possible contradictions. Privacy issues need to be aligned with organisational values and regulatory standards. The fact that some information is available publicly does not automatically imply that it can be freely published, e.g, publishing information of a criminal conviction that violates privacy. Temporal issues should be considered: will data that is public now also be automatically public in ten years? Archiving of automated stories is to be considered including what are the country-specific requirements.

Selection of facts

In many cases algorithms used for news automation are created by someone who is not a journalist or an editor. This has implications for news criteria and the selection of stories. As a general rule, the media outlet using new automation has to make sure that editorial decisions are made either by journalists, or according to journalistic principles, even when using external solutions.

Self-regulation

There are usually country-specific ethical guidelines for journalistic self-regulation. As of now, as far as we are aware, there are no self-regulation guidelines (e.g. from professional societies) specifically designed for automated content.

Self-regulatory guidelines apply to all journalistic content, implying that they are applicable to automated content. However, the increased amount of automated content can create contradiction with the existing guidelines, thus calling for a re-interpretation of them.

For example: if media organisations fail to provide understanding of how their automated content has been produced, it can be in violation of the International Federation of Journalists (IFJ) guidelines stating that stories should only be reported "in accordance with facts of which [the journalist] knows the origin". Suppressing essential information regarding the story is also against the IFJ Code of Principles.

Copyright and liability

When news organizations develop, implement and use automated content, they have to consider its legal liability. According to an article written by Lewis, Sanders and Carmody the main things to consider are the complicated matter of determining fault in a case of algorithm-based libel, and the inability of news organizations to adopt defences similar to those used by Google and other providers of algorithmic content (2018).

In addition, local laws and regulations need to be pursued and internalised as well as all agreements between the media company and the possible data provider.

Regulation and guidelines for media publishing generally apply to automated content, according to country-specific standards and legislation. Media outlets should have a common understanding on how these rules are applied to automated content in their publications. If automated processes for creating content are introduced, it is advisable to familiarise and approve the implementation with parties concerned. It should also be made clear how the data has been collected and handled, and how the data has affected the outcome of the article.

By-line

When a story has been produced fully or partly by an automated process, it should be made clear to the readers, listeners and viewers. However, editorial practices are not consistent here. For instance, the RADAR cooperation between the British Press Association and Urbs Media show that newspapers prefer to use the by-line of the reporter that wrote the template. A decision to consider is where to draw the line – what counts as automated?

Data criticism

All data should undergo the same process of verification as all other information and sources that journalists use in their work. Newsrooms should be aware of, and accept, the ways the data has been collected and that a system of continuous evaluation is put into place. Both governmental, public and private data should be scrutinised using the same parameters and evaluated accordingly.

Personalisation

Personalisation is an added value in automated news production, but it contains a risk for creating segregation between audience members. This can be prevented in automatic news generation by selecting facts that are of interest to an individual, while including facts that are collectively viewed as important by the newsroom.

There are several questions that media outlets have to be able to tackle regarding automated and personalised news: should a member of the audience know that the article they are reading is personalised information; and what information is this personalisation based on – individual, member of a certain group, age, living area or something else?

Adding interactivity to personalisation is a way of aiding transparency, and can give insights into audience preferences.

Corrections policy

According to current understanding, media outlets are responsible for all articles they publish, even when an article has been fully produced by automation and personalised. Possible mistakes – caused by data, algorithm or human error – are always the outlet's responsibility.

Chapter 7

7. The future of news automation

Technological development of Natural Language Generation systems happens fast and this form of artificial intelligence is bound to produce exciting results in many fields of society. The media business, however, appears to be largely excluded from this development. Andreas Graefe, an experienced German professor in the field with his own NLG startup, says that despite the hype about "robot journalism" we haven't really moved on in the last three years.

David Caswell, an American entrepreneur and founder of Structured Stories who is now Executive Product Manager of BBC News Labs in London, is even more disillusioned. The industry as well as journalism scholars have exaggerated how far technological development has actually gone. NLG is still in its early stages and we are lacking a complete theory of natural language in the foreseeable future. Noam Chomsky as well as other linguistics experts have no prospect for any fundamental understanding of what language is and what the model behind it is. This is reflected in today's news automation, for instance at the Associated Press.

"Five years ago, there were many bold predictions about how automated journalism will develop. From claims that 90% of news will be automated to Pulitzer prizes for automated content. In reality, not much has changed. Progress is steady but slow."

- Andreas Graefe, Endowed Sky Research Professor at Macromedia University (interview with Lindén, 2018)

The hype is a general problem connected to the promise of Artificial Intelligence and as noted by Amber Case – the technology and culture researcher – we need to separate the kind of super-advanced AI that we have encountered in sci-fi films from the process automation that is largely an extension of the Industrial Revolution (Case, 2018). This is also a view broadly shared by the research group behind Immersive Automation. The future of news automation will focus on engineering and process-oriented applications of AI that can address the clear needs of news information consumers within narrow, well-understood domains.

"They are essentially template based. They are not unsophisticated, but they are, in no way, intelligent. There is no machine learning. There is no artificial intelligence, of any kind, behind them."

- David Caswell, Executive Product Manager at BBC News Labs (interview with Lindén, 2018)

7.1. Deconstruction of journalism: 7 conclusions

The future is most certainly bright but there is a lot of work to do, starting with the basics. The future of automation lies in deconstruction of the fundamental principles of journalism. That means breaking down journalistic work into the actual information artefacts and micro processes so as to analyse what can be automated and what are inherently human tasks. Through carefully breaking down tasks, hybrid systems combining automated and human effort in complementary ways will pave the way for higher efficiencies and lower costs while maintaining quality and resisting commodification (Diakopoulos, 2019).

Through a round of discussions with experts on news automation, combined with our own thinking, we have come to the following sobering conclusions about the realistic challenges surrounding the effective implementation and deployment of automated content:

■ Automatically generated texts beyond the most basic templating systems are often still prone to error. Effective automated content systems must be cognizant of the alien and unfamiliar errors automation can produce, and inject appropriate editorial oversight, management, and maintenance into the process.

■ NLG systems are still quite unsophisticated and their extendability outside texts on sports, real estate or finance is limited by several factors. It is hard to design and contextualise narratives in advance for subjects where the outcome is uncertain, like politics. Careful design and engineering is necessary to deploy automated content even within narrow, well understood, and clearly-defined domains.

■ The availability of interesting and useful data is a crucial issue as there are strong private interests that try to control and commercialise data. The only operational stories that news organisations have tackled so far have been the obvious ones that already have a

"That is the pathway to many different new uses of journalism, and new business models for journalism, and new ways for sense making in journalism."

- David Caswell, Executive Product Manager at BBC News Labs (interview with Lindén, 2018)

lot of data pre-assembled, such as financial and sports data. But, as the case of British RADAR shows, open data can be a rich source for news automation. Media companies should campaign more for open public and private data. Alternatively they should develop strategies for creating and partnering to obtain unique datasets, as this is key input for generating interesting and commercially promising content. New technical capabilities for automatically gathering data in targeted domains via voice-based AI could drastically lower costs, creating altogether new use cases for automated content (Diakopoulos, 2018).

■ News automation provides media companies with an opportunity to expand their businesses outside traditional news. However, there will probably be heavy competition from new entrants to the field that are better equipped to experiment with new features. New entrants will be able to experiment without being weighed down by the ballast of the past, technological debt or organisational forms of path dependency. Legacy news organizations may benefit from establishing entrepreneurial incubators for automation projects that are unencumbered by traditional workflows.

■ The flexibility of NLG systems is still limited and versions for chatbots or talking/listening machines, such as Alexa, require a lot of expensive development work. This is one reason why Alexa and similar services are provided only in major languages.

■ Personalisation of automated news is the dream of every publisher. That requires extensive profiles of customers as well as predictive models based both on online and offline behaviour. With GDPR and consumer awareness regarding privacy on the rise, media companies may be disadvantaged compared to established and broadly accepted services such as Google and Facebook. In the longer term, media companies must decide whether they want to participate in the deep user modelling needed to extensively personalize automated content.

■ It might very well be that the most useful applications of news automation is for drawing insights from large datasets that journalists can use as raw material for creating interesting stories. Automation will help journalists and editors create the news with automated analysis, partially written texts and other sophisticated tools. However, the risk is that in practice automated processes, instead of relieving journalists and editors from routine tasks, will create even more work going through the output.

7.2. Future scenarios

In our own project, we involved the steering group of Immersive Automation in some forecasting to see where news automation is heading. We used a method called Backcasting, which has been utilised in many fields, particularly within service design. Backcasting is a navigating tool where participants navigate the paths from desired futures back to the current situation.

One of the groups envisioned a future where media companies and editorial staff have many more tools at their disposal to help their work, for instance with story discovery. The question was how AI can assist the reporter in his or her work. The strengthening field of human computer co-creation was of interest to this group and was discussed.

Another group took on the question of data as a prerequisite to automation and AI. How can the large enough scale be guaranteed, especially in small countries with data scarcity? What should the vision be and how should it be realised? The participants also stressed the importance of cooperation and were pondering whether institutions such as Statistics Finland could have an important role in generating data, or whether this should be left to commercial actors.

We also believe that the arrival of 5G will provide altogether new opportunities for creating and distributing immersive and meaningful content – text, video, sound, social signals – based on personal interests, time, location and activity. This form of news is targeted specifically at real time needs with snippets of information generated automatically.

It can also be other forms of signals. In Germany, NOZ Medien/mh:n Medien together with the company Datenfreunde is working on a development project funded with money from Google's Digital News Initiative called "Ambient News". The term 'ambient' has been used in media and journalism to describe the

"So how would your lamp be illuminated in the morning? When you have a big traffic jam going on, is it more red, or is it more green? What is being shown on your (smart) mirror in the morning? Is it relevant to you, is it not? That is what we are trying to figure out because a lot of content could be much more data-driven than it is today. It could be much more automated and brought to the right audience."

- Nicolas Fromm, Digital Manager at NOZ Medien (interview with Lindén 2018)

"He (the journalist) knows when they wake up, basically, when they have to go to work and where they work and where they live. So we can decide what kind of information for each person."

- Marco Maas, CEO of Datenfreunde GmbH (interview with Lindén 2018)

ubiquitous nature of news in today's society (Hermida, 2010). The project focuses on transforming journalistic content into data and converting the data into social signals in smart homes. Nicolas Fromm, digital manager at NOZ Medien/mh:n Medien, explains the operational logic. The first experimental phase at a local newspaper focuses on smart light bulbs but the intention is to also use smart mirrors.

"The experiment is designed so that one journalist at the newspaper in Flensburg is connected to 10 readers that act as testers," explains Datenfreunde CEO Marco Maas. "Experiments like these and many more exemplify the potential of automated content generation, but as previously noted, whether legacy media will show the way forward remains to be seen."

Chapter 8

8. Footnotes and sources

The Immersive Automation is a research and development project that aims at automating news production, and is looking for new ways of making it more automatic. It also looks at the ecosystem of news production. In Immersive Automation a research consortium of data scientists, linguists, and journalists work together with researchers and media companies to bring forward a solution for producing engaging, data driven content. The project's aim was to create a roadmap and a demonstration of a future news ecosystem based on automated storytelling, intense audience engagement and user experience. The project started in early 2017 and continued until the end of May 2018.

Immersive Automation was funded by Business Finland, The Media Industry Research Foundation of Finland, The Swedish Cultural Foundation in Finland. The media companies involved as well as the research organisations participating in the project are: VTT Technical Research Centre of Finland, University of Helsinki, Sanoma, Alma Media, Conmio, Keskipohjanmaa, Kaleva, Streamr and KSF Media.

Authors: Lindén, Carl-Gustav; Tuulonen, Hanna; Bäck, Asta; Diakopoulos, Nicholas; Granroth-Wilding, Mark; Haapanen, Lauri; Leppänen, Leo; Melin, Magnus; Moring, Tom; Munezero, Myriam; Sirén-Heikel, Stefanie; Södergård, Caj; Toivonen, Hannu; Wuori, Naomi

Sources

Akst, D. (2013). "Automation anxiety." The Wilson Quarterly, Summer 2013. Available at: http://archive.wilsonquarterly.com/sites/default/files/articles/AutomationAnxiety.pdf [Accessed 17.2.2018]

Andersen, P. M.; Hayes, P. J.; Huettner, A. K.; Schmandt, L. M.; Nirenburg, I. B. & Weinstein, S. P. (1992). "Automatic extraction of facts from press releases to generate news stories." Association for Computational Linguistics. Proceedings of the third conference on Applied natural language processing, pages 170–177.

BDZV/SCHICKLE, Trends der Zeitungsbranche 2017 Berlin, 1. Februar 2017.

Bouayad-Agha, N.; Casamayor, G.; Mille, S. & Wanner, L. (2012). "Perspective-oriented generation of football match summaries: Old tasks, new challenges." ACM Trans. Speech Lang. Process., 9(2):1–31.

Boyd, D. & Crawford, K. (2012). "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon." Information, communication & society, 15(5), 662-679.

Case, A (2018): "How to Design a Better Internet: Calming Technology, Humanizing the Algorithm, and a New Xerox PARC." Medium. Available at: https://medium.com/@caseorganic/how-to-design-a-better-internet-calming-technology-humanizing-the-algorithm-and-a-new-xerox-parc-e4d702be8787 [Accessed 14.8.2018]

Clerwall, C. (2014). "Enter the Robot Journalist." Journalism Practice, 8:5, 519-531, Available at: https://doi.org/10.1080/17512786.2014.883116 [Accessed 28.8.2018]

Diakopoulos, N. (2018). "There are a lot of rote tasks a good AI interviewer could do for you." Columbia Journalism Review. Available at: https://www.cjr.org/tow_center/artificial-intelligence-reporting-interviews.php [Accessed 28.8.2018]

Diakopoulos, N. (2019, forthcoming): "Automating the News: How Algorithms are Rewriting the Media." Harvard University Press.

Govik, R. (2018). "The homeowners bot." Available at: https://medium.com/MittMedia/the-homeowners-bot-36d2264e2d34 [Accessed 21.2.2018]

Graefe, A., Haim., M., Haarmann, B. & Brosius, H-B. (2016). "Readers Perception of Computer-Generated News: Credibility, Expertise, and Readability." Journalism. https://doi.org/10.1177/1464884916641269

Guel, S. (2009). "News production automation." CSI Magazine. Available at: http://www.csimagazine.com/csi/News-production-automation.php [Accessed 27.2.2018]

Hansen, M.; Roca-Sales, M.; Keegan, J.M. & King, G. (2017): "Artificial Intelligence: Practice and Implications for Journalism". Columbia University Academic Commons, https://doi.org/10.7916/D8X92PRD

Hermida, A. (2010). "Twittering the news: The emergence of ambient journalism." Journalism practice, 2010, 4.3: 297-308.

IBM (n.d.): "Advancing AI with Project Debater." Available at: https://www.research.ibm.com/artificial-intelligence/project-debater/research.html [Accessed 15.8.2018]

Jung, J., Song, H., Kim, Y., Im, H., & Oh, S. (2017). "Intrusion of software robots into journalism: The public's and journalists' perceptions of news written by algorithms and human journalists." Computers in Human Behavior, 71, 291-298.

Kent, T. (2015): "An ethical checklist for robot journalism." Medium. Available at: https://medium.com/@tjrkent/an-ethical-checklist-for-robot-journalism-1f41dcbd7be2 [Accessed 6.8.2018]

Kodjo, E. (2017). "Artificial intelligence can't solve every problem in the media, but it can take care of these." Medium, Global Editors Network. Available at: https://medium.com/global-editors-network/artificial-intelligence-cant-solve-every-problem-in-the-media-but-it-can-take-care-of-these-bb633474c39 [Accessed 17.2.2018]

Latar, N. L. (2015). "The Robot Journalist in the Age of Social Physics: The End of Human Journalism?." The New World of Transitioned Media: Digital Realignment and Industry Transformation, 2015, pp. 65–80. Wiesbaden: Springer.

Lavenda, D. (2016). "Artificial Intelligence vs. Intelligence Augmentation." Available at: https://www.networkworld.com/article/3104909/software/artificial-intelligence-vs-intelligence-augmentation.html [Accessed 22.2.2018]

Lecompte, C. (2015). "Automation in the Newsroom. How algorithms are helping reporters expand coverage, engage audiences, and respond to breaking news". Nieman Reports. Available at: http://niemanreports.org/articles/automation-in-the-newsroom/ [Accessed 12.2.2018]

Leppänen, L.; Munezero, M.; Granroth-Wilding, M. & Toivonen, H. (2017, a). "Data-Driven News Generation for Automated Journalism." In Proceedings of the 10th International Conference on Natural Language Generation (pp. 188-197). Available at: http://www.aclweb.org/anthology/W17-3528 [Accessed 21.8.2018]

Leppänen, L.; Munezero, M.; Sirén-Heikel, S.; Granroth-Wilding, M. & Toivonen, H. (2017, b): "Finding and expressing news from structured data." In Proceedings of the 21st International Academic Mindtrek Conference (pp. 174-183). ACM Available at: https://www.cs.helsinki.fi/u/htoivone/pubs/Leppanen_et_al_Mindtrek_2017.pdf [Accessed 21.8.2018]

Lewis, S. C. & Westlund, O. (2015). "Big data and journalism: Epistemology, expertise, economics, and ethics." Digital Journalism, 3(3), 447-466.

Lewis, S. C.; Sanders, A. K.; Carmody, C. (2018): "Libel by Algorithm? Automated Journalism and the Threat of Legal Liability." Available at: http://journals.sagepub.com/doi/10.1177/1077699018755983 [Accessed 15.8.2018]

Lindén, C. & Karlsson, S. (2018). Personal communication.

Lindén, C. (2017). " "Robot Journalism": The damage done by a metaphor." Data Driven Journalism. Available at: http://datadrivenjournalism.net/news_and_analysis/robot_journalism_the_damage_done_by_a_metaphor [Accessed 26.2.2018]

Los Angeles Times (2014). "Quakebot." Available at: http://www.latimes.com/local/lanow/earthquake-27-quake-strikes-nearwestwood-california-rdivor-story.html [Accessed 7.4.2017]

Melin, M., Bäck, A., Södergård, C., Munezero, M. Leppänen, L. & Toivonen, H. (2018): "No landslide for the human journalist – An empirical study of computer-generated election news in Finland." IEEE Access, Early access, DOI: 10.1109/ACCESS.2018.2861987

Mendez-Nunez, S. & Trivino, G. (2010). "Combining semantic web technologies and computational theory of perceptions for text generation in financial analysis." In Fuzzy systems (fuzz), 2010 IEEE international conference on, pages 1–8. IEEE.

Mori, M. (1970/2012). "The uncanny valley." IEEE Robotics & Automation Magazine, 6(2012). Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6213238 [Accessed 7.5.2018]

Myles, S. (2018): "How Can We Make Algorithmic News More Transparent?". Associated Press. Presentation at the conference Algorithms, Automation and News, Munich, 22nd May 2018.

Nesterenko, L. (2016). "Building a system for stock news generation in Russian." In Gardent, C. & Gangemi, A. editors, Proc. of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016), pages 37–40, Stroudsburg, PA. ACL.

Picard, R. (2010): "Value creation and the future of news organizations: Why and how journalism must change to remain relevant in the twenty-first century." Lisbon: Media XII.

Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., & Sykes, C. (2009). "Automatic generation of textual summaries from neonatal intensive care data." Artificial Intelligence, 173(7-8), 789-816.

Prakash, S. (2017). "Keynote: Journalism & Technology: Big Data, Personalization & Automation." Available at: https://www.youtube.com/watch?v=PqMvxo89AQ4 [Accessed 7.3.2018]

Reiter, E. & Dale, R. (2000). "Building natural language generation systems." Cambridge University Press.

Reiter, E., Robertson, R., & Osman, L. M. (2003). "Lessons from a failure: Generating tailored smoking cessation letters." Artificial Intelligence, 144(1-2), 41-58.

Reuters (2017). "Reuters Institute Digital News Report". Available at: http://www.digitalnewsreport.org/ [Accessed 26.2.2018]

Sharman, D. (2018): "Four journalists join PA robot reporting unit as scheme expands." Available at: https://www.holdthefrontpage.co.uk/2018/news/four-journalists-join-pa-robot-reporting-unit-as-scheme-expands/ [Accessed 27.8.2018]

Sundar, S. S., Waddell, T. F. & Jung, E. H. (2016). "The hollywood robot syndrome: Media effects on older adults' attitudes toward robots and adoption intentions." Paper presented at the The Eleventh ACM/IEEE International Conference on Human Robot Interaction, 343-350.

Sundar, S.S., (1999). "Exploring Receivers' Criteria for Perception of Print and Online News." Journalism & Mass Communication Quarterly 76 (2): 373–386.

Sundgren, T. (2017). "Soldr personalization service — MittMedia innovation for survival." Available at: https://medium.com/MittMedia/soldr-personalization-service-MittMedia-innovation-for-survival-fda26f7cdbee [Accessed 21.2.2018]

Suthaharan, S. (2014). "Big data classification: Problems and challenges in network intrusion prediction with machine learning." ACM SIGMETRICS Performance Evaluation Review, 41(4), 70-73.

The Washington Post (2016). "The Washington Post experiments with automated storytelling to help power 2016 Rio Olympics coverage." Available: https://www.washingtonpost.com/pr/wp/2016/08/05/the-washington-post-experiments-with-automated-storytelling-to-help-power-2016-rio-olympics-coverage/?utm_term=.8dd58dd87d77 [Accessed: 7.3.2018]

The Washington Post (2017). "The Washington Post leverages automated storytelling to cover high school football." Available at: https://www.washingtonpost.com/pr/wp/2017/09/01/the-washington-post-leverages-

heliograf-to-cover-high-school-football/?utm_term=.2318729cad5c [Accessed 7.3.2018]

Theune, M.; Klabbers, E.; de Pijper, JR.; Krahmer, E. & Odijk, J. (2001). "From data to speech: a general approach." Natural Language Engineering, 7(01):47–86.

Tuulonen, H. & Dai, Z. (2018). Personal communication.

Tuulonen, H. (2017). "A possibility, a threat, a denial? How news robots affect journalists' work practices and professional identity." University of Gothenburg, MA Thesis in Investigative Journalism.

United Robots (2017). "Rosalinda for sports." Available at: http://www.unitedrobots.se/produkter-1/ [Accessed 4.4.2017]

Wamba, S.F.; Akter, S.; Edwards, A.; Chopin, G. & Gnanzou, D. (2015). "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study." International Journal of Production Economics, Volume 165, July 2015, Pages 234-246.

van der Kaa, H. A. J. & Krahmer, E. J. (2014). "Journalist versus news consumer: The perceived credibility of machine written news." In Proceedings of the Computation+Journalism conference New York.

Wasp3D News (n.d.). Available at: http://www.wasp3d.com/news.html [Accessed 27.2.1018]

Wölker, A. & Powell, T. E. (2018). "Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism." Journalism. https://doi.org/10.1177/1464884918757072

Yle (2018): "Voitto-robotti avataan kaikkien käyttöön." Available at: https://yle.fi/uutiset/3-10125596 [Accessed 26.5.2018]

Zuboff, S. (1988). "In the age of the smart machine: The future of work and power." New York: Basic Books.

Acknowledgements

This report wouldn't have been possible without the research and hard work done by the Immersive Automation project, spearheaded by Carl-Gustav Lindén and his team. The project was financed by Business Finland (former Finnish Funding Agency of Innovation Tekes), The Media Industry Research Foundation of Finland, The Swedish Cultural Foundation in Finland, the media companies involved as well as the research organisations participating in the project: VTT Technical Research Centre of Finland, University of Helsinki, Sanoma, Alma Media, Conmio, Keskipohjanmaa, Kaleva, Streamr and KSF Media.

About the Authors

Lindén, Carl-Gustav

Carl-Gustav Lindén is the project leader for the Immersive Automation research project, and an Adjunct Professor (Docent) at the Swedish School of Social Science, University of Helsinki, Finland. His particular interests are journalism and media innovations, as well as media business models.

Tuulonen, Hanna

Hanna Tuulonen is a PhD student at the University of Helsinki and a freelance journalist. In the Immersive Automation research project, she worked as a project planner. Hanna is interested in Chinese data-driven journalism and news automation and their impact on journalism and media both inside and outside China.

Bäck, Asta

Asta Bäck is a principal scientist at VTT Technical Research Centre of Finland, and in the Immersive Automation research project she focuses on news ecosystems.

Diakopoulos, Nicholas

Nicholas Diakopoulos is an Assistant Professor in the School of Communication at Northwestern University where he is Director of the Computational Journalism Lab (CJL). He is also a Tow Fellow at Columbia University School of Journalism as well as Associate Professor II at the University of Bergen Department of Information Science and Media Studies.

Granroth-Wilding, Mark

Mark Granroth-Wilding is a research associate at the Department of Computer Science, University of Helsinki, Finland, with expertise in Artificial Intelligence, in particular: Natural Language Processing, Computational Music Analysis and Computational Creativity.

Haapanen, Lauri

Lauri Haapanen is a post doc-researcher at the University of Jyväskylä, Finland. Lauri is interested in media products and journalistic work processes from a linguistic point of view.

Leppänen, Leo

Leo Leppänen is a doctoral student at the Department of Computer Science, University of Helsinki, Finland, with expertise in Natural Language Generation and Data Analytics.

Melin, Magnus

Magnus Melin was recently a research scientist at VTT Technical Research Centre of Finland, working mainly with Linked Open Data. Currently he is employed by the Mozilla Developer Network.

Moring, Tom

Tom Moring is a Professor Emeritus of Communication and Journalism at the Swedish School of Social Science, University of Helsinki, Finland. Tom specialises in political communication, broadcasting, and minority languages and media.

Munezero, Myriam

Myriam Munezero is a text and data analyst as well as a researcher in natural language processing and natural language generation.

Sirén-Heikel, Stefanie

Sirén-Heikel is a PhD-candidate in media and communication studies at the Faculty of Social Sciences at the University of Helsinki, Finland. Previously she has worked as a journalist in both print and broadcast.

Södergård, Caj

Caj Södergård is a research professor in digital services and media at VTT Technical Research Centre of Finland Ltd. Caj has a worked within image processing, big data, and artificial intelligence for more than 30 years, mostly with applications within the media sector.

Toivonen, Hannu

Hannu Toivonen is a professor of computer science at the University of Helsinki, Finland, since 2002. He works in the areas of artificial intelligence and data science, more specifically in computational creativity, data mining as well as analysis and generation of natural language.


WAN IFRA
World Association of Newspapers and News Publishers