



An Evaluation of the Pilot Application of Artificial Intelligence to Witness Statement and Report Generation at Hertfordshire Constabulary

Dr Paul Walley, Director of Learning,

Dr Helen Glasspoole-Bird, Research Fellow

Report Date 23/3/2025

Centre for Policing Research and Learning
The Open University
Walton Hall
Milton Keynes
MK7 6AA

To contact the authors email paul.walley@open.ac.uk



Contents

Executive Summary.....	4
Introduction	4
Literature	4
Methodology	4
Findings	4
Conclusions and Recommendations	4
Introduction	5
RVR and ADA	6
MG11 Statements.....	7
Literature	8
Literature Scoping Results.....	8
The Public Sector and AI.....	8
Policing applications.....	10
Ethics	13
Official policy documents	14
Section Summary	15
Methodology.....	16
Interview Methodology	16
Report Quality Methodology	17
Grammar, spelling and readability	17
Linguistic Quality.....	18
Performance Data	19
Results.....	21
Interview Findings	21
Officer involvement – leading the interview.....	21
Officer involvement – checking the statement	23
Narrative.....	24
Capturing non-audio content.....	26



Engagement with the witness	27
Perceived efficiencies and cost savings.....	28
Section summary.....	30
Report Quality Findings	31
MS Editor Scores	31
Linguistic Analysis Outcomes.....	32
Figure 1 The Frequency of Register Labels (% of sample)	33
Performance Data Findings	37
Discussion	38
Defining output quality	38
Comparing AI & Human Statements.....	39
Productivity.....	40
Figure 2 A comparison of intended and actual processes	43
Development Costs	43
Key findings	45
The effectiveness of ADA	45
RVR	46
Implications of the use of ADA	46
Other Observations	47
Recommendations – Hertfordshire	47
Recommendations – Policing	48
Concluding remarks.....	49
Limitations	50
References	51
Appendices	55
Appendix 1:	55



Executive Summary

Introduction

This report produces an effectiveness evaluation of the Version 1 of an Artificial Intelligence (AI) Application, referred to as “ADA”, in Hertfordshire Constabulary to produce witness statements and related documents from the audio of Rapid Video Response contact with victims of domestic abuse.

Literature

The evaluation reviewed the existing literature about AI applied to this work and identified a range of critiques. For example, the problem of AI hallucinations, where the AI invents untrue content, needs to be addressed. Other issues include the lack of non-audio content in AI-generated statements.

Methodology

We conducted a range of user and stakeholder interviews, analysed the quality and linguistic style of the AI output and studied productivity and process performance data to reach conclusions.

Findings

The readability of AI statements was acceptable and there were few grammatical or spelling errors, especially once the initial errors had been eliminated. However, the AI was poor at writing the statements with a clear context and narrative about the specific incidents. Readability was low but just about acceptable. The AI used words not normally used by witnesses. Statements were mainly academic in style and contained small errors and omissions, partly hallucinated by the AI. The AI provided good assistance to those with difficulties in writing statements, but experienced officers produced better work. The findings about the productivity impact concluded that some gains were counteracted by the need for additional quality checks.

Conclusions and Recommendations

We concluded that the test of concept had not yet been made but a version 2 of the AI should produce better results. The final section contains further conclusions and recommendations both for the project and lessons learned for policing.



Introduction

Artificial Intelligence (AI) is increasingly seen as a major source of productivity improvement in information-based processes for at least the next decade in both public and private organisations (Kalai et al., 2024). This applies very much to policing where the last decade or so of austerity, with restrictions on police budgets, has focused attention on the need to improve productivity. The National Police Chiefs' Council priorities for policing have moved in this direction. Police work is also becoming more data intensive. Practically all police investigations will now involve digital evidence, such as data from mobile phones and body-worn cameras, and all of this information needs to be stored, processed, studied and reported, often with some urgency. Processes that have traditionally been paper-based are now digital and there is still opportunity to improve the ways in which much of the information used by police is processed. AI applications are often seen as a way of improving how this data is analysed and reported and there is appetite in the police to embrace this.

This report summarises the early independent evaluation of one pilot application of AI, referred to as “ADA” (Anathem Digital Assistant), that is being used to produce a series of documents related to incidents of domestic abuse, including the MG11 Witness Statement. This work sits alongside the evaluation being conducted by Hertfordshire Constabulary and uses some of the same data. We have conducted additional research and analysis in compiling this report. The purpose of our evaluation is to understand the impact of the AI on productivity and quality in generating these statements and reports. The evaluation uses a combination of interviews, analysis of AI outputs and process data to identify the challenges, benefits and improvement opportunities that exist at this early stage of the AI development.

We address three research questions in the study:

RQ1: What are the users and other stakeholders reporting as the advantages, disadvantages and issues with the use of ADA?

RQ2: What are the measurable differences between the quality of the output of ADA when compared with outputs written without AI assistance?

RQ3: What evidence is there about the impact on productivity and workload when AI is used to generate statements?

The report is divided into five main sections. First, we present the findings from a scoping study of existing academic evidence about police applications of AI. We limit this study to general literature on issues such as ethics of AI in policing and then the use of AI as a productivity tool. We do not cover other areas of use such as facial recognition, data analytics or forecasting. Our second section provides an explanation of the mixed research methods employed during this study. The qualitative and



quantitative results are presented in a third section, followed by a discussion of the results and their implications. The final concluding section summarises the key findings and makes recommendations for the AI project. We include lessons more generally for UK policing to consider when addressing the future adoption of AI productivity tools.

RVR and ADA

The ADA application has been trialled using the audio generated using GoodSAM by Rapid Video Response (RVR) contact with victims of domestic abuse. GoodSAM initially launched for life saving purposes in 2013 to change the response to cardiac arrests, this technological development has been adopted by the police for a range of services. RVR is used to screen the urgency of domestic abuse cases; non-urgent cases are dealt with via remote video contact with victims rather than in-person attendance by officers. RVR has already been demonstrated higher victim satisfaction, higher arrest rates and improved trust in confidence with police (Rothwell et al., 2022) when compared with conventional approaches to domestic abuse incidents. This is partly because it provides a better opportunity to collect evidence (Koppensteiner et al. 2022), identify hidden harm (Nicholls et al., 2013) and addresses a perennial problem of downplay of abuse (Felson, 2006). In cases where RVR is used, an officer interviews the witness via the video link and is expected to help officers produce a witness statement in real time without any break in contact. This is an attempt to capture all evidence and testimony promptly, avoiding issues associated with withdrawal of complaints in the time between an incident and the statement being taken. Statements are signed by witnesses electronically.

ADA is used in conjunction with RVR, taking the audio from the RVR contact and automatically transcribing this (using the transcribing tool). The transcript is then used to generate a series of related documents: the MG11 witness statement, the DARA risk assessment and handover summaries. A feature of the application is that it shows some of the sources of risk and “points to prove” when generating these statements. Unlike many other AI applications, ADA shows its working when generating output. ADA does not compile any output until the interview has concluded. This means ADA cannot provide real-time advice, such as pointing out missing items of evidence etc. while the interview is occurring.

It should be noted that ADA does not actually make decisions, such as a final risk assessment, but puts the information into a suitable format for officers to use. For a typical case, ADA will take about 2-3 minutes to produce an initial MG11 from the completed audio, before the officer checks this for accuracy and usability. The officer and witness still have to agree on the final version of the statement before it is signed electronically by the witness.



Our report only evaluates version 1 of ADA. During the data collection phase of this work a completely revised version of ADA has been developed and is now ready for the first trials.

MG11 Statements

A witness statement must meet certain requirements for it to be acceptable to courts, using a standard pro-forma. The statement is intended to represent the voice of the witness, using their own words, comparable to an oral testimony in court. It has to be factual, pointing towards the sources of evidence where claims are made. The term “seen, heard and felt” is often used in instructions to officers and witnesses when compiling a statement. The language style is therefore different to many formal reports. For example, where a statement includes what an offender has said this should include unredacted swear words. (Most transcribing tools automatically redact this language).



Literature

This section provides a narrative summary of the literature found during a rapid review of the literature into the application of policing and the wider public sector. It is limited to articles that have been found in the standard library databases where the articles contain “artificial intelligence” or “AI” in titles, key words or abstracts and reference to either “policing” or related words (“police” etc.). Additional grey literature has been found using Google Scholar and from known reference sources including the NPCC and the College of Policing. Exclusion criteria were developed; articles more than 10 years old (due to the rapid pace of change) and generic AI applications, such as chatbots were excluded. The purpose of the search is to establish what evidence already exists that provides any evidence base for the effectiveness of AI when using in well-established policing processes and any related issues that may affect evaluation, adoption, development or implementation. The work is structured into relevant themes. Note that we have focused our literature summary on work that offers the greatest relevance to this study. Hence the content here does not contain any detail on other unrelated police AI applications such as facial recognition.

Literature Scoping Results

In total 36 articles were directly found using the search protocol. These articles referenced many others which were also scrutinised for relevance. Articles have been clustered into the following sub-groups:

1. Public sector-related articles
2. Articles on studies of policing applications
3. Ethics, policing and AI
4. Official policy and guidance within policing

The Public Sector and AI

There is a small body of literature that focuses on the use of AI in the public sector. Wirtz et al. (2019) first state that the productivity potential of AI, that is causing commentators to predict consequential significant economic growth, can also be applied to public organisations. Their scoping study of the literature from the public sector found just 14 studies that addressed AI applications, with a total of 30 articles ultimately found using additional search mechanisms. The papers they found covered five topics:

1. AI Government service
2. Working and social environment influenced by AI
3. Public order and law-related to AI
4. AI ethics
5. AI government policy

The papers of relevance to this study were included in a discussion of the trade-offs associated with AI. The use of AI to support surveillance technologies (i.e. face recognition etc.) is seen as a positive in terms of public protection but does raise civil liberties issues.

Wirtz et al. (2019) conclude:

“Against the background of emerging applications for the public sector that promise great public value, major challenges are arising with regard to AI responsibility, as well as social and ethical issues, which potentially threaten value creation for public service providers and agencies (Quraishi et al., 2017; Ransbotham et al., 2017).”

The paper produces a list of examples of applications of AI under the following categories:

1. AI Knowledge management systems
2. AI Process Automation Systems
3. Virtual Agents
4. Predictive Analytics and Data Visualisation
5. Identity Analytics
6. Cognitive Robotics and Autonomous Systems
7. Recommendation systems
8. Intelligent Digital Assistants
9. Speech analytics
10. Cognitive Security Analytics and Threat Intelligence

A separate study (De Sousa et al., 2019) shows a similar list of applications of AI in the public sector based upon a literature review, noting that the US and India are the countries where most of this work is being conducted. The UK is one of ten countries at the top of this list in terms of the number of studies.

The Wirtz et al. (2019) paper concludes with four separate challenges for the adoption of AI:

Technology Implementation

The work provides a discussion of a wide-ranging set of both technical and managerial challenges for implementation including AI safety, system/data quality and levels of integration.

AI Laws and Governance

The potential for AI to cause harm through poorly controlled use features in many articles, requiring carefully designed governance mechanisms.



AI Ethics

The issues of ethics will be covered in depth later in this report. A key point is that the use of AI for decision-making is problematic where there is no human override or mechanism by which the decision logic can be scrutinised.

AI Society

Perhaps inevitably there are concerns that AI will have some detrimental effects on society including the consequences for workforce substitution.

A later paper by the same team of authors (Wirtz et al., 2021) produces a more detailed scoping review of the same literature with a slightly different set of inclusion and exclusion criteria. The general conclusions are similar, but it is noteworthy that policing applications are given specific mention in their conclusions:

“There are many risks, such as AI-based citizen surveillance (Feldstein, 2019; Weyerer & Langer, 2019) or predictive policing (Asaro, 2019; Hiroshi, 2020; Langer, 2020), that are controversial but often seem speculative in their consequences due to their limited actual use cases to date. Concerning the area of regulation, Kuziemski and Misuraca (2020, p. 2) state that “the public sector’s predicament is a tragic double bind: [. . .] to govern algorithms, while governing by algorithms”. Accordingly, the general handling of AI within Government is a systematic challenge.”

The issues around collaboration in AI development are also covered in the literature (Campion et al., 2022). They mention issues such as security concerns and data sharing issues as the more tangible challenges. Other comments point towards implementation and organisational issues.

One paper (Haesevoets et al., 2024) presented three studies of the public acceptability of the use of AI in public services. They concluded that the public want AI to have some say in public sector decisions but prefer more human input for more technical decisions. They either want no role or an advisory role for AI in decision-making and then only when decisions are less ideologically charged. Schiff et al., (2023) also point out some of the reluctance by the public for AI to be involved in decision-making, especially if this is driven at a national level. They refer to “bureaucratic proximity” of decision-making as a public preference.

Policing applications

This section will look at the papers that are based upon police applications of AI for report generation and the issues around them.

Adams (2024) suggests, in a US context, that police officers do not necessarily have good report-writing skills and there are potential advantages in having AI-assisted



report-writing available. The potential for improved accuracy in reports is of particular interest as this can feed through to better outputs and stronger cases. Adams also points to the balancing of risks, managing both the advantages and downsides of AI adoption. A number of Large Language Model (LLM) applications are described in this work, focusing on the evolution of applications based around ChatGPT. One enhanced GPT method is Truleo ([Automate Your Reporting Workflow | Try Our A.I.-Powered Report Narration for Free | Truleo](#)). This application allows officers to leave a voicemail that converts the audio into a report ready for final checking. This is secure as it is provided using “GovCloud”. The work also compares the GPT method to a more human-based “recorded line protocol” where officers send in audio that is more manually transcribed and sorted.

Adams et al. (2024) present the results of a randomised control trial of a report writing tool called “Draft One” produced by Axon. This technology takes the audio from body-worn cameras and converts it into an incident report ready for checking and corrections. There are some built-in features that prompt officers to check details and insert information into the draft, increasing engagement with the report. Draft One is based on ChatGPT.

Their initial observation is that officers already use “boilerplate prose” (see Dement and Inglis, 2024) and templates when writing reports. This impacts on timesaving as some of the actions of the AI merely replicate this. However, the quality of the output was noteworthy:

“While our study found no significant time savings—contrary to the marketing claims surrounding AI—efficiency should not be the sole focus. Report quality remains a persistent concern in policing, with long-standing issues related to poor spelling, grammar, voice, and tone. AI assistance has the potential to address these issues, and Axon’s internal study suggests that their Draft One system produces reports with improved terminology and coherence while maintaining similar levels of completeness, neutrality, and objectivity.” (Adams et al., 2024).

The negative aspects of the use of AI concerned the comparative standardisation of reports. The impact on the legal system needs to be considered:

“AI-generated reports may be perceived as more uniform or polished, which could influence how they are interpreted or valued by different stakeholders. This could lead to positive outcomes, such as increased credibility and readability, but also negative consequences, such as a reduced sense of transparency or authenticity.”

The Draft One system has been criticised by those who see AI as part of a threat to civil liberty. The ACLU paper (Stanley, 2024) first points out the problems associated with AI technology development: in particular the bias created by sourcing from the internet:

“Because LLMs are trained on something close to the entire Internet, they inevitably absorb the racism, sexism, and other biases that permeate our culture.”

They suggest that even when filters are introduced, they can be bypassed, and problems can still remain. Their four main issues are:

1. AI suffers from “hallucinations” and therefore makes things up. It also contains hidden bias that may not be corrected by human intervention.
2. Evidential and memory problems - video is selective about what is captured. Other perspectives might not be recorded.
3. Transparency and discovery problems, such as the potential lack of ability to interrogate evidence. AI applications need to have creativity switched off.
4. Loss of “disciplinary” function in report writing, e.g. the need to justify actions such as stop and search.

Policing reports are used by legal professionals and one of the potential challenges is for AI-generated reports to be accepted by the legal profession. The main risk is that anything generated by AI is formally challenged in court as an unacceptable or erroneous piece of evidence. Harasta et al. (2024) see the legal profession using AI in the following ways:

1. **Summarisation**

Much legal work involves summaries of multiple documents or providing precis of larger legislative proposals. This involves considerable time and effort if done by people without any AI support.

2. **Translation**

Where reports in one language need to be translated.

3. **Legal question answering**

AI could be used to develop either chatbots or expert systems that provide advice about a particular legal situation.

4. **Legal reasoning**

AI could be designed to work out more complex aspects of legal reasoning (beyond the above recollection of existing law).

5. **Legal research**

Similar to literature reviews in an academic context much legal work is about finding/discovering information and sharing this.

6. **Access to justice**

AI could be a low-cost solution to the provision of legal advice to a wider proportion of the population.

7. **Legal judgment prediction**

AI could be used to assess whether or not a case is viable.



Spring et al. (2022) have found that, within the legal profession, the development of AI is varied, but there is currently a focus on high volume, back-office applications. There is other literature that studies alternative applications of AI in a policing context. There are many articles covering the use of face recognition (see Guo & Kennedy, 2023), covering both the technology and ethics aspects of its use. Other areas of research activity include the use of AI in crime prediction, police risk assessment, criminal use of AI and the use of AI in crime prevention.

Ethics

The effectiveness and efficiency of AI in policing can be seen in the management of large datasets - for example, to identify vulnerable groups for prevention purposes, screening vehicle number plates, the use of chatbots for non-emergency requests and using facial recognition tools to match CCTV images to police databases. However, privacy and bias concerns have been raised associated with these uses including the lack of transparency and consent (Berryhill et al., 2019). Despite these challenges, there remains significant appetite for the use of AI tools in policing. As Lewis (2021) highlights, the challenge lies in deploying AI in a way that maintains public trust and confidence while enhancing police services.

The adoption of AI in policing introduces new ethical dilemmas. Dechense et al. (2019) outline a set of six concerns that must be addressed by Dutch police in their use of AI to ensure moral responsibility. They must uphold the high-level ethical principles of: accountability; transparency; privacy and data protection; fairness and inclusivity; human autonomy and agency; and robustness and safety. Similarly, an analysis of AI ethics documents from public, private, and NGO sectors suggests a growing global consensus on five core ethical principles: social responsibility; transparency; bias and fairness; privacy; and safety and reliability (Schiff et al., 2021). Reflecting these concerns, the NPCC's (2023) Covenant for Using AI in Policing has prompted police forces across England and Wales to establish ethics panels and oversight boards. These initiatives aim to maximise AI's potential benefits while assessing risks and ensuring responsible implementation (Oswald et al., 2024).

While AI offers the potential for greater efficiency, the use of machine-generated statements raises critical concerns. LLMs are being developed to improve the quality of police reports, but Stanley (2024) questions their training processes, arguing that bias and a lack of transparency in statement generation could undermine their credibility. These concerns are particularly significant, as they could affect the legal status and admissibility of AI-generated statements in court. The software for AI-generated police statements is still in development, with early implementations in California and the Netherlands raising concerns over inadequate safeguards (Murphy & O'Brien, 2024). Adams (2024) emphasises the ethical necessity of officers verifying AI-generated reports to ensure accuracy and reliability, reinforcing the importance of maintaining

human oversight. This also allows for the inclusion of non-verbal, observational data that AI may overlook.

As public awareness and apprehension about AI's impact continue to grow (Hogg, 2024), it is crucial that ethical principles are translated into policing practice through rigorous evaluation of the use of AI rather than being shaped primarily by industry interests.

Official policy documents

UK Policing already has guidance on the use of AI (NPCC, 2023). The guidance brings together the Digital and Science & Technology strategies for policing to articulate the ambition for policing to develop AI, addressing the following objectives:

1. To drive policing efficiency.
2. To counteract the benefit of AI to criminals who might wish to exploit its capabilities.
3. To maintain public confidence through standards, ethics and oversight.

Built from other sources, the NPCC sets out the following principles:

- Lawful use: The use of AI must comply with legal and other regulatory standards.
- Transparency: The public must be made aware of the use of AI, it can be scrutinised, and the workings of the AI should be demonstrable.
- Explainability: Users must be able to explain how a decision or recommendation was reached.
- Responsibility: There must be responsible use of AI and where outputs are undesirable, the use must be stopped. This includes where there is a risk of bias or harm.
- Accountability: Appropriate governance systems must be in place. There is detailed guidance on Governance issues.
- Robustness: Data used, e.g. to train AI, must be accurate and reliable.

The College of Policing (College, 2024) has also produced guidance on the building of AI tools. The first section of the guidance concerns the “use case” and suggests promising areas for the application of AI, including:

- Automated transcription
- Automated document redaction
- Synthesis of complex data sets
- Enhanced search
- Support for digital enquiries



Ethics issues are explained carefully. Compliance, risk and stakeholder engagement are all explained as part of the process of development. Arguably, one of the most pertinent issues is the advice on commissioning where they request due diligence to avoid duplication of development activity. One particular recommendation is:

“When deciding what new tool/system to commission, forces should always try and build on what has been tried and tested elsewhere.”

The document also provides detailed advice about testing and evaluation of AI developments, providing an evaluation toolkit. There are some clear warnings about bias, especially where a tool replicates disproportionality.

Section Summary

This section has provided a highly selective summary of the literature focusing on the use of AI in productivity applications. We are using this to inform our methodology and have identified that the technical issues of hallucinations, errors and non-audio content of witness interviews are important elements to consider. The focus on ethics in much of the literature also provides a key dimension to consider. Many of the ethics are probably less relevant to highly transactional processing of information when compared, say to the issues of facial recognition or some decision-making tools. We will be aware of the potential for ethics issues, especially in relation to the impact of AI on witnesses and users.



Methodology

A mixed-methods approach is used to complete the evaluation. Three separate types of assessment were used:

1. Semi-structured interviews were held with users, police stakeholders and developers of the AI to obtain an understanding of their perceptions of the utility of the application.
2. The quality of the reports was assessed using well-established methods of understanding report readability and linguistic style. This will be explained in the relevant section below.
3. RVR performance data was provided by the development team, so that the research team understood, as much as possible, the likely impact on call duration and outcomes.

It should be noted that the project was on “pause” at the time of the study (December 2024 to March 2025) and so we were restricted to assessing the output from version 1 of the ADA. This meant that sample sizes of documents to assess were small and these samples were documents that were either test transcripts not related to a real case or a small number of reworked transcripts from cases that had already closed. There were no samples from cases that were passing through the Criminal Justice System. As such, only a few measures had statistically significant results, but they can be used to inform future assessments on what features to assess when considering the next generations of AI output.

Interview Methodology

The internal evaluation team identified a sample of interviewees. This included senior officers so that the context of the development could be understood, in relation to force strategy and objectives. Questions for the evaluation are largely based around the approaches recommended from these sources: the College of Policing guidelines (College, 2024), the NPCC Covenant for the use of AI (NPCC, 2023), 45 questions around the ethical use of AI (Lewis, 2021) and methodology from Ferguson (2024). A representative of the ADA developer company, Anathem, was also interviewed about the future direction of developments and their perceptions of the existing impact on productivity.

Additionally, we interviewed three officers who have used RVR and ADA so that we could establish the practical impact of ADA. Included in the list were other team members who were responsible for implementing ADA and having to adapt systems and procedures to accommodate the innovation. This included services such as IT security personnel. We have provided further details of the semi-structured questions in appendix 1.

Report Quality Methodology

Report quality was formally analysed in two ways:

1. Microsoft Editor was used to provide grammatical, spelling and readability scores for sample witness reports.
2. Linguistic analysis of a sample of RVR incidents was conducted using “Biber Tagging” software.

Throughout this section and the rest of the report we refer to AI-generated reports as AI reports and those that didn’t use AI as “human” reports, purely for reasons of conciseness.

In both above cases sample statements from the pilot work were assessed as stand-alone pieces of analysis. A small sample (4) of real historic cases were taken where the original report was compiled without AI assistance. The AI can be retrospectively applied to generate an AI version of these statements for comparison. A further sample of eighteen randomly selected AI or human statements taken from the initial test transcripts were put through the same analysis.

Grammar, spelling and readability

Standard word processing programmes, such as MS Word, now contain embedded error identification and correction tools. These are extremely useful for comparison purposes when studying the differences between AI and human text. In this case MS Editor was used to produce its overall formal Editor Score (a percentage for each document) and routine counts of spelling, grammar, clarity, conciseness, formality and punctuation errors. The same feature can be used to generate further document statistics, including the mean length of paragraphs, sentences and words. This data is then computed into two measures:

1. **The Flesch Readability Score**

This is a simple formula that was developed over 75 years ago (Flesch, 1948) that looks at the average sentence length and the number of syllables per word. The score typically ranges from 0-100, with a score of 60-70 regarded as “plain English”. Although not widely used as an absolute standard, a readability score of 45 or over is a legally minimum standard for some US insurance policies.

2. **The Flesch-Kincaid Grade Level Score**

This measure uses the same data, weighted differently, to match the readability of text to (US) school grade levels. Table 1 provides a summary of how the data is interpreted and what type of text this compares with in practice. This measure will give an indication of any differences in AI and human text.

Table 1 The Flesch Reading Ease Grading System

Score	Grade	Words per sentence	Syllables per 100 words	Example
90-100	5	8	123	Learn to read books
80-90	6	11	131	The Gruffalo
70-80	7	14	139	Harry Potter
60-70	8-9	17	147	
50-60	10-12	21	155	Jurassic Park
30-50	College	25	167	A Brief History of Time
0-30	College Grad	29	192	An academic paper

Linguistic Quality

It is generally believed that AI applications such as chatbots proficiently substitute for human text, although the length and complexity of such conversations will be much shorter and simpler in nature when compared to the more detailed witness reports assessed here. The aim of this analysis is to understand any differences between linguistic styles of AI and human text. In this case the text output is created for a specific purpose and has a set of rules that are legally required. The statement must be in the witness's own words and hence use first person grammar.

Here we use the methodology proposed by Sardinha (2024) to assess the linguistic qualities of the generated statements. This method builds upon the work of Biber (Biber, 1988; Friginal & Biber, 2016). This work is based on the idea of a linguistic register, whereby the style of writing is influenced by the context or purpose of the text being produced.

Language choices follow norms that correspond to the register. AI applications do seem to be able to emulate some types of register, for instance when being asked to write poetry. The software used in this study uses six dimensions or measures of linguistic style (Nini, 2019):

1. Involved versus Informational Discourse

This measures the density of the information within the text. Academic prose, for example, is informationally dense whereas casual conversation contains less information.

2. Narrative versus Non-Narrative Concerns

This is a measure of whether there is a story contained within the text. These are likely to include past tenses and third person pronouns "he stole her phone..."



3. **Contextual Reference**

This is a measure of whether the context is an important part of the text, such as narrative of a sport event. Formal reports tend to use wordy and complex phrases in order to be precise. Accessible conversations tend to use simpler phrases and shorter words.

4. **Overt Expression of Persuasion**

The use of words such as “must”, “should”, “could” etc. (modal verbs) indicate that an author is expressing a personal opinion, commonly seen in professional communications.

5. **Abstract versus Non-Abstract Information**

This measures whether a statement has active clauses. A statement such as “someone stole the phone” is active whereas “the phone was stolen” is passive.

6. **Informational Elaboration**

This measure judges whether additional information is provided, often within a short section of text. An example would be “The man, *who was wearing a mask*, smashed the window.”

The variations in these dimensions lead to a set of twenty-three separate styles of writing, referred to as registers. Examples of registers including press reportage, academic prose, face-to-face conversations and planned speeches. The software used, called MAT, takes either a piece of text (in a word .txt document form), or a folder containing multiple documents, and grades this for the register.

Numerical scores are achieved across each of the above dimensions, where each can be assessed for the best fit for linguistic style. The combination of scores also provides an indication of best fit of overall style.

There is much commentary in the media about the use of the “Turing Test” when assessing the effectiveness of LLMs to replace human written text. The original test proposed by Alan Turing was concerned more about AI being seen as intelligent as humans, rather than simply mimicking human speech or writing. We are not concerned about either interpretation of this test. If AI produces an acceptable, but recognisably “AI-assisted”, output it may not matter. The question is whether or not the linguistic qualities meet the needs of the witness statement without excessive correction by officers.

Performance Data

Given that the pilot work was on hold at the time of the data collection for this report we were unable to compile any performance data in real time. We are having to refer back to data collected in the earliest stages of the pilot work and have to rely on the accuracy of the data collection process. The collection of performance data around RVR has been thorough since its adoption, due to the automatic collection of call data statistics,



and we can state with some confidence the impact that RVR has on productivity. However, there is no fully documented comparisons of RVR vs RVR+ADA performance. Only a small amount of data is available for the testing of version 1 of ADA. The main data we can use is the aggregated call times for RVR calls, split between ADA and non-ADA calls. The sample size of actual ADA incidents is too small to make any hard judgements about measures of time saved in writing reports. We are entirely reliant on the qualitative accounts from the officers using ADA to provide an insight into efficiencies. It should be noted that the call durations and times taken to write reports are extremely variable due to the large differences in the content, complexity of the cases and the number of offences that may be reported in one call. Hence, small variations in the data would not be statistically significant.



Results

In this section we divide our findings into three parts, based upon the three methodological approaches. We devote considerable time to presenting the outcomes of our user interviews. We present sample data from the accuracy and linguistic analysis separately. Finally, we report the performance data that was provided for us.

Interview Findings

To address the research questions including the advantages and disadvantages of using ADA, the quality of AI-generated statements and ADA's impact on productivity and workload, we present the qualitative findings from the interviews by the following themes:

- Officer involvement, both in terms of leading the interview and in checking the ADA-generated statement
- The narrative included in the statement
- How non-audio content was captured
- How the use of ADA may have impacted on engagement with the witness
- The perceived efficiencies and cost savings when using ADA.

Reference to “user” relates to those interviewees who used the ADA software as part of the evaluation.

Officer involvement – leading the interview

For human and AI generated statements, officer involvement includes leading the interview with the victim and then writing or editing the statement before asking the victim to sign as an agreement that it reflected their account. We explored how the use of ADA may influence or change the ways in which an officer is involved in the process of interviewing and statement generation.

Most of the officers we interviewed commented that knowing that ADA was running in the background allowed them to focus more on their interaction with the victim rather than having to concentrate on typing notes. The following sections focus on engagement and non-audio content to expand on how the use of ADA may impact on statement quality.

Our interviewees had a good understanding that ADA was a tool, and that correct and clear information would need to be put in to get a quality output. One officer explained that they had to use their experience to “steer the conversation into the relevant points,” rather than let the victim “ramble on,” to ensure that ADA picked up on the relevant points. Although mindful of the information that ADA needed to capture, this officer was clear that using the software would not impact on how they spoke to the victim:



“The thing to remember with ADA is it doesn't write statements for you. It's a statement writing tool. You know, you kind of incorporate it within the statement writing process... in terms of how I speak to people would remain the same.”

However, another officer explained how using ADA changed the way in which they structured their interview. With knowledge of the five-part structure of the statement and what information ADA would need to populate the report template, the officer reordered their questioning to fit this:

“Because I knew it was capturing everything, I changed the way I spoke to people a little bit in terms of the order I'd go through things. The victim wouldn't necessarily notice that... I'd structure my interaction with the victim a bit like taking a statement. So in taking the details for the report, I would ask questions in the IPAC statements format. So just getting an introduction of them, all the people that are going to be involved, all the places. Then getting them to give me an account where I'd ask questions, digging a bit deeper into the information.”

Whilst this helped to reduce the time required to edit the statement, others raised concerns that the approach may become “quite robotic” and that it might limit the amount of information that the officer gets. They commented that the first priority should be for an officer to build rapport with the victim as this often leads to greater trust and disclosure. They emphasised that the quality of the report is determined by the quality of the interviewing technique - asking the right questions and developing the skill of follow-up questions designed to elicit more evidence or understanding about the case.

There was a consensus among those we interviewed that only experienced officers should use ADA. This was because experience would ensure that officers could build rapport with the victim and know what evidence they need to support the “points to prove” associated with a particular offence. One officer quantified this commenting that “no officer under two years would be using ADA.” This officer was clear that ADA is not going to replace an experienced officer:

“It takes an experienced officer to use ADA... You know, you can't have a probationer that uses ADA and, you know, think that the quality of an experienced officer is going to come out of it... I'd say you have to have an experienced officer using ADA to get the output of an experienced officer.”

Another officer expressed concern that those using ADA may presume that it can do more than it is programmed to do which might reduce the quality of their questioning. The statement was generated after the interview had finished (in one report stated as being between three to ten minutes) so the officers could not see any of the statement content in real time to check what had been captured. This officer believed that using AI will “make people lazy” and could “de-skill” officers:

“We're removing the ability for them to interact, engage and ask testing questions. If the point at which you correct a statement is after an AI has written it, do you ask the right



questions? ...when I'd write a statement, I'll be writing it and I'll be asking those questions. You've noticed where you might want to expand, and you have that to and fro."

This officer was clear that with the software in its current stage of development, "the police officer produces the best witness statement."

Officer de-skilling was also raised as a concern by a non-user interviewee in case the technology failed:

"I think before an officer uses ADA, they should be fully competent... We know what technology is like. If technology is not there, then we're going to have to go back to writing statements 100%. We need officers who'll always be able to take a statement - as good in person as what ADA produces or of a standard that makes sense."

Officer involvement – checking the statement

Although most officers identified ways in which ADA sped up the generation of statements, all officers emphasised the importance of keeping the "human-in-the-loop" when checking and editing statements before sending to the witness for agreement and sign-off. Early iterations of the first version of ADA which one officer described as "a very raw product with quite a lot of errors," led to some problems in the generated statement, many of which were quite easy to spot. For example, the software initially picked up on background noises and had to be trained to focus on just one voice at either end of the conversation. It initially replaced swear words with a series of asterisks which was not appropriate for the specific purpose of a witness statement as this detail should be included for context. Hallucinations where words that had not been spoken by the witness appeared in the transcription. A record of errors kept by users of ADA, noted that this included random words including "Akhilesh Yadav." These early issues had been fed back to the software developers and had been resolved quickly. However, following these early fixes, other errors and problems began to emerge which the developers traced back to the re-coding associated with the initial fixes.

Learning how to use the AI software was described as a "straightforward process" and one of the more user-friendly pieces of technology that had been introduced into policing in a while. However, despite its easy functionality, practise using ADA was identified as important in understanding how it generated statements from the transcript. One officer identified the value of having designated officers working in the RVR team so that the cumulative understanding of using ADA was not lost. One officer who had used the software extensively highlighted the importance of having the opportunity to "use it and learn it" which was compromised by staff turnover:

"Officers come here for three or six months and then they're going back out to shift and they're replaced by someone else. So currently, you've got officers who, you know as they're kind of getting into the prime of recognising how to use ADA and get the best



results out of it and then they're picked up and put back out on shift where they don't use ADA and someone else has been put in their seats.”

Our interviewees highlighted that when checking the veracity and quality of a statement, an officer's priority was to ensure that key evidential content linked to an offence's "points to prove" was included. They reported that this information was not always picked up by ADA in more complex situations or where multiple people and places had been mentioned. In part, this had led to a request from those who had used ADA to build additional functionality into version 2:

“So the points to prove. The point of that was simply just once the officer has decided what offence it is, they would click that the points to prove would be generated and it would take the information from the transcript and say this covers this point... this covers this point...this covers this point. So at no point is it making any suggestions about what offence has been committed.”

This officer emphasised the importance of making it clear to the software developers that the officer must be the one making the decisions on what offences have taken place and that the software was not involved in the decision-making process.

When asked whether this functionality might lead to statements becoming some sort of checklist, most interviewees did not believe this would be the case for statements. However, one officer reflected on a meeting with those who had used ADA where a checklist had been suggested for the summary of the statement:

“There would almost be a screen where we could, it would almost have things almost like a checklist where we would select the relevant people, we would almost like select the relevant information and it would be very much like we'd go down and we'd almost like tick boxes tick off what we thought should and shouldn't be included.”

An important aspect about the quality of an AI-generated statement which was mentioned by several of our interviewees is the extent to which it is a true reflection of the witness's narrative and reported in their "voice". This is discussed in more detail in the next section.

Narrative

This section refers to the way in which the statement reflects the witness's account of the incident and the way in which they recounted this. A related section on non-audio content and how this is integrated into the statement follows.

Many of our interviewees highlighted the importance of ensuring that the final statement reflected both the witness's account and their "voice" including the choice of vocabulary and the witness's phraseology. This is of particular importance due to the legal status and the role of the statement in court. One interviewee described the statement as "the most powerful piece of evidence to any criminal case." Several of our interviewees referred to words that ADA had generated and that the witness had not said nor were likely to use:



“On one or two occasions it may be used vocabulary that I don't think is consistent of what the victim would have used.”

An example of this was the word “insidious” which the officer knew had not been said during the interview. Officers noted that whilst an ADA-generated statement might accurately reflect the meaning of the witness’s account, apparent differences with the way the witness speaks in court could compromise the perceived veracity of the statement.

Officers reported that ADA did not always accurately capture details of complex situations where witnesses mentioned lots of people and places. There was agreement between interviewees that there might have been a benefit in starting the software trial with more straightforward crimes as domestic abuse cases were often detailed and complicated. However, this area was chosen as “the best option...for the purpose of controlling the testing.” The ADA-generated statement included information from the transcript despite its lack of relevance. This was referenced by our interviewees as “completely immaterial” and which had “no benefit within an evidential document at all,” as one officer explained:

“the AI was unable to differentiate between people that are relevant for the criminal conduct...So, for example, you know, for domestic, you might talk about the offender ex-partner, but you might reference your new partner and that...but actually the criminal offence has nothing to do with the new partner. The AI statement would without fail, introduce every single person that had been referenced in the call and then it was completely unable to differentiate people. Same with locations.”

Some officers explained how it would take time to delete or edit irrelevant information to promote accuracy and focus. They also reported that they had to edit timelines as ADA did not always accurately record these coherently. For example:

“They would jump around, so it wouldn't provide consistently a narrative from A to B to C to D to E. It would kind of jump between it would try and group offences together, but in doing so it would jump around. Actually, in terms of readability, you know from reading as a criminal conduct case, it didn't flow in the way that we would expect or that would make logical sense either to us, to the CPS, in terms of how it was structured.”

Reflecting on the editing process, interviewees commented on ADA’s lack of sense-making and because “it hasn’t got that human brain,” its inability to differentiate between what was important evidentially and what was just people “rambling on”. An interviewee noted that the skill of an officer is “knowing when to weight certain things” and that “the AI couldn't work out what was important and what was not important”. One officer commented that ADA needed some “guiding steps” to help it identify which parts of the transcript are relevant. This links to the above comments about how version 2 of ADA could help to identify relevant text to support the “points to prove” for the offence.



Although concerns had been raised about inaccuracies, extraneous information and tone of voice in the ADA-generated statement, all officers who had used the software were assured that an edited version would be of the necessary quality and reflect the witness's account. An officer explained:

“The final statement should never really come out as a misrepresentation because the officer reads it and then the person signing the statement reads it in full so you know it almost goes through two checks to make sure that the things that are in the statements should be in there.”

However other interviewees who were further removed from the use of RVR raised concerns about the verification check to ensure that the use of ADA does not become “wholly automated”. They gave this caution:

“It becomes very easy to stop doing that human check when you become reliant upon the automation and you know it's sound in terms of its security and delivering a good outcome in terms of the individual work that it's automating. So that's something I want to keep a close eye on because we don't want to fall foul of the legislation and also we do need to make sure that where we're using this technology, somebody's still checking the outputs.”

There was consensus among the interviewees that some form of quality control of the ADA-generated statements would be of value if the software was adopted for use by the force.

Capturing non-audio content

We were interested in how ADA made sense of non-audio content from the interviews. This includes observational data about the witness including any injuries, their demeanour or emotional state – particularly if linked to specific content of their narrative. It also includes the witness's environment that could add useful information about their context including items captured in the video's view (whether intentionally or unintentionally placed in view) and the sight and sound of any other people in the witness's setting.

The officers we interviewed deemed any evidence of injuries to be the most significant non-audio content that they would need to capture. In human generated statements, officers note any injuries they observe and ask the victim about these. They might also ask the victim to take a photograph of their injury and upload it via the app used to capture the RVR.

All statements rely on officers verbally reporting on any injuries they could see during the interview or asking the victim to describe their injuries or “list your injuries for me”.

Other officers agreed that they would also take this approach:

“With a product like ADA, you would have to say to them, kind of ‘what injuries do you have? Can you describe it?’ to get that full information. Because if the information isn't said out loud, then the system wouldn't, we'll never know”.



However, one officer reported that they would not phrase their questioning in a way that required the victim to list their injuries as they were not “doing it consciously for ADA” but would embed information about injuries into their questioning to ensure that it was captured in the AI statement. For example:

“I can see you’ve got a bruise. Tell me a bit about how that happened. Is that from the offender?”

If any non-audio detail was missed during the interview, officers explained that they could enter descriptive content into text boxes within the body of the AI generated statement before the MG11 is finalised into the format for court.

“...if there is anything that I feel it [ADA] hasn't picked up on that you know the necessary human element would do, such as you know someone's demeanour, the way someone's presenting, then I can add that information into the statements.”

The function of editing or adding in text after the statements has been generated by ADA was highlighted as particularly useful for officers to add comments about the witness’s emotions - how they presented during the interview or their reflections on how the incident impacted them. An officer explained:

“There's a part in the statement called the VPS - the Victim Personal Statement - where it would be explained about how this event has affected them, and that's where you'd put a lot of that information in. An officer would write in a certain way so the emotion would be able to come across. The AI wouldn't be able to do that because it wouldn't know that the person was upset while giving that information.”

Another officer explained how they would capture affective aspects of the incident by asking specific questions towards the end of the interview including “What impact has this had on you?” and “How has it made you feel?” Although this officer may ask these questions when not using AI, it was a good way for them to ensure that information about a witness’s emotions was captured on the audio.

Engagement with the witness

For transparency purposes, the witness was informed that AI was being used during the interview to transcribe the audio and to generate a statement (there was also a declaration of its use in the statement). Although we have not got witnesses’ perspectives on the use of AI, we did question our interviewees whether statements generated by ADA led to any differences in their perceived engagement with witnesses.

Some officers reported that when using ADA, their engagement with the witness was qualitatively better:

“I was engaging with them more and they were feeling more seen and heard because when I’m using ADA, I can sort of look at the person on the screen.”

Greater eye contact, smoother communication and giving “their full attention to the person,” were benefits that officers reported because they no longer had to multitask



and take notes. This was reflected in a comment by one officer who usually took three to four pages of notes but when using ADA, had less than one page:

“I think that it [ADA] does allow you to focus a bit more on the person that you're speaking to... Sometimes if you've got someone that is talking at 100 miles an hour, like some people do, I feel like a lot of the time I'm taking my notes and I'm always catching up on what I'm typing and possibly missing information that they're saying to me. So I'm having to ask them to repeat things a lot and then you can sort of feel like that person then feels like maybe they're not being listened to necessarily. So ADA definitely helped with that to prevent, you know, so much of that, you know, disjointed communication.”

However, one officer challenged claims about improved engagement which had been discussed among the team who had trialled the software. This officer noted that there had not been any previous problem with witness engagement when taking human statements and questioned whether any witness had commented on lack of eye contact. This officer reasoned that greater eye contact and apparent fuller attention could not be used as a rationale for using AI for statement generation.

Another way in which officers found that ADA helped to promote better engagement with the witness was in the way that it freed up their cognitive capacity. Not having to think about whether they had captured the witness's narrative - and not having to type this out in real time - allowed officers to think more carefully about the questions they asked. One officer reported that their questioning had improved because, “I felt like there was more headspace to be able to do that.”

It also enabled one officer to look at historic information associated with the case and follow up on previous experiences which might have been missed if their attention had been focused on taking notes.

Freeing up cognitive capacity was particularly helpful for one officer who had been diagnosed with dyslexia at the age of ten. This officer reported that a lot of attention was “immediately drawn away from the witness” when concentrating on typing notes. They claimed that that using ADA “definitely let me engage with the witness more,” and thus allowed them to build greater rapport. Another officer commented that they felt reassured that ADA was “catching everything” and therefore was not distracted by concerns that they were missing some of the witness's statement, especially if the witness was talking quickly.

Perceived efficiencies and cost savings

Most of our interviewees who had used ADA reported that the process of preparing statements had sped up. This was due to the developers making early improvements to the software and the effect of user familiarisation with the software. Despite the time taken by officers to edit the ADA-generated statements and to add in text related to observational data, there was consensus among our interviewees that ADA had “the potential” to be a time saving tool for statement generation, particularly for



straightforward cases involving just one offence. Commenting on time savings, one interviewee made this observation:

“Your more of a straightforward ones would be, would be fine...it would depend on how complex the case was, how many victims, how many people would be introducing into the statements, how many different scenes, the more complex it got, the more convoluted it got.”

One officer who had considerable use of the software had seen how the process of ADA-generated statements was becoming faster and more accurate. The officer with dyslexia noted that they had already experienced “huge” time saving with a belief that ADA also improved the quality of their written statement. It was also noted that whilst statement generation could save time and that more statements would be produced per day, there were limitations related to quantity of output due to case fatigue:

“I think I could definitely do more with ADA however there's is a plateau - there's a sort of an emotional strain to dealing with domestics one after the other, all day, every day.”

An interviewee who had not used the software, but worked closely with those who had, commented about whether any time was saved by using ADA:

“I don't think there was much but I think we had to be savvy to the point that we were developing that. So when we were using it, we were still developing it and we just very much saw the potential there.”

One officer reflected on their use of ADA and was clear that it had taken them more time to edit a disjointed statement, particularly when the case was complex or involved a complicated timeline, than if they had written it without AI. The process may also take longer if an officer has a certain way that they like to write statements or write at a particularly high standard. An officer made this observation:

“Editing to ensure it's the voice of the victim and whilst it might be of greater quality than some officer's statements, others wanted to rewrite and ... I'd say in the initial output it might be that the quality of the statement maybe might not quite meet if an officer writes an exceptional statement all the time - possibly the initial output might not meet that standard, but like I said, there is the option for them to go in and change it to reflect their usual standard if needed, and if they feel like they can write a manual statement faster than doing that through ADA. Then you know that those officers might not benefit.”

Interviewees explained that they were waiting for version 2 of the software which should include additional functionality which they believed would promote the accuracy and quality of the generated statement – and thus reduce the time taken to check it. This included a “landing page” to check output before statement generation which would allow officers to check the accuracy and quality of the content and reduce the number of documents generated thus saving on data storage costs. One user explains what they have requested for the next version of ADA:



“The major thing was the accessibility of the officer being able to decide what's important and what's not. So one major thing that we asked for was a landing page. So after the transcription has taken place, the officer would then be able to check the data so things like the people, the location, the times, anything like that before anything was actually generated.”

One officer noted that ADA had greater potential than seen thus far in the generation of the statement (the MG11 is the main focus) and a brief summary (the DARA is not yet up to standard). It could avoid “double-keying” to generate other files associated with the case that officers were required to complete. However, another interviewee explained how this would be “further down the line” as the police computer system currently used by the force would need to be better able to draw data together from different places.

Interviewees agreed that version 2 might reduce time for statement generation and free officers up to respond to the next call. However, time savings overall were questioned given the importance of checking the statement and building in a quality control stage. One officer also outlined how, in its current state, the CPS would not be able to process an increased number of statements, especially if other forces were using similar software - thus it would push a bottleneck further down the sentencing process. An interviewee who had not used the software explained how they could understand the appetite among front line officers for automated statement generation and believed that this would be the norm in the future, however they questioned whether this would lead to time saving across the whole process. They also highlighted the importance of good records management by deleting data when appropriate to save money on data storage. As this was an “intangible” aspect of AI tools, it was not always considered a priority.

Other specific financial costs mentioned in the interviews included licences for the RVR and Anathem’s transcription costs.

Section summary

Overall, there was no concern about the principle of using AI to generate witness statements and there was agreement that automation could speed up specific processes. Interviewee responses indicated that the benefits would mean greater productivity rather than any threats to jobs. In terms of the ethics associated with the use of AI for statement generation, one user of the software noted:

“The AI can't be making any decisions. I think if we if we allow it to start making decisions or suggestions, then ethically I think it's wrong. If all it is doing is reordering information that is set into it, then no, I don't believe there is [any ethical concern]. And in terms of the equality side of things? The use of ADA with the majority of the protected characteristics doesn't have a negative or positive effect. It's a completely neutral zone.”

There was consensus between our interviewees that ADA had great potential to increase productivity, however proof of concept had yet to be established in terms of time saved. Problems with the early iterations of version 1 had been solved quite easily although concern remains about the quality and veracity of the witness's narrative, particularly for complex cases. There is still an appetite for the use of AI and interviewees are looking forward to trialling version 2.

Report Quality Findings

MS Editor Scores

Two sets of scripts were used as a paired comparison whereby an original transcript that already had a human statement was also put through ADA. This allows a direct comparison to be made between the AI and human accounts of the same incident. Scores for clarity, spelling, grammar etc. record the number of errors, so low scores are better than high scores.

Table 2 AI vs Human Edit Scores

	AI1	Human 1	AI2	Human 2
Editor score	98%	92%	95%	96%
Writing	Formal	Formal	Formal	Formal
Spelling errors	0	0	2	0
Grammar errors	1	6	1	4
Clarity	10	1	15	1
Conciseness	1	1	3	1
Formality	2	1	2	3
Resume	2	0	1	2
Sentences per paragraph	13.6	1.7	12.2	2.3
Words per sentence	22.9	28.4	19.4	23.7
Characters per word	4.7	4.1	4.7	4
Flesch Reading Ease	50	61	50.7	64.8
Flesch-Kincaid Grade	11.9	11.8	11	10.1
Passive sentences %	14.7	6.2	24.5	3.5

We used this data to inform our larger study of a wider range of AI and human statements. Fourteen AI and eight human statements were individually assessed using MS Editor and then the scores were put into corresponding data sets for t-tests to establish any significant differences. The main findings are in table 3 below.

Table 3 Main differences in report quality

Measure	Mean		Standard Deviation		Significantly Different? Y/N	P value
	AI	Human	AI	Human		
Grammar errors	3.86	8.75	3.01	4.92	Y	0.0086
Concision errors	1.24	3.75	1.38	3.92	Y	0.043
Words per sentence	18.7	26.8	2.96	11.0	Y	0.016
Word length	4.62	4.1	0.14	0.18	Y	0.0001
Reading ease	53.1	63.1	4.56	10.66	Y	0.0061
Grade level	10.4	10.99	1.25	4.25	N	--

The results indicate that human statements use shorter words in longer sentences, but this creates a more readable script. The human scripts contain twice as many grammatical errors as the AI scripts and are less concise. There was no measurable difference between the types of script in terms of the grade level. We also noticed that the paragraph lengths (measured by number of sentences) were very evidently different, but this was explained in two ways. First, given that the human scripts had longer sentences, authors tended to break up the text more frequently. Second, we believe officers writing scripts had a different protocol for formatting the script, with a greater tendency to divide sections of the testimony up into smaller chunks.

We have not reported the spelling error scores in the above table. We found that most spelling errors were due to the use of US spelling rather than mistakes. This can be easily corrected.

It is also worth noting that the Standard Deviation of all of the measures is consistently higher for human scripts. This should not be surprising because the scripts were written by a range of different authors, and a variety of outputs is to be expected. It would be worth conducting further study of the quality of reports based upon the authors' levels of experience in report writing.

Linguistic Analysis Outcomes

Most of the linguistic analysis was conducted on two pooled sets of scripts one of which was the collection of all AI generated scripts and the other as the Human Written

Scripts. We can therefore report the styles of each corpus of work, but we also looked at the variability by studying the outcomes for individual scripts.

For the individual scripts we studied the linguistic styles reported across all of the six dimensions. To convey the sense of which styles were most prevalent we studied the frequency of reporting of the styles from within the data set. Figure 1 shows this result for both types of script.

Figure 1 The Frequency of Register Labels (% of sample)



The figure shows that “academic prose” was cited as the style six times more frequently in the AI texts than in the human texts. As a general trend, formal styles of writing were more common on the AI texts. We note the informal styles of general fiction and personal letters featured more frequently in human texts. The Conversation register was not used at all by the AI statements.

These findings are also reflected in the summary styles for each corpus. Table 4 shows the styles reported for each dimension and the overall style reported.

Table 4 AI/Human corpus comparisons

Dimension	AI	Human
1. Involved	Broadcasts	General Fiction
2. Narrative	General Fiction	Prepared Speeches
3. Context	Academic Prose	Press Reportage
4. Persuasion	Personal Letters	Personal Letters
5. Abstract	Press Reportage	Press Reportage
6. Informational	Academic Prose	Conversations
Combined 1-6	General Narrative Exposition	Imaginative Narrative



The table does indicate some marginal trend towards formal writing by AI but there are some ways in which these are similar. For example, both sets of texts avoided a scientific style of reporting information (category 5) and instead presented information as if it were a press report.

We then conducted tests for statistical significance across the same analysis, incorporating all sources of statements into this analysis. Table 5, below, shows the box-whisker plots of the corpus data and the commentary of the similarities and differences. It provides the evidence of just two statistically significant differences in the styles of text. The AI statements appear to have information packed more densely. The conversational style of human reports is significantly different. The other main difference is in the style of addressing the situation or context. Human texts are written fully in the context of the domestic abuse incidents and potential offences. The AI texts seem to treat this more as a context-independent presentation of facts.

Table 5 Statistical differences in linguistics

<p>Dimension 1</p> <p>Dimension score</p> <p>1</p> <p>AI Human</p>	<p>Involved vs Informational Discourse</p> <p>There was a measurable difference between the scripts by this dimension, with little overlap in the individual scores. The AI script contained packed information that was formally presented. The Human scripts were more conversational, but there was a wider range of scores within the sample.</p> <p>Outcome: Different $p=0.0005$</p>
<p>Dimension 2</p> <p>Dimension score</p> <p>1</p> <p>AI Human</p>	<p>Narrative</p> <p>Given that statements are supposed to be written in 1st person, you would expect scores here to be low. However, the statements are unusual in that they are historic accounts (past tense), which affects scoring. There is no significant difference here. The human scripts exhibited wider variation.</p>
<p>Dimension 3</p> <p>Dimension score</p> <p>1</p> <p>AI Human</p>	<p>Context</p> <p>It is clear by these scores that AI focuses less on context than the human authors. This is one area that is consistent with critique of AI – that the AI does not reflect the context or purpose of the statement.</p> <p>Outcome: Highly significant difference $p = 0.0001$</p>



<p>Dimension 4</p> <p>Dimension score</p> <p>1</p> <p>■ AI ■ Human</p>	<p>Persuasion</p> <p>Human and AI scripts were similar each other is this respect. High scores represent a style that is expressing a personal point of view. It is perhaps surprising the scores aren't higher.</p> <p>No difference in the scripts.</p>
<p>Dimension 5</p> <p>Dimension score</p> <p>1</p> <p>■ AI ■ Human</p>	<p>Abstract Information</p> <p>Neither set of scripts presented information in a style that represented scientific information. This is probably the correct outcome.</p> <p>No difference in the scripts.</p>
<p>Dimension 6</p> <p>Dimension score</p> <p>1</p> <p>■ AI ■ Human</p>	<p>Elaboration</p> <p>The scores in this dimension were low, indicating that time was being taken to explain events slowly and carefully.</p> <p>There was no statistically significant difference in the scripts.</p>



Performance Data Findings

Hertfordshire Constabulary use Power BI to monitor the performance of the Force Control Room (FCR). The Power BI data we present in our results have been supplied to us by the FCR, not collected by us.

Over the two years 2023-2024, Hertfordshire Constabulary received 40,558 calls classed as Domestic Abuse. Of those 40% were classed as “immediate” and officers would be sent to the scene of the incident. RVR operators dealt with 5,693 incidents over that time. Each incident attended takes on average 4 hour and 18 minutes of officer time. The average RVR call uses 43 minutes of officer time.

Only a small proportion of the RVR calls actually require a witness statement to be generated. 70% are managed within RVR without a statement required, with 11 % of the calls requiring a statement of some sort.

Investigations saw an increase of 6 minutes during the GoodSAM call, additionally the time to complete admin increased by 48 minutes when using ADA. Non-Crime calls saw an increase of 2 minutes during the GoodSAM call, additional the time to complete admin increased by 12 minutes when using ADA. Hence, presently, call duration was 13% longer when ADA was used. The expected increase in workload per incident when the above figures are combined is 18.5 minutes if ADA is used.

The PowerBI data records monthly averages of how many incidents an officer will complete in a shift. This average ranges from 2.37 to 3.71 incidents per shift. Hence, the use of ADA would add between 43 and 67 minutes of extra work to a shift. (The number of incidents completed per officer in a shift is therefore likely to fall.)



Discussion

Defining output quality

The results we present here raise the question of how we define quality when assessing the output from this type of knowledge-based process. In this case the requirements go significantly beyond the production of an adequate summary of text or information. This output has to tell a clear story in the words of a witness and also create the right evidence to pass the basic tests that would lead to a useable prosecution file. In this case we suggest the following quality dimensions:

Context – the output has to be clear about the overall background, including who the people in the account and the reasons for the case.

Narrative – the output has to tell the story of what happened, focusing on the relevant aspects of the original account and putting these into a sensible storyline.

Errors, omissions and hallucinations. In the context of AI-generated outputs the biggest risk is that the AI will invent evidence, some of which is plausible unless detected and removed. We cross-checked the statements made by interviewees with hallucinations recorded in the development error log. The following samples are hallucination examples from this error log:

“she is a woman of average height and brown hair.” (no description given)

“He is a tall man with a stocky build and has a known history of violence, although he has never been violent towards me.” (not in the interview)

“The incidents occurred at various locations, including my aunt’s house, which I used as a safe place when I became homeless after the split. The house is in a quiet residential area, and I stayed there to avoid Amy. The weather was clear on the days of the incidents, and there were no obstructions to my view when I witnessed Amy and Jordan outside the house.” (none of this detail provided)

“he would “jump me.” This threat occurred approximately two weeks ago, around 18:00.” (no time stated)

Even then, these errors are “discoverable” and could affect outcomes. Hallucinations, whilst present at the start of the work, were more easily prevented as time passed. The interviews stated that it was common for information to be omitted or overlooked. Hence, one quality dimension is the completeness of the report, perhaps including all “points to prove” as an essential component of a witness statement. The presence of hallucinations does in part rely on the memory of the interviewer, who might not notice that an additional piece of detail has been added.

Readability – the account has to be written in a style that is both accessible and representative of the words of the witness. The AI seems to provide a more consistent but less representative output.

Technical – The output of AI scores well on technical errors, but higher levels of grammatical errors in human statements seem to matter less. Other measures such as spelling errors showed comparability.

Comparing AI & Human Statements

Table 6 below provides three anonymised and disguised sections of AI and human statement text taken from the same witness interview. These sections illustrate some of the issues raised by our results.

Table 6 AI and Human Text for the same statement components

AI Text	Human Text
On a Sunday morning, JOHN DOE physically assaulted me by attempting to strangle me. This incident occurred at our home and was provoked by his anger towards me for locking my phone, which he demanded access to.	On Sunday 18th February 2024 I told JOHN in our home and told him I was leaving as I couldn't take this anymore, JOHN had asked for more money and when I refused he got aggressive, he tried to take my phone from me and pinned me down by putting his hands around my neck.
During the assault, I did not resist, fearing that any resistance would result in more severe violence. Thankfully, I did not lose consciousness or struggle to breathe as he ceased his attempt to strangle me of his own accord.	I didn't fight JOHN off as I know when I try and resist he gets angrier, JOHN didn't squeeze and I didn't have any trouble breathing the incident only lasted a few seconds and JOHN released his hands from my neck.
When confronted about his financial actions, JOHN DOE threatened to commit suicide. On a particular Saturday night, he acted on his threats by taking a large number of tablets in an attempt to harm himself. During this incident, I called emergency services (999) for assistance, but JOHN DOE took the phone from my hand and disconnected the call. He then made himself sick, which resulted in the expulsion of the tablets he had ingested, and he did not suffer any lasting harm from this attempt.	On Saturday 17th February 2024 JOHN attempted to take his own life at our family home by taking tablets I called 999 but he knocked the phone out of my hand and then made himself be sick.



The interviews highlighted the problems that AI text tended to use words that witnesses would not use in statements. They were also concerned that the narrative did not generally reflect how witnesses would explain incidents. The linguistic analysis does confirm much of these concerns because we did find that AI uses longer words and tends to be more academic in style. The comparison of the above AI and Human statements allows us to illustrate these concerns. There are several observations about this text comparison that can be made

- The human written statement is more specific about the dates of when incidents have occurred.
- The AI version choose longer words such as “consciousness”, “expulsion” and “resistance”.
- The AI uses “ceased his attempt” instead of choosing to say something like “stopped” as an example of the difference in writing style. The phrase “of his own accord” is out of place/unnecessary.
- The accounts offer slightly different views about what happened. For example, the officer recorded the phone being knocked out of the witness’s hand. The AI suggests the phone was taken and the call disconnected. This type of detail is important.
- The AI account of the taking of tablets offers unnecessary explanations of the event.

We conclude that the results from the linguistic analysis validate the concerns of the officers using ADA in regard to the quality of the AI generated reports. It is also worth reflecting on how much of the AI text a diligent officer would want to re-edit so that it presented more accurate and realistic account of the incidents. We suggest that this example could have half of the text reworked to make it more reflective of the person’s account. Under time pressure would the AI text be checked to ensure the account of the phone incident was accurately written?

Productivity

The impact of ADA on productivity was one of the key questions that we have attempted to address in this study. The initial user interviews gave us a wide range of opinions ranging from significant reductions in time taken to generate reports through to easier to do it manually. The differences were largely based on the range of abilities and levels of experience of the officers. We believe that the time taken to *generate* draft reports is much faster, but this is partly counteracted by a significant increase in the time needed to *check* the reports for errors. We are concerned that an additional and highly significant extra step will be added later, before the file handover stage, when other officers need to go through these reports more thoroughly to ensure no errors are passed to the CPS and further into the system. The report that RVR call duration is



longer when using ADA is an important consideration, but we believe that longer calls with witnesses could be an indication of greater care being taken to obtain all relevant information before the report is generated, rather than a negative impact on the process and its productivity.

There is also extra work created when using AI because of the need to disclose all versions of the documents generated and provide additional documentation. Where cases go to court there is a need to disclose the existence of the first version of the transcript. If amendments are made to the transcript, that version needs to be provided so that defence and prosecution can see what changes have been made. There is likely to be a minimum of three additional documents that the Officer in the case will have to notify CPS of, then redact and supply to them on request.

Any future assessment of later versions should also consider how any savings in officer time are realisable, either through fewer numbers of officers needed within the RVR team or through opportunities to do other work when a case has been cleared. We also raise the potential for “work intensification” if officers are expected to quickly generate reports and move on to new demand under the pressure of a less well-resourced department. To counterbalance this we appreciate that, for many officers, the report writing phase of an incident is the least liked aspect of the job and supporting technology is likely to be popular at some level.

In many situations, practice and experience of performing a task leads to better quality and productivity. Where officers work in RVR/ADA for short periods of time there could be fewer productivity gains, due to the learning process and then leaving when fully productive. It is unlikely that people would want to stay in RVR on a longer-term basis so more learning/gains will be made if ADA is used more widely outside RVR, with officers using it regularly.

Ultimately any productivity assessment needs to include the whole journey of an incident from the initial call through to final outcome. Within the data presented in the previous section we saw that the proportion of DA referrals to the RVR unit that became cases for referral to the CPS was about 11% of the total call demand. It must be factored in, therefore, that only a relatively small proportion of the overall demand currently would be impacted by ADA. We have not addressed any other downstream impacts on workload or productivity as a result of this change. The following questions also need to be considered in future assessments of workload impact and productivity. As a result of changes to the quality of reports and file build:

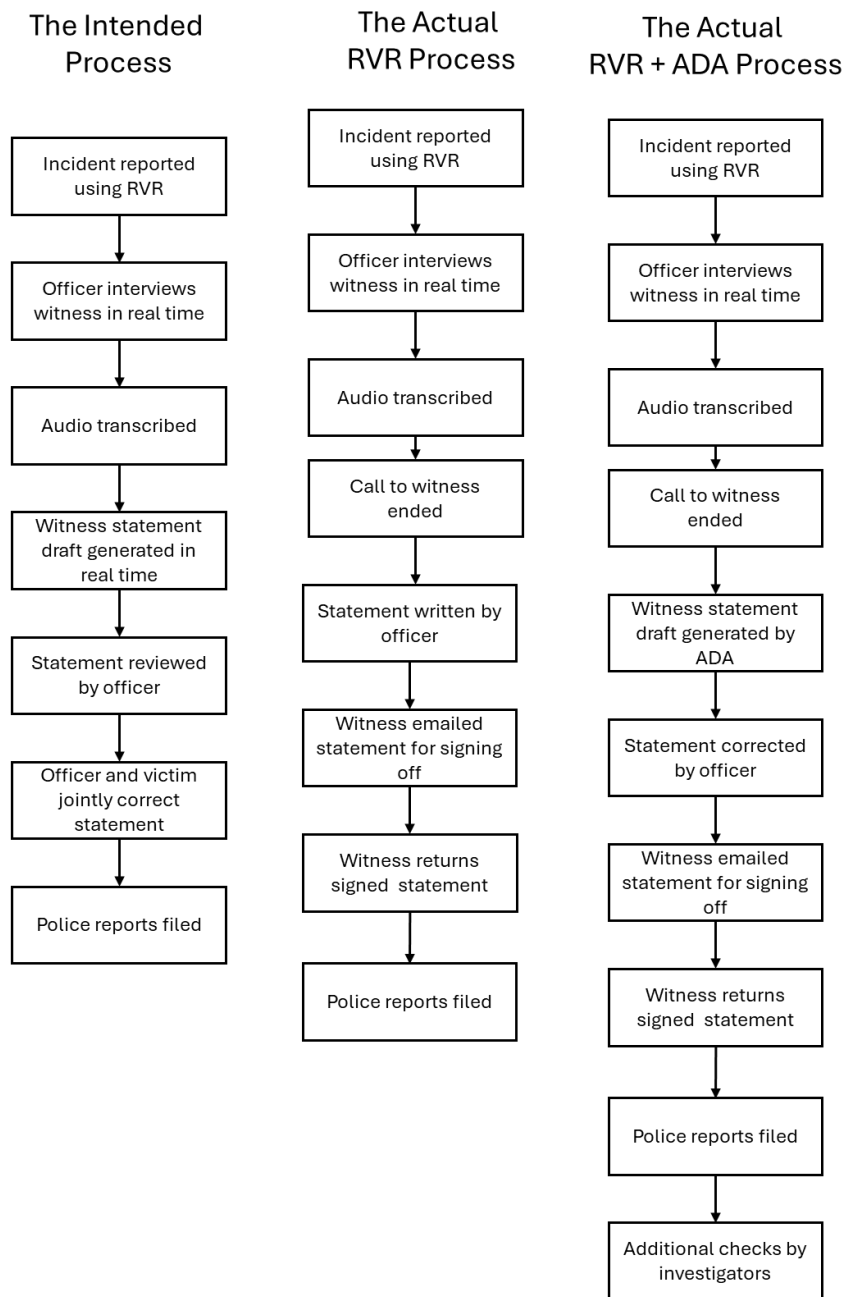
- Is there a significant change to demand or workload within the following investigation process?
- Is there a change in the number of files being sent to the CPS?



- Is there a change in the workload of returned files from the CPS as a consequence of AI-written reports?
- Does the number of prosecutions increase?
- Is there a change in the proportion of “guilty” pleas, thus reducing other workloads?
- Is there a change in the time officers spend in trials because of any changes to the quality of the file build and prosecution cases?

We received some suggestions that an additional checking process was taking place after the initial file build, that was not previously present. If it becomes the norm to perform a second downstream check, then the workload here does need to be factored in. We were provided with the official process maps of how the system should work but these did not fully reflect what we found. Figure 2 shows the difference between the intended process and the process as described in interviews.

Figure 2 A comparison of intended and actual processes



Development Costs

We noted that Hertfordshire Constabulary have put in a significant amount of officer time to help develop the early versions of ADA. Figures provided to us based around time sheets suggest the equivalent of two FTEs of officers (across a range of levels of seniority) have been needed to develop ADA this far, not including senior officer supervision. To bring ADA up to a standard where this is useable there will need to be more investment in time before the full benefits are realised. There are multiple developments of similar AI applications elsewhere and also parallel developments of



ADA functionality in other forces. As the College of Policing point out, there is a need to avoid unnecessary duplication of effort where possible. Presently we feel that the race to achieve working versions of AI reporting tools, with competition between forces and between AI companies, risks absorbing a considerable amount of police resource unnecessarily. The ultimate business model of what savings are realised, how these are achieved, and the external costs of data processing and storage need to be thought through carefully.



Key findings

The effectiveness of ADA

- ADA has so far produced witness statements with an acceptable standard of readability and few spelling or grammar errors. It does not currently produce work that matches the average standard produced by the police officers we compared ADA with.
- The linguistic style of ADA is currently more formal or academic in style than statements written by officers, but this might be improved over time.
- A key weakness in any AI, demonstrated by ADA, is the limitations concerning the understanding of the context of the output. The narrative produced by the system is not of a good standard compared to existing policing standards.
- The context of domestic abuse cases was an interesting choice of starting point because of the complexity of these cases. The system would cope far better with simpler, less interconnected cases, such as shoplifting incidents. However, the marginal benefits of using ADA for simpler tasks might not be as great. This is because simpler, template driven or robotic approaches could be as effective.
- The benefits of ADA are not universally spread. Officers who were either dyslexic or had other difficulties in writing reports were more enthusiastic and estimated higher levels of time saved. Those with experience of writing reports, or higher writing skills/standards, reported less time saved as they tended to spend longer checking and correcting errors in the ADA output.
- The data suggests that AI applications such as ADA will ultimately result in a standardisation of statements, both structurally and linguistically. This means that the standard will move more towards the average, with fewer poor or excellently written statements.
- The system was gradually improved to remove AI hallucinations to a great extent. However, we still believe there will be a risk of hallucinations appearing in statements and there will need to be a change of emphasis in the role of officers towards review and error checking/quality control.
- The biggest risk of the output concerns subtle errors, omissions or hallucinations that are more difficult to detect when reviewing work under time pressure. There is a secondary risk that checking will not occur or be ineffective and this could have significant consequences for the outcomes of legal cases.
- At present there was mixed (inconclusive) evidence about the level of productivity improvements achieved by ADA. It is important that the costs for the whole journey of the statement from interview to court use are considered when finally deciding whether there are cost savings. This will partly be influenced by outcomes such as CPS levels of acceptance of statements, changes in



prosecution rates (and hence police time needed in file build and court activity), any changes to the rates of guilty pleas, defence responses to AI generated testimony and conviction rates.

RVR

- Whilst the research was not asked to specifically comment on the effectiveness of the RVR, it was clear from the data provided that the system as a stand-alone use of technology was effective. There was good data available to demonstrate the impact on productivity on the force as a whole (reduced in-person attendance).
- We did note that the protocol for the use of RVR is not adhered to. Users were taking a pragmatic approach to breaking contact with the witness between interview end and report completion. Our expectation was one continuous contact with the witness up to the point of signature. If future versions of ADA were to perform well (with low risk of errors and omissions) then it could help maintain the protocol if desired.
- Most of the users of RVR only do so for short periods of time, such as officers who are on restricted duties. This means that there is a high turnover of users of the technology. In the short term this has implications for the effectiveness of less experienced users and supervision and training requirements.

Implications of the use of ADA

- ADA was seen as easy to use but the system will only be fully effective if a wide range of officers/investigators are trained in its use and correct procedures.
- The use of ADA changes the interview process. It is an opportunity for the officer to focus on the interaction with the witness, with less worry about recording all details. However, the audio does not automatically include any visual evidence, such as injuries to the witness or events occurring in the background. Therefore, it changes the need for more of an interview narrative about evidential aspects that are visual in nature. Interviewers will need training or guidance about how to best do this.
- These developments present an opportunity to improve the nature of contact with witnesses, through a better engagement process.
- Another consequence is the need to disclose all versions of the transcript and statement produced by officers when using ADA. This impacts on workload for the Disclosure officer and the CPS.



Other Observations

- The work has highlighted a clear difference of opinion about the speed and direction of travel in the use of AI in policing. The technical perspective is that the use of AI will spread quickly and extensively in policing, often replacing officers and staff. In contrast, users see a more limited application of AI, with its purpose to support policing rather than replace policing activities.

Recommendations – Hertfordshire

Version 2 of ADA is now ready for testing and continued development, and we are satisfied that there is a case to continue with this work. We make the following points about the next steps:

- Clarify what ADA is intended to do in the medium term. Is this, as the name suggests, a “digital assistant” or is the intention for it to completely replace the officer in writing drafts of reports?
- Review how the use of ADA is impacting on the functioning of RVR, including if the protocol needs to be refined.
- Consider making a judgement on whether simpler tasks would be better for this type of AI, e.g. shoplifting incidents using body-worn camera audio.
- Monitor the time spent on checking the ADA outputs by officers and consider whether any fail-safe procedures need to be introduced to ensure checks are done.
- Study the different impacts on officers, based around literacy and the report-writing skills or experience.
- Study the workload impact downstream, especially the time spent on quality control.
- Increase the involvement of CPS if possible, to ensure that the process is fully redesigned to suit all parties. This is a two-way process because there may be additional benefits for the police if it helps reduce file return issues and the acceptability of outputs downstream.
- Avoid any duplication of effort with other Forces who have more recently become involved in ADA development work.
- Revisit training needs of officers’ interviewing techniques and statement writing.



Recommendations – Policing

Although the comments below are beyond the original scope of the work, we thought it useful to provide some reflection on the findings for policing in general. These are issues raised more than definitive findings, but hopefully valuable:

- The underlying AI strategy, especially the productivity drivers for AI adoption, need to be clarified.
- There could be better direction in terms of what type of technology should be used for policing applications. Could some of this work be done by simpler robotic processing or even templates? Are the marginal benefits of AI currently worthwhile?
- Forces need to be clear about the ways in which AI should provide assistance. For example, any technological enhancement away from robotic processes towards AI generated advice or decision-making is a big step from risk and ethics perspectives.
- The overall cost impacts of AI need to be better understood, especially around realisable benefits. The long-term implications for the expenditure on data storage and data processing also need to be factored into funding calculations.
- The development of AI needs to be much more of a collaborative venture within policing, otherwise a lot of police time will be spent on duplicated development work.



Concluding remarks

This study asked three research questions. Below we provide a quick summary answer to each of them:

RQ1: What are the users and other stakeholders reporting as the advantages, disadvantages and issues with the use of ADA?

The main advantage of ADA is the speed with which a draft witness statement can be compiled, once an interview has concluded. Most users however felt that the key disadvantage was the output quality of the statements that required a lot of rework. It did provide assistance to those who struggled with report-writing and there would be general support for the automation of the statement-writing process by users.

RQ2: What are the measurable differences between the quality of the output of ADA when compared with outputs written without AI assistance?

The key measurable differences between AI and human text mostly concern readability and the way in which the context and narrative of an incident are reported. AI text is less readable and the present documents are only marginally above the chosen minimum required standard for official documents such as insurance policies. The AI texts do not necessarily tell the correct story of an incident before they are corrected.

These findings are supported in the qualitative interviews and also by direct comparison of AI and human witness statements generated by the same interview content. The language and phrases used by the AI is readily detectable and appears to be clumsily written with unnecessary phrases and complex, unnatural wording.

RQ3: What evidence is there about the impact on productivity and workload when AI is used to generate statements?

There was no consistent evidence in the first version of ADA that it would lead to productivity savings. The call durations using ADA were slightly longer, report generation and checking took significantly longer and there was a likelihood that additional checks would be needed to guard against hallucinations, errors and omissions.

Overall, the study was conducted to see if there was sufficient evidence of proof of concept for ADA. Our conclusion is that version 1 of the AI has not provided this proof, but the imminent version 2 may improve significantly in terms of performance.



Limitations

This is very much a pilot evaluation of an early version of an AI tool for specific use in policing. We conducted a limited number of interviews and had a limited quantity of sample data. A larger sample of real statements and more of a comparative sample of AI and human written text from the same incidents would have been helpful. The pause in the study activity prevented this from happening. We also had only secondary Power BI data provided, so some statistical tests on the data were not possible, although the data provided was robust.

Acknowledgements

We are grateful to Hertfordshire Constabulary for their cooperation and work in helping assemble the data. We also thank the officers and staff for the time they spent in interviews and time redacting statement samples. The evaluation is part of a “Test and Learn” grant obtained by Hertfordshire Constabulary and we thank the funding body for their support.



References

- Adams, I. T. (2024) Large language models and artificial intelligence for police report writing. In *CrimRxiv*. <https://doi.org/10.21428/cb6ab371.779603ee>
- Adams, I.T., Barter, M., McLean, K. Boehme, H.M. and Geary Jr, I.A. (2024) No man's hand: artificial intelligence does not improve police report writing speed. *Journal of Experimental Criminology*. <https://doi.org/10.1007/s11292-024-09644-7>
- Asaro, P. M. (2019) AI ethics in predictive policing: From models of threat to an ethics of care. *IEEE Technology and Society Magazine*, 38(2), 40–53. <https://doi.org/10.1109/MTS.2019.2915154>
- Berryhill, J., Heany, K.K., Clogher, R., and McBride, K. (2019) “Hello, World: Artificial intelligence and its use in the public sector”, *OECD Working Papers on Public Governance*, No. 36, OECD Publishing, Paris, <https://doi.org/10.1787/726fd39d-en>.
- Biber, D. (1988) *Variation Across Speech and Writing*, Cambridge University Press, Cambridge.
- Campion, A., Gasco-Hernandez, M., Jankin Mikhaylov, S., & Esteve, M. (2022) Overcoming the Challenges of Collaboratively Adopting Artificial Intelligence in the Public Sector. *Social Science Computer Review*, 40(2), 462-477. <https://doi.org/10.1177/0894439320979953>
- College of Policing (2024) *Guidance to police forces on building AI tools and systems*, College of Policing, p26.
- Dechesne, F, Dignum, V., Zardiasvili, L., and Bieger, J. (2019) *AI & ethics at the Police, Towards Responsible use of Artificial Intelligence in the Dutch Police*, Report, v 1.2, Leiden University. <https://hdl.handle.net/1887/85954>
- Dement, C., and Inglis, M. (2024) Artificial intelligence-assisted criminal justice reporting: An exploratory study of benefits, concerns, and future directions. *Criminology & Criminal Justice*, 17488958241274296. <https://doi.org/10.1177/17488958241274296>
- de Sousa, W.G, Pereira de Melo, E.R, De Souza Bermejo, P.H., Sousa Farias, R.A., Gomes, A.O., (2019) How and where is artificial intelligence in the public sector going? A literature review and research agenda, *Government Information Quarterly*, Volume 36, (4), <https://doi.org/10.1016/j.giq.2019.07.004>
- Feldstein, S. (2019) *The global expansion of AI surveillance* (Vol. 17). Carnegie Endowment for International Peace.
- Felson, R., Messner, S., Hoskin, A and Deane, G. (2006) Reasons for reporting and not reporting domestic abuse to police. *Criminology*. 40 (3) 617-648. Available through: ARU Library website <library.aru.ac.uk>



Ferguson, A.G., (2024) *Generative Suspicion and the Risks of AI-Assisted Police Reports* (July 17). Available at SSRN: <https://ssrn.com/abstract=4897632>

Flesch, R. (1948) A New Readability Yardstick. *Journal of Applied Psychology*, 32, 221-233.

Friginal, E. and Biber, D. (2016) Multi-dimensional analysis. Eds. In: Baker, P., Egbert, J. (Eds.), *Triangulating Methodological Approaches in Corpus Linguistic Research*. Routledge, Abingdon, 73–89.

Guo, Z., and Kennedy, L. (2023) Policing based on automatic facial recognition. *Artificial Intelligence and Law*, 31, 397–443 <https://doi.org/10.1007/s10506-022-09330-x>

Haesevoets, T., Verschuere, B., Van Severen, R. and Roets, A. (2024) How do citizens perceive the use of Artificial Intelligence in public sector decisions?, *Government Information Quarterly*, Volume 41, Issue 1 <https://doi.org/10.1016/j.giq.2023.101906>.

Harasta, J., Novotná, T. and Savelka, J. (2024) *It Cannot Be Right If It Was Written by AI: On Lawyers' Preferences of Documents Perceived as Authored by an LLM vs a Human* <https://doi.org/10.48550/arXiv.2407.06798>

Hiroshi, M. (2020) Human-centric data protection laws and policies: A lesson from Japan. *Computer Law & Security Review*, 105487. <https://doi.org/10.1016/j.clsr.2020.105487>

Hogg, R. (2024) Delivering Artificial Intelligence in Policing, 3rd Edition. *Black Marble's Essential Guide Series*. Available at <[Black Marble | Delivering AI Intelligence in Policing - 3rd Edition](#)>

Kalai, M., Becha, H. & Helali, K. (2024) Effect of artificial intelligence on economic growth in European countries: a symmetric and asymmetric cointegration based on linear and non-linear ARDL approach. *Economic Structures* Vol. 13, No. 22 <https://doi.org/10.1186/s40008-024-00345-y>

Koppensteiner, M., Matheson, J. and Plugor, R., (2022) *Improving access to support services for victims of domestic violence: demand for services and victim outcomes*. [pdf]. Available at:< https://www.researchgate.net/publication/365149783_The_Impact_of_Improving_Access_to_Support_Services_for_Victims_of_Domestic_Violence_on_Demand_for_Services_and_Victim_Outcomes>

Kuziemski, M., & Misuraca, G. (2020) AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy*, 44(6), 101976. <https://doi.org/10.1016/j.telpol.2020>.

Langer, P. F. (2020) Lessons from China-the formation of a social credit system: profiling, reputation scoring, social engineering. In *The 21st Annual International Conference on Digital Government Research*. Seoul.



Lewis, M. (2021) *Digital Policing: The Ethical Issues Arising from Digital Policing*, South Yorks PCC, pp16.

Murphy, S. and O'Brien, M. (2024) '[Police officers are starting to use AI chatbots to write crime reports. Will they hold up in court?](#)' *The Associated Press*, August 26.

Nicholls, T., Pritchard, M., Reeves, K. and Hilterman, E. (2013) Risk assessment in intimate partner violence: a review of contemporary approaches. *Partner Abuse* [e-journal] 4(1), pp 76-168 Available at: < <https://domesticviolenceresearch.org/risk-assessment/>>

Nini, A. (2019) The Multi-Dimensional Analysis Tagger. In Berber Sardinha, T. & Veirano Pinto M. (eds), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 67-94, London; New York: Bloomsbury Academic

NPCC National Police Chief's Council (2023) *Covenant for using AI in Policing*, v1.1, NPCC London. Available at <[Covenant for Using Artificial Intelligence \(AI\) in Policing](#)>

Oswald, M., Paterson-Young, C., McBride, P., Maher, M., Calder, M., Gill, G., Tiaks, E. and Noble, W., (2024) *Ethical review to support Responsible Artificial Intelligence (AI) in policing: A preliminary study of West Midlands Police's specialist data ethics review committee*, Braid Report, Northumbria University: September 2024. Available at <[Ethical review to support responsible AI in policing: a preliminary study of West Midlands Police's specialist data ethics review committee - BRAID UK](#)>

Quraishi, F. F., Wajid, S. A., and Dhiman, P. (2017) Social and ethical impact of artificial intelligence on public: A case study of university students. *International Journal of Scientific Research in Science, Engineering and Technology*, 3(8), 463–467.

Ransbotham, S., Kiron, D., Gerbert, P., and Reeves, M. (2017) Reshaping business with artificial intelligence: Closing the gap between ambition and action. Edited by *MIT Sloan Management Review*. [https://sloanreview.mit.edu/projects/reshaping-business-with-artificial-intelligence/\(open in a new window\)](https://sloanreview.mit.edu/projects/reshaping-business-with-artificial-intelligence/(open-in-a-new-window))

Rothwell, S., McFadzien, K., Strang, H., Hooper, G. and Pughley, A. (2022) Rapid Video Responses (RVR) vs. Face to Face Responses by Police Officers to Domestic Abuse Victims: A Randomised Controlled Trial. *Cambridge Journal of Evidence-Based Policing* [e-journal] 6, 1–24 <https://doi.org/10.1007/s41887-022-00075-w>.

Sardinha (2024) "AI-generated vs human-authored texts: A multidimensional comparison" *Applied Corpus Linguistics Volume 4, Issue 1*, April.

Schiff, D., Borenstein, J., Biddle, J. and Laas, K. (2021) "AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection," in *IEEE Transactions on Technology and Society*, vol. 2, no. 1, 31-42, doi:[10.36227/techrxiv.14109482](https://doi.org/10.36227/techrxiv.14109482)

Schiff, K. J., Schiff, D. S., Adams, I. T., McCrain, J., and Mourtgos, S. M. (2023) Institutional factors driving citizen perceptions of AI in government: Evidence from a



survey experiment on policing. *Public Administration Review*, 0(0).

<https://doi.org/10.1111/puar.13754>

Spring, M., Faulconbridge, J. and Sarwar, A. (2022), „How information technology automates and augments processes: Insights from Artificial-Intelligence-based systems in professional service operations”, *Journal of Operations Management*, Vol. 68, Iss. 6-7, Pages 592-618.

Stanley, J. (2024) AI Generated Police Reports Raise Concerns Around Transparency, Bias, ACLU Report, December, [AI Generated Police Reports Raise Concerns Around Transparency, Bias | ACLU](#)

Weyerer, J. C., and Langer, P. F. (2019) Garbage in, garbage out: The vicious cycle of AI-based discrimination in the public sector. In Y.-C. Chen, F. Salem, & A. Zuiderwijk (Chairs), *Proceedings of the 20th Annual International Conference on Digital Government Research*. Dubai.

Wirtz, B.W., Weyerer, J.C. and Geyer, C. (2019) Artificial Intelligence and the Public Sector—Applications and Challenges, *International Journal of Public Administration*, 42:7, 596-615, <https://doi.org/10.1080/01900692.2018.1498103>

Wirtz, B., Langer, P.F. and Fenner, C. (2021) Artificial Intelligence in the Public Sector - a Research Agenda, *International Journal of Public Administration*, 44:13,1103-1128, <https://doi.org/10.1080/01900692.2021.1947319>

Appendices

Appendix 1:

1a) Interview questions for officers who have used the software

This interview focuses on officers' use of the AI software (training, process, technicalities and initial evaluation) and how it might have changed the way the officer interacts with the victim. It also explores officers' thoughts about the tool and their attitudes towards AI, including ethical considerations.

What is your role in this project?

Process map – can we go through the functionality and any documentation you can share with us on it?

Functionality:

There have been a number of iterations of ADA:

How many?

How has the functionality changed/will change:

- **First attempt**
- **Current usage**
- **Imminent further development**
- **Long term**

How accurate is this table representing the process as it stands?

Steps in the statement generation process

Step	Action	Comment
1	Incident reported	Police have an option to either visit (with travel time) or switch to a video call. The call can record the context.
2	Real time RVR interview	The RVR is used to immediately obtain an account of the incident by the victim, which is recorded.
3	Audio transcription	The audio track from the video is transcribed using fairly standard software.
4	Transcription entered into AI tool	The transcript is entered into the AI tool, usually with only the minimum of checking for the accuracy of the transcript.
5	Witness statement generated	A draft witness statement is compiled usually taking a few minutes.
6	Statement reviewed by officer	This is reviewed for obvious errors by the officer before it is shared with the complainant.
7	Joint review by officer and victim	The officer and complainant jointly agree the accuracy of the statement.
8	Final reports generated	The statement and other reports are generated and filed.



The complainant perspective: (see later for further questions)

What is the service experience for the complainant:

1. Without ADA
2. With RVR but not ADA?
3. With neither

Did you let the victim know that their statement was being generated by AI?

- Any negative or positive opinions?
- Any perceived differences in outcomes for the victim?

What has been the process of development between you and Anathem?

What has been the division of labour between Herts and Anathem?

- How did they train the software to fit with Herts and points to prove?
- What did you give to Anathem to start with – statements etc?

How much interaction has there been between you and Anathem?

What investment do you think you have put in on this? Time and money.

- Understand the whole suite of documents involved in the report. What does it do, what doesn't it do.

Did you identify anything to feed back to the software developers to change for the next version?

- What was the difficulty/problem that you hope the next version will solve?
- What do you hope the next version will offer that is different from the version you used?
- Do you think that the next version will improve the accuracy of the statement? How?

The AI captures speech but not any observations of the victim or the room that they are in. How did you note this before using AI and what do you do now?

- How do you integrate any observation notes into the AI generated statement?
- Do they give prompts so that this is picked up in the statement?

How do you verify that the AI-generated statements align with the victim's emotions, language, and intent?

Once the victim has finished giving their statement, can you explain what happens next.



- Do you think the AI generated statement captures the emotions of the victim effectively?
- Is this as effective as when you were writing the statement/any difference?
- How is the victim involved in checking the statement?

Technicalities

- Checking the statement/ correcting the statement – process and functionality. How straightforward is this? If you are unsure if something is represented correctly, do you go back and listen to the audio/ask the victim to clarify?
- How do you clarify and change the statement if it doesn't capture what you thought the victim said?

What safeguards are in place to prevent the AI from generating biased or potentially harmful statements?

- Have there been hallucinations?
- Are there mis-reads/factual errors?
- Are there types of omissions?

To what extent has the system changed the dialogue between officer and complainant?

- Verbalising either injury or other visual content?
- Instructing the victim?
- Checking statements

How much of a learning curve have you been on?

What training did you have before using the AI tool?

- What technical skills did you need as part of the development?
- Are there any tools in the software that you don't use/don't make full use of? Why?

What training do others need in order to use this?

- What changes in behaviour are needed when adapting to a new system?
- How much training will be needed? What type?
- Is there need to practice in a classroom situation first?

Since using AI to generate the statement, what interaction do you have with the victim during the interview/ how does this compare with how you did it before?

- Do you engage differently with the victim because you are not having to make notes/type up the statement?
- If so, how. How is this different?
- Do you think this interaction changes the nature of the statement in any way?



- How do you use the functionality of the prompts and the points to prove - do you have a laptop in front of you? Engagement with the victim.
- Would you say that the AI plays a part in the investigative process? Or just capturing what is said?

Compared with the previous process, to what extent do you think that the AI generated statement is accurate?

Before this project, had you had any experience of working with AI in your job/outside of work?

- Follow-up – positive or negative experience. Attitude towards AI as a tool and its benefits.

Do you have any ethical concerns about this method of generating statements?

- Explore any ethical concerns.

Finally, I understand that there are wider concerns about AI generated statements being used in Court. Do you understand what these concerns are?

- Explore linked to any ethical concerns that the officer might have.

1b) Interview questions for senior officers & staff AI development

This work focuses on the application of AI in police forces specifically for the development of productivity tools. The questions we are asking are in this context. However, there may be some links with the use of AI for higher-level decision-making and broader applications such as face recognition.

How advanced do you think the use of AI is at Herts?

What existing AI applications are widely used within the force?

What technical issues (if any) have you had to address with these applications?

- Is the ethical use of AI a widely discussed issue? Governance?

What is the planned direction of travel for the development and adoption of AI based applications in Herts?

- Are there specific areas you will develop/avoid?
- Is there a vision of what Herts use of AI will look like in, say, 10 years?



What are the drivers for the strategy?

- NPCC guidelines?
- Internal initiatives
- Other external pressures/constraints

How do you expect AI to impact on productivity?

Will AI create the need for large scale process change?

What implementation challenges do you see in developing greater use of AI in the force?

Are there specific workforce issues that you will need to address?

- The recruitment of technical staff?
- The development of existing analysts/others
- The training of officers/staff in the use of new applications?
- Training in AI awareness etc.