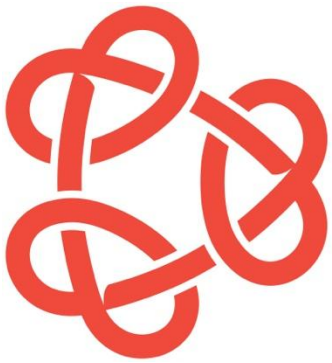


University of Staffordshire

PhD - Computer Science

Automated Digital Forensics – Towards acceptance of Machine Learning and Artificial Intelligence-based automation within forensic analysis of image files

Candidate: Christopher HAWKINS

Supervisors: Dr Russell Campion & Mr Chris Howard

A thesis completed in partial fulfilment of the requirements for the award of Doctor of Philosophy in Computer Science at the University of Staffordshire

Version: 30 Date: SEPTEMBER 2026

Abstract

Digital devices are the modern evolving zeitgeist. They increasingly dominate every aspect of daily life, and without them it is increasingly difficult to participate in society. As a result of this the Digital Forensic (DF) field is one of the most challenging and ever-present domains within law enforcement. The evidential potential of these devices increases with every new advancement. However, the effect of these devices is that they take longer to examine correctly, they are more complex and there are more of them to examine.

There have been ongoing developments to introduce automation technologies in Artificial Intelligence (AI) and Machine Learning (ML) to attempt to solve this urgent need for increased efficiency. However, these advancements have come with increased concern over the governance of these tools and how this can be effectively used whilst maintaining legal standards and principles. This has been achieved until this point with the requirement for human verification of any use of automation technologies.

This thesis proposes a novel term of '*True Automation'* to refer to the concept of relying upon an automated software or device and the analysis that it conducts without needing a human-in-the-loop component, in an attempt to maximise the gains in terms of efficiency which can be obtained through the use of automation technologies. In addition, the thesis proposes a novel assessment metric dubbed the *Practical Reliability (PR) Metric*, for the synthesising of technical accuracy and reliability into a single usable and understandable metric that can be for the rapid and accurate assessment of automation technologies.

Finally, a novel assessment framework for the assessment of automated tools within the image analysis domain is proposed, with accompanying documentation and guidance as to how it can be used to assess tools in a real-world environment.

The requirements for this framework are obtained from a combination of secondary and primary research methods, making use of academic research in the field, as well as the input from serving practitioners in the DF unit of a UK Home Office Police Force.

Rigorous testing of both the framework and the metric is conducted using both expert review and technical testing.

Framework testing involves expert validation, review of the thresholds set and technical review.
This testing shows that, following appropriate revision, all 8 of the proposed gate thresholds were accepted. Additionally, practitioner acceptance testing indicates that 92% of survey respondents and 73.9% of interview respondents would be willing to use automated tools without human review if they passed the framework.

Metric testing shows that current tool developers have not published the data necessary to conduct meaningful F1 and Matthews Correlation Coefficient (MCC) calculations. Therefore, the metric is tested against 17 image manipulation detection methods. This testing shows scores ranging 0.3240 to 0.4960. All results fall within the expected 0 – 1 range but were below the required passing threshold. However, the produced testing results showed that the required simplified score was achieved, accurately representing the required F1 and MCC values, whilst providing a value which could easily be interpreted by non-technical practitioners.

Following these results, the thesis concludes that the framework is effective and suitable at the task and would be viable for assessment of automation tools, and the metric is effective at simplifying complex model assessment metrics whilst maintaining technical relevance. As a result, it is shown that the adoption of '*True Automation*' is possible with appropriate safeguarding. However, cultural, legal and ethical considerations remain which can only be addressed by increasing collaboration between the technical and forensic domains.

Acknowledgements

My sincere and eternal thanks go to my fiancé Helen, my parents Michael and Caroline, and my brother Daniel, who have had to deal with my worrying and stressing about this thesis forever.

I would like to express the sincerest gratitude to my supervisor team, Mr Chris Howard and Dr Russell Campion, who have put up with my questions and procrastination for far too long.

Sincere thanks are also given to everyone at *the Home Office Police Force* – without your contributions, thoughts and opinions, this would have remained simply an idea.

I would finally like to thank my assessor team, and the University of Staffordshire, for the support and guidance throughout this process.

Dedication

I would like to dedicate my PhD Thesis to my amazing soon-to-be wife Helen, without whom this thesis would never have happened.

Table of Contents

ABSTRACT II
ACKNOWLEDGEMENTS IV
DEDICATION V
TABLE OF CONTENTS VI
LIST OF FIGURES IX
LIST OF TABLES XI
LIST OF EQUATIONS XI
LIST OF ABBREVIATIONS & DEFINITIONS XII
PUBLISHED WORKS AND ACADEMIC CONTRIBUTIONS DURING PROJECT XIII

AUTHOR'S DECLARATION XIV

1.0 INTRODUCTION 1

1.1RESEARCH BACKGROUND 1
1.1.1 PROBLEM STATEMENT 6
1.2TRUE AUTOMATION DEFINITION 7
1.3PROBLEM DOMAIN JUSTIFICATION 9
1.3.1FOCUS GROUP TO DETERMINE PROBLEM DOMAIN 9
1.3.1.1PARTICIPANT PROFILES 10
1.3.1.2QUESTIONS AND FORMAT 11
1.3.1.3RESULTS, ANALYSIS AND CONCLUSION 13
1.4ETHICS STATEMENT 14
1.5RESEARCH AIM 14
1.6HYPOTHESIS 14
1.7RESEARCH QUESTIONS 15
1.8RESEARCH OBJECTIVES 15
1.9DELIVERABLES 16
1.10FRAMEWORK DEFINITION AND PROJECT DESIGN 16
1.11CONTRIBUTIONS TO KNOWLEDGE 18
1.12PROJECT MAPPING 23
1.13STRUCTURE OF THESIS 24
1.14 CHAPTER SUMMARY 27

2.0 RESEARCH METHODS 29

2.1 RESEARCH METHODOLOGY 29
2.2 OBJECTIVE / STRATEGY MAPPING 32
2.3 RESEARCH APPROACH 33
2.3.1 PROJECT MANAGEMENT 35
2.4 RESEARCH METHOD & RESEARCH DESIGN 38
2.4.1 PHASE 1 – RESEARCH 38
2.4.2 PHASE 2 – DEVELOPMENT 40

2.4.3 PHASE 3 – TESTING 56
2.4.4 – TESTING METRIC SUMMARY 61
2.5 CHAPTER SUMMARY 61

3.0 SYSTEMATIC LITERATURE REVIEW 62

3.1 THE NEED FOR INCREASED EFFICIENCY IN DIGITAL FORENSIC ANALYSIS 62
3.2 MACHINE LEARNING ALGORITHMS AND ARTIFICIAL INTELLIGENCE 65
3.3 THE USE OF MACHINE LEARNING AND AUTOMATION IN DIGITAL FORENSICS 67
3.4 EXISTING WORKS AND TOOLS 69
3.5 ISSUES FACED BY AUTOMATION IMPLEMENTATION INTO DIGITAL FORENSICS 70
3.6 EXISTING DIGITAL FORENSIC AND INVESTIGATIVE FRAMEWORKS 73
3.7 LEGAL AND ETHICAL CONSIDERATIONS OF *TRUE AUTOMATION* WITHIN DIGITAL FORENSICS 77
3.7.1 RESPONSIBILITY, TRUSTWORTHINESS, ACCEPTANCE AND SUSTAINED ACCURACY 79
3.7.1.1 RESPONSIBILITY 79
3.7.1.2 TRUSTWORTHINESS 80
3.7.1.3 ACCEPTANCE 81
3.7.1.3.1 OVERALL ACCEPTANCE 82
3.7.1.4 SUSTAINED ACCURACY 82
3.8 GAP ANALYSIS 83
3.9 CHAPTER SUMMARY 87

4.0 THE REAL-WORLD PRACTICAL STATE OF DIGITAL FORENSIC-BASED AUTOMATION WITHIN CRIMINAL JUSTICE IN THE UK 89

4.1 ORGANISATIONAL STATE OF AUTOMATION 89
4.2 PRACTITIONER STATE OF AUTOMATION 93
4.3 LAW ENFORCEMENT CONCERNS OVER AUTOMATION USE 106
4.4 CHAPTER SUMMARY 109

5.0 A FRAMEWORK TO IMPLEMENT DIGITAL FORENSIC-BASED AI AND ML AUTOMATION WITHIN THE UK'S CRIMINAL JUSTICE SECTOR 110

5.1 FRAMEWORK PREMISE AND OPERATION 110
5.2 INTERVIEW, SURVEY AND FOCUS GROUP ANALYSIS 112
5.3 A NOVEL DF DOMAIN ASSESSMENT METRIC – *PRACTICAL RELIABILITY* 116
5.4 A NOVEL ASSESSMENT FRAMEWORK TO IMPLEMENT *TRUE AUTOMATION* INTO THE UK'S DF DOMAIN 124
5.5 GATE EXPLANATIONS & FRAMEWORK FUNCTION 125
5.6 FRAMEWORK ASSESSMENT MATRIX 148
5.7 CHAPTER SUMMARY 148

6.0 TESTING & VALIDATION OF FRAMEWORK 150

6.1 TESTING METHODOLOGIES 150
6.2 TESTING METHOD SELECTION 151

6.3 Testing and Validation of Framework 153
6.3.1 Expert Validation 153
6.3.2 Gate Passing Metrics 167
6.3.3 Technical Testing 174
6.3.4 Comparative Testing Use and Suitability Summary 189
6.4 Testing Conclusions and Analysis 189
6.4.1 – Future Work 191
6.5 Chapter Summary 191

7.0 – TESTING & VALIDATION OF *PRACTICAL RELIABILITY* METRIC 193

7.1 Expert Validation 193
7.1.1 Methodology Justification and Participant Selection 193
7.1.2 Testing Scope 194
7.1.3 Testing Results 194
7.1.4 Testing Conclusions 195
7.2 Technical Testing 196
7.2.1 Testing Method 196
7.2.2 Testing Scope 196
7.2.3 Testing Results 197
7.2.4 Testing Summary and Analysis 199
7.3 – Testing Conclusions and Analysis 201
7.3.1 Future Work 202
7.4 Chapter Summary 202

8.0 – PROJECT EVALUATION & CRITICAL REVIEW 204

8.1 Project Scope 204
8.1.1 Stakeholders and Police Force 204
8.1.2 Demographic Diversity 205
8.1.3 Problem Domain 206
8.2 Research Methodology 207
8.2.1 Lack of Triangulation 207
8.2.2 Researcher Bias 208
8.2.3 Sample Sizing 209
8.2.4 Alternative Research Methods 210
8.3 Framework Design 211
8.3.1 Dependence 211
8.3.2 ISO Requirements 212
8.3.3 Missing Element 213
8.3.4 Interview and Survey Misunderstanding 214
8.4 Testing & Validation 215
8.4.1 PR Metric Testing 215
8.4.2 Focus Group Testing 216
8.4.3 Gate Passing Metric Testing 218
8.5 PR Metric 219
8.5.1 PR Metric Scoring 219

8.5.2 Edge Cases 220
8.6 Operational Challenges 220
8.6.1 Internal Resourcing 220
8.6.2 Defence and Disclosure 221
8.7 Evaluation Summary 222
8.8 Additional Validation 223

9.0 – RESEARCH CONCLUSIONS AND RECOMMENDATIONS 226

9.1 Context and Motivation 226
9.2 Research Aim 227
9.3 Hypothesis 228
9.4 Research Questions 228
9.5 Research Objectives 230
9.6 Deliverables 231
9.7 Contributions to Knowledge 233
9.8 Analysis 235
9.9 Recommendations 236
9.10 Future Work 239
9.11 Final Word 242

10.0 REFERENCES 243

11.0 APPENDICES 254

Appendix 1 – Focus Group Notes 254
Appendix 2 – Ethics Forms 257
Appendix 3 – Survey Information Sheet and Question Sheet 268
Appendix 4 – Interview Information Sheet 278
Appendix 5 – Interview Question Script & Answer Recording Document 282
Appendix 6 – Testing Focus Group Participant Information Sheet 286
Appendix 7 – Testing Focus Group Domains Script 291
Appendix 8 – Primary Research Conducted in 2021 294
Appendix 9 – Framework Testing Focus Group Results 295
Appendix 10 – Metric Passing Value Testing Results 301
Appendix 11 – PR Metric Expert Testing Results 304
Appendix 12 – Additional Validation Results 307

List of Figures

Figure 1 - Messages between Carter and Roy (USA Gov, 2014) 3
Figure 2 - Automation and ML/DL/AI (Jarrett & Choo, 2021) 7
Figure 3 - Artefact Design and Project Process 17
Figure 4 - Thesis Structure 24
Figure 5 - Research Onion (Saunders, Lewis and Thornhill, 2019) 30
Figure 6 - Initial Gantt Chart 36

Figure 7 - Project Management 37
Figure 8 - Literature Review Process....39
Figure 9 - FOI Request Wording 41
Figure 10 - SPSS Variables Setup 42
Figure 11 - Variable Setting within SPSS....43
Figure 12 - Variable Setting within SPSS for Refusal Code 43
Figure 13 - Survey Information Sheet....46
Figure 14 - SPSS Survey Variables 50
Figure 15 - Interview Codebook 56
Figure 16 - Number of Officers Concerned about the number of devices being submitted for analysis....64
Figure 17 - Number of Officers concerned about the length of time being taken for analysis 64
Figure 18 - AI and ML Venn Diagram (Malandruccolo, 2024) 65
Figure 19 - Machine Types and Techniques (Sarker, 2021) 66
Figure 20 - Current AI-Based Digital Forensic System (Hall, Sakzad and Choo, 2022) 71
Figure 21 - Collated FOI Responses 90
Figure 22 - Refusal Reason Breakdown 91
Figure 23 - Cost Exclusion Response....91
Figure 24 - Standard Deviation of Survey Results 94
Figure 25 - Visualisation of Survey Awareness Question 95
Figure 26 - Visualisation of Survey Tool Awareness Question....95
Figure 27 - Visualisation of Survey Tool Use Question 96
Figure 28 - Visualisation of Survey Frameworks Awareness Question 96
Figure 29 - Visualisation of Survey Automation Willingness Question....97
Figure 30 - Visualisation of Survey Willingness Question....97
Figure 31 - Visualisation of Survey Reliance Question....98
Figure 32 - Pearson Chi-Square Test....99
Figure 33 - Frequency Tables for Survey Question on Tool and Automation Awareness....100
Figure 34 - Frequency Table for Survey Tool Use Question....100
Figure 35 - Frequency Table for Survey Frameworks Awareness Question....101
Figure 36 - Frequency Table for Survey Automation Acceptance Questions....103
Figure 37 - Results of Interview Questioning on the use of Automation 104
Figure 38 - Visualisation of Interview Question on whether a framework would solve concerns....106
Figure 39- Framework Creation Information Flow 111
Figure 40 - A Novel DF Tool Assessment Framework 124
Figure 41 - Framework - Focus Group Versio 157
Figure 42 - Likert Results of Framework Testing 158
Figure 43 - Definition Inclusion Example....161
Figure 44 - Colour Coding Legend 161
Figure 45 - Passing Level Acceptance Rate....171
Figure 46 - Autopsy 3rd Party Module Search Result 181
Figure 47 - Autopsy Function Pipeline (Autopsy, 2018)....182
Figure 48 - Original DF Gate Definition 188
Figure 49 - Interview Respondent Representation 190

List of Tables

Table 1 - Participant Mapping for SFP Focus Group....11
Table 2 - Contributions to Knowledge....19
Table 3 - Project Mapping Matrix....23
Table 4 - Objective & Methodology Mapping....32
Table 5 - Literature Review Criterion....40
Table 6 - Seven Steps Approach (Adeoye-Olatunde and Olenik, 2021)....51
Table 7 - Machine Learning Types (Sarker, 2021)....66
Table 8 - Existing Investigative Frameworks....73
Table 9 - Automation Accuracy Confusion Matrix....117
Table 10 - PR Assessment Matrix....129
Table 11 - Time Saving Assessment Matrix....131
Table 12 - Transparency Assessment Matrix....133
Table 13 - Verification Assessment Matrix....136
Table 14 - Report Generation Assessment Matrix....139
Table 15 - Updating Assessment Matrix....141
Table 16 - Tool Tie-In Assessment Matrix....143
Table 17 - Malleability Assessment Matrix....146
Table 18 - Framework Metric Matrix....148
Table 19 - Framework Validation Participant Information....155
Table 20 - Technical Testing Summary....186
Table 21 - Framework Testing Comparison....189
Table 22 - PR Metric Scores (Kong et al., 2024)....199

List of Equations

Equation 1 – Accuracy Calculation....117
Equation 2 - Precision Calculation....118
Equation 3 - Recall Calculation....119
Equation 4 - F1 Score Calculation....119
Equation 5 - MCC Calculation....120
Equation 6 - MCC[1] Calculation....120
Equation 7 - PR Metric Calculation....123

List of Abbreviations & Definitions

ACPO – Association of Chief Police Officers

AIMS - Artificial Intelligence Management System

CSAM – Child Sexual Abuse Material

CJ – Criminal Justice

DF – Digital Forensics

DFU – Digital Forensics Unit

DL – Deep Learning

ESR – Early-Stage Review

FOI – Freedom of Information

Framework – *"A Basic Structure underlying a system, concept or text."* (Cambridge Dictionary, 2022)

GB - Gigabyte

GDPR – General Data Protection Regulations

HMCTS – His Majesty's Courts & Tribunal Service

MCC – Matthews Correlation Coefficient

MLA – Machine Learning Algorithm

MSR – Mid Stage Review

NDA – Non-Disclosure Agreement

NPCC – National Police Chiefs Council

PR – Practical Reliability

RAM – Random Access Memory

SFP – Staffordshire Forensic Partnership

SLR – Systematic Literature Review

SOP – Standard Operating Procedures

TB – Terabyte

TED – Tell, Explain, Describe

UK – United Kingdom

UKAS – United Kingdom Accreditation Service

XAI – Explainable Artificial Intelligence

Published Works and Academic Contributions During Project

Reviewer and Panel Member for *APIIT International Research Conference 2024 – Colombo, Sri Lanka*

Hawkins, C.P. (2024) As an ethical hacker, I can't believe the risks people routinely take when they access the internet in public, The Conversation. Available at: http://theconversation.com/as-an-ethical-hacker-i-cant-believe-the-risks-people-routinely-take-when-they-access-the-internet-in-public-240599 (Accessed: 9 December 2024).

Hawkins, C. and Cahill, S. (2024) 'A BESPOKE LABORATORY ENVIRONMENT FOR THE DELIVERY OF PRACTICAL TECHNICAL-BASED COMPUTING TRAINING AND EDUCATION TO REMOTE AND BLENDED LEARNERS: A PRACTICAL IMPLEMENTATION', *ICERI2024 Proceedings*, pp. 9804–9813. Available at: https://doi.org/10.21125/iceri.2024.2477.

Panel Chair: ICERI 2024 Panel – Flipped and Blended Learning – Seville, Spain

Professional attendance at the 2025 conference on *Use of AI and Drones in Policing* – University of Staffordshire, Stoke-On-Trent, Staffordshire

Professional attendance at the 2025 presentation on *AI and Digital Inclusion* – Good Things Foundation

Professional attendance at the 2025 conference on *Policing, Automation and AI Governance* - BluePrism and AWS

Professional attendance at the 2025 conference on *Think Digital Government* – Think Digital Government Service

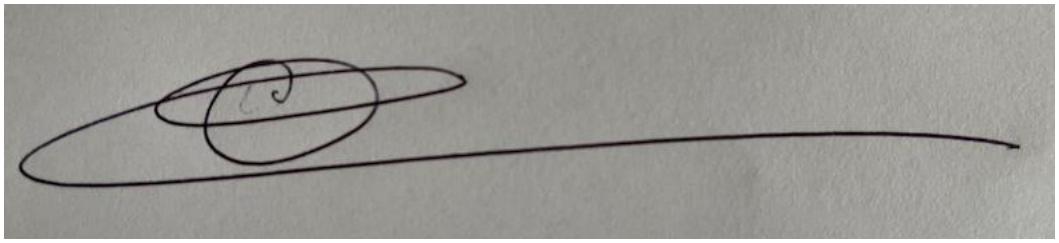
Professional attendance at the 2025 conference on *DeepFake Detection Research Briefing* – Accelerated Capability Environment (ACE), Home Office

Author's Declaration

I declare that this thesis is my own, and has not been previously submitted for any other degree at the University of Staffordshire nor any other institution.

No other person has contributed to this thesis, other than within primary research as stated within the thesis.

Signed:



C HAWKINS

16/03/2026

1.0 Introduction

The following chapter introduces the project, and provides the research background, motivation, as well as the initial problem domain specification and justification.

The chapter also introduces the first contribution to knowledge in providing the definition of '*True Automation*'.

As well as serving as an overall introduction, the chapter provides the basis and context for all subsequent chapters of the work and gives the reader the information that is needed to understand and inform the remainder of the project report.

The chapter provides the problem definition first, before moving to the '*True Automation*' definition to assist in understanding how the problem definition is being addressed within the project. The specific domain is then explored, with a focus group being explained and conducted, this ends with the selection of image analysis of the problem domain based on this focus group and the knowledge of the researcher.

Following this establishment, the chapter moves to explaining the required research components such as the research aim, hypothesis, questions and objectives. Specific requirements such as contributions to knowledge are also given. The chapter then ends with an explanation as to how the remainder of the thesis has been structured.

1.1 Research Background

Society is increasingly dependent upon electronic communication devices in a variety of forms, from traditional computers and laptops, to tablets and phones, with the propagation of these devices being extreme to the point that there are more people on the planet that own a mobile phone, than own a toothbrush (Bentson, 2024). These devices now form an integral part of daily life for people all over the world, with increasing capability and complexity coming as a natural extension of society's developing reliance upon them.

This prevalence does not only extend to the ownership of devices, but people are also making use of these devices at record rates, with an upward trend on display. In fact,

as of April 2025, the global average screen time per day was 6 hours and 40 minutes, an increase of 30 minutes since 2013 (Duarte, 2025), additionally we are using devices from a younger age, with research showing that 98% of United Kingdom (UK) under-2-year-olds view screens on a typical day (Fish *et al.*, 2026).

A natural change because of the pervasiveness of these devices, is the relationship that users have with them, and how they are used, as well as to what extent they permeate lives. Particularly, the use of mobile devices has changed with users moving from seeing them as purely voice or text communication mediums to being a key part of interpersonal relationships and how people interact with the world around them. This is evidenced with the method by which people form relationships, with meeting partners online outpacing all other means of introduction in 2019 (Rosenfeld, Thomas and Hausen, 2019).

This shift has wide ranging ramifications across all of human society, not the least of which is law enforcement and forensic investigations. As the use and presence of devices has increased, the capabilities of these devices have also increased – the smartphone has become an immersive digital environment which contains a wealth of digital evidence about suspects and victims which would have previously been difficult or impossible to obtain. An example of this is the 2014 case of Michelle Carter and Conrad Roy. During this case, it was found that the vast bulk of the relationship between the two subjects was almost exclusively conducted via mobile phones and text messages, which afforded the judicial system an in-depth look into the relationship as it was seen by those subjects. This enabled the prosecution and conviction of Carter for encouraging the suicide of Roy, something that would have been much harder to achieve without the level of insight afforded by the forensic analysis of the messages sent between them, an example of this is shown below in Figure 1:

7/12/14 5:01 PM	Incoming	From: +15088384879 Michelle Carter	Well wait the generator is gonna be on because you'll be passed out so they'll know you used carbon posioning	Conrad	SMS Messages
7/12/14 5:01 PM	Incoming	From: +15088384879 Michelle Carter	Dead*	Conrad	SMS Messages
7/12/14 5:01 PM	Incoming	From: +15088384879 Michelle Carter	Its not loud is it?	Conrad	SMS Messages
7/12/14 5:05 PM	Outgoing	To: +15088384879 Michelle Carter	Not really lmao	Conrad	SMS Messages
7/12/14 5:05 PM	Incoming	From: +15088384879 Michelle Carter	Okay good	Conrad	SMS Messages
7/12/14 5:08 PM	Incoming	From: +15088384879 Michelle Carter	Are you gonna do it now?	Conrad	SMS Messages
7/12/14 5:09 PM	Outgoing	To: +15088384879 Michelle Carter	I'm home.	Conrad	SMS Messages
7/12/14 5:09 PM	Incoming	From: +15088384879 Michelle Carter	Okay	Conrad	SMS Messages
7/12/14 5:13 PM	Outgoing	To: +15088384879 Michelle Carter	Ahhh	Conrad	SMS Messages
7/12/14 5:14 PM	Incoming	From: +15088384879 Michelle Carter	What?	Conrad	SMS Messages
7/12/14 5:14 PM	Outgoing	To: +15088384879 Michelle Carter	Idkk I'm stressin	Conrad	SMS Messages
7/12/14 5:15 PM	Incoming	From: +15088384879 Michelle Carter	You're fine, it's gonna be okay. You just gotta do it babe, you can't think about it	Conrad	SMS Messages
7/12/14 5:16 PM	Outgoing	To: +15088384879 Michelle Carter	okay. okay I got this	Conrad	SMS Messages
7/12/14 5:17 PM	Incoming	From: +15088384879 Michelle Carter	Yes you do, I believe in you. Did you delete the messages?	Conrad	SMS Messages
7/12/14 5:17 PM	Outgoing	To: +15088384879 Michelle Carter	Yes but ur gonna keep messaging me?	Conrad	SMS Messages

Figure 1 - Messages between Carter and Roy (USA Gov, 2014)

This has obvious judicial benefits, however, it also increases the strain and requirements placed upon DF analysis and the teams that undertake this, due to the increase in the amount of data present on said devices, as well as the volume of devices that are being submitted.

A further implication of this societal shift which impacts upon the DF domain is that criminality is shifting towards the use of digital media, and is actively employing advanced and digitally enabled counter-forensic techniques, further advancing the problems that are facing the DF domain in respect of time taken for analysis, and the level of information that is available to investigators (Andreja, 2025).

A recent and impactful example was the world-wide EncroChat and Sky ECC systems that was used by criminal entities to send messages with perceived impunity. The use of this system was massively impactful for law enforcement across the world, as traditional DF processes could not penetrate the encryption in use on the devices. This required the use of advanced covert cyber techniques to address, leading to significant legal and ethical challenges discussed in *Chapter 3.7* to this thesis (Andreja, 2025).

The above notwithstanding, an unexpected benefit that can be obtained in these cases is that the individuals involved in the use of these systems felt confident and assured in the defences that they had employed, meaning that when this was broken, millions of messages were obtained by law enforcement which were openly discussing a range of serious criminality, leading to a significant evidence pool being obtained.

This leads into concerns that have been consistently raised within the field of DF, including publications as far back as Rogers and Seigfried, (2004), which have continued into the modern arena as shown in Luciano *et al.* (2018), Gentry and Soltys (2019), Sunde and Dror (2019) and Kim and Lee (2020).

These works state that whilst there are several issues that face the DF environment, there are some key, and persistent, challenges identified which require addressing. The identified key issue is that investigations are increasingly including the examination of digital devices, which leads to investigators having to deal with potentially extremely large numbers of devices (Patil *et al.*, 2022). Further to this, due to the amount of data that is held on these devices, examination of even small devices can take a significant amount of time. These problems combine resulting in the fact that DF Laboratories are struggling to cope with the level of demand being placed upon them (Rappert, Wheat and Wilson-Kovacs, 2021), which necessitates a change in strategy and the adoption of newer solutions and technologies (Qadir and Varol, 2020).

Breaking this down further, one of the most significant and growing areas of DF analysis is the requirement for image analysis (Holt, Bossler and Seigfried-Spellar, 2022). There are a multitude of reasons for this, which will be explored in this thesis, including not only the increasing abundance of devices capable of storing and sharing images, the increasing existence of tools that make use of AI and automation technologies for the purposes of image analysis (Grossman and Grimm, 2025), but also the increasing prevalence of “DeepFakes” and AI-generated imagery increasing the accessibility and prevalence of images which constitute an offence under UK Law (Verdoliva, 2020), as well as the technology to make such images being far more accessible to everyday users than has been seen in previous years (Hawkins, 2025).

Additional to the above technical points, there exists further issues such as the psychological trauma that is inflicted upon digital forensic examiners when they must manually review and classify images of offence, especially indecent images of children (Strickland, 2022).

It is logical to expand upon this and state that these concerns would not act in isolation. The psychological trauma inflicted upon examiners would interact with the

increasing workload outlined above, adding additional strain to a system already struggling to keep pace with changing technology and criminality, as well as investigative practice.

One of the prominent solutions proposed in the most recent academic works aiming to solve these problems is the introduction of ML into DF processes (Qadir and Varol, 2020). However, this has not been without issues, the main being that there is a lack of applicable datasets which limits the effectiveness, and research, in relation to the combination of these fields (Bhatt and Rughani, 2017).

This thesis builds on the research of Qadir and Noor (2021), in which they conducted a review of ML and its use within the DF field, following their review, they proposed a basic framework for the use of ML within the field. However, this was very limited and did not seek to involve the law-enforcement professionals in its creation, nor did it consider whether their framework for an automated tool would ensure that the findings of that tool would be admissible in court. Additionally, this thesis expands upon the existing framework considerations around Explainable Artificial Intelligence (XAI) conducted by (Khalid, Iqbal and Fung, 2024), in which the authors stated that in order for DF to continue making use of automation technology in any meaningful manner, the need for XAI must be addressed to ensure key elements such as Transparency, Accountability and Reproducibility when using AI tools.

As of the time of writing this thesis, there does not appear to be any current research, nor work, towards developing a workable and viable framework against which ML, or a tool using ML, could be assessed in terms of its applicability to criminal justice and the extent to which the court system could rely on such a tool in lieu of human analysis. As stated above, the XAI work conducted by Khalid, Iqbal and Fung (2024) points out the problem and takes steps towards addressing this but still requires a human-in-the-loop component which has a significant effect on the time-saving potential that is desperately required by the DF domain. The researcher, as part of his role, has attended multiple conferences, events and discussions around increasing the use of AI and automation technology within law enforcement, and within these discussions, a key omission has been the testing, verification and assessment of tools that are being considered for use within the domain. This omission was emphasised

at the 2025 ACE conference on Deepfake analysis and the criminal justice system, in which it was pointed out that there is no current formal, nor informal, system for the validation of AI tools for use within either the court system, nor the wider Criminal Justice System (Hawkins, 2025).

The result of this lack of uniformity and clarity is that there is a lack of engagement and trust of automated processes within the UK Criminal Justice sector and Court system. This issue was showcased in the 2026 UK Government White Paper on Policing, which discussed the use of AI in depth, and introduced a new system of nationalisation for the use of AI within Policing (Home Office, 2026).

This thesis looks to address this concern by creating a novel framework against which those tools could be assessed, and which would provide the Criminal Justice system evidence to justify any such future reliance on truly automated analysis of image files within the realm of DF analysis.

1.1.1 Problem Statement

The above chapter states the background to this research, and the overall problems that are facing the DF domain following the increase in the prevalence and complexity of devices. The chapter also shows that in response to the concerns raised, the DF domain has moved to increase the use of innovative technology such as AI and ML.

The issue that is raised is that the ongoing requirement for human oversight either reduces or eliminates the efficiency benefits that are obtained.

Accordingly, this thesis seeks to answer the question as to whether it would be possible to remove the requirement for this oversight and rely on automated tools to perform their function without needing to be checked by a human user. The project looks to address whether this would be acceptable from a user perspective, what assessment and assurance would be required, and whether current technology is able to meet these requirements.

The thesis follows DF domain trends in using AI, XAI and ML as tools to obtain a result. The project does not aim to utilise these technologies as assessment tools, rather how they themselves and their use within DF specific tools can be assessed

and relied upon to produce reliable and trustworthy results. As this is a novel use, the focus on the thesis is on how this assessment could be completed, with the main deliverable being the assessment framework. Due to the limitations of the thesis (*Chapter 8*) it was considered that the produced framework (Figure 40) would need additional stakeholder engagement to be fully developed. As such, the practical adoption of this framework is considered beyond the scope of this thesis but is identified as a key future work.

1.2 True Automation Definition

When considering the applicability of such concepts to DF, or any process, the differences between *AI, ML* and *Automation* (Figure 2), and how these concepts interact, must be explored. It is important to define these terms and the concepts that will be used at the outset of this thesis.

In a broad sense, and the sense that will be used for this thesis – whilst the concepts of ML and AI focus upon the formation of conclusions and results based on input data, with *automation or Robotic Process Automation (RPA)* referring to the robotic process of repeating specified actions in a defined order, without the necessity of including a human user in that process (Jarrett and Choo, 2021).

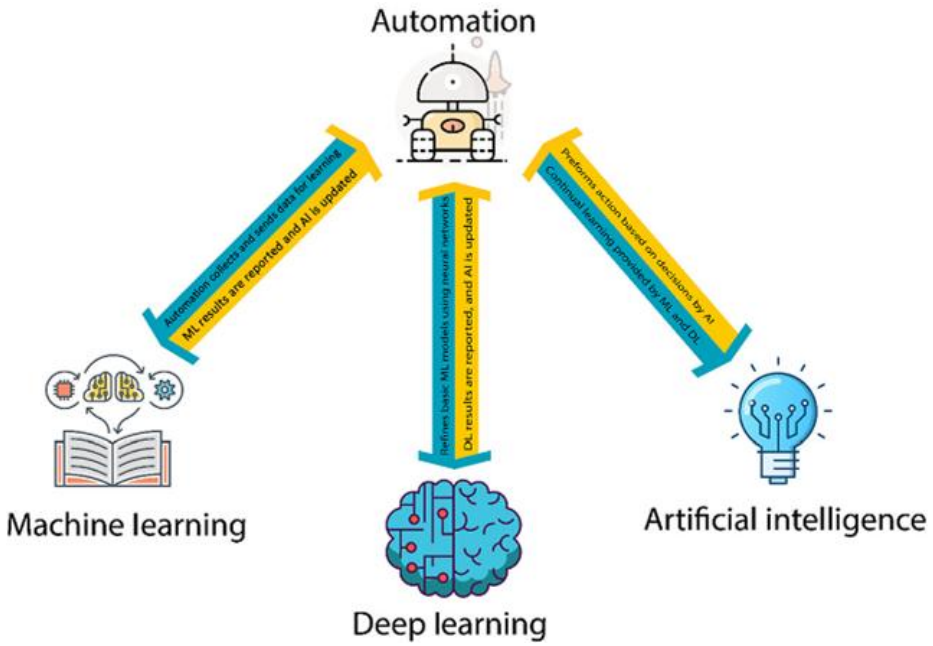


Figure 2 - Automation and ML/DL/AI (Jarrett & Choo, 2021)

As the DF domain has developed several tools have been created to assist in the analysis of suspect image files (Qadir and Noor, 2021). However, historically tools

have not met a usable level of automation – neither making use of advanced computing paradigms and capabilities – nor being relied upon by investigators as they conduct only base level analysis or extraction of files. Almost all analysis of image files within the context of criminal justice investigations makes use of human investigators to review tool findings, including when those tools themselves are the elements performing analysis. The purpose of these reviews is to corroborate and verify the findings of the tools, and ensure that evidence has been correctly located, or negated as the situation warrants, a process that has been called *Human-In-The-Loop (NPCC, 2026).* However, this also introduces several issues such as the potential for error and the time taken, as previously discussed (Platzer, Stuetz and Lindorfer, 2014).

This practice, whilst appropriate for the security checking of tool results, has the result of curtailing meaningful gains for the criminal justice sector from the advancement of technology and automation capability. Potential for gains in terms of time, efficiency and for preventing harm to investigators in respect of their need to review potentially harmful material that has been located as part of an investigation, are lost when there is a need for a human to manually verify the results of that automated capability.

However, as technology in the realms of AI and ML-based models has increased in complexity and reliability, there has been a recent move towards adoption of those technologies, with tools such as *SkinSheriff* working to integrate computational analysis which would allow for a consideration of removing the human element from criminal analysis (Platzer, Stuetz and Lindorfer, 2014).

As a result of the above, the researcher has coined the term *'True Automation'* to describe a conceptual paradigm in which ML and AI software systems are designed, and relied upon, to conduct their specific function or task entirely without human intervention. In this context, '*True Automation*' refers to solutions that achieve their stated objectives independently, removing the need for any human involvement in the process.

This differs from traditional RPA as discussed above in that the novel term allows for the use of, and reliance on, AI and ML based intelligent tools without the need for

human review, whereas traditional RPA simply allows for the conducting of repetitive actions without any intelligent feature set or analysis.

This is the term that will be used throughout this thesis.

1.3 Problem Domain Justification

Within the domain of DF, there are multiple disciplines and areas that could be considered – the domain is wide ranging and covers multiple disciplines (Solomon *et al.*, 2011).

As such, when reviewing the scope of this thesis, as well as the limitations and applicability of any findings from the research, and the final produced framework, it is necessary to narrow the scope of this to a defined domain within the DF field.

In order to achieve this, the author's own knowledge and experience within the law enforcement domain, as well as primary research in the form of a targeted focus group was used.

1.3.1 Focus Group to determine Problem Domain

The focus group that was conducted had two significant criteria for the directing of the discussions that were held.

- Objective 1 - The most common use of automation within the DF field
- Objective 2 - The challenges facing the DF field

These factors directed the discussions that were held at the time of the group taking place.

In order to ensure that the results were captured in both reliable and usable format, the questions that were recorded on a *Notion* document, and the responses were recorded as they were spoken. The raw notes that were taken during the course of the focus group are contained as *Appendix 1*.

Although it would have been preferable to record this directly, this was not possible due to ethical considerations and the wishes of the participants.

Once the results of the focus group were obtained, the results were analysed and the key themes were identified, which were used to answer the two objectives that are outlined above. This is detailed in the following sections.

1.3.1.1 Participant Profiles

The focus group that was conducted contained eight members, made up of various professionals involved in the Staffordshire Forensic Partnership (SFP).

These members included serving Police Officers, Supervisors, Retired Investigators and teaching staff from Staffordshire University involved in delivering DF-based material.

At the time that the focus group was conducted, a specific decision was made to keep the sample size specific to the SFP, this led to the stakeholder profile being limited in diversity and viewpoint.
Although this constitutes a potential limitation, this decision was specific to the intended purpose of the focus group, which was not to create generalisable findings but to provide early-stage, practitioner-informed, justification for the selection of an appropriate problem domain for the thesis.
Accordingly, it was decided to keep the participant pool restricted to operational practitioners. This ensured that discussions remained grounded in real-world investigative requirements, avoiding the potential for the dilution of findings through commercial interests, vendor perspectives or broader technological considerations unrelated to operational practice. Specific, detailed, personal data was not obtained from the participants, as this was not required and would have led to concerns around the ethics of the research and was also a subject of concern from the participants.

Although this could lead to concerns regarding participants suitability, and their ability to comment on the questions and topics that were posed, their presence within the SFP means that they are involved in the DF field to the point that these concerns are not problematic. Hence, no additional information was required. There are no considerations being made regarding length of service, or difference in opinion based on role or experience level and therefore, although it was considered, no additional information was required nor recorded.

The participants have been coded, with a breakdown of the different positions shown below in Table 1.

Table 1 - Participant Mapping for SFP Focus Group

Participant Organisation	Participant Number	Position
University of Staffordshire	1	Teaching Staff
Law Enforcement (Non-Officer)	2	SFP Administrative Staff
Law Enforcement	3	Investigative Supervisor
	4	Serving Investigators from a variety of investigation units involved in the investigation of sexual offences.
	5	
	6	
	7	
	8	

1.3.1.2 Questions and Format

Within the focus group session that was held, the researcher asked the group several open-form questions, making use of the Tell, Explain, Describe (TED)-style open question format to encourage open form answers, with more specific questions asked to obtain direct answers where required (Tümen Akyildiz and Ahmed, 2021).

This was done as, at this stage of the project, qualitative data was required in order to capture the key themes that were being identified by the participants – as the aim of the focus group was to obtain detailed information about how automation is used within DF, as well as the domains that are most commonly seen, qualitative data was decided to be the most effective way of ensuring that usable information was obtained (Morgan, 2018).

The following is a list of the questions that were asked, including an explanation as to the rationale for each question;

- Tell me about any use of automation that you are aware of within DF.
  This question was intended to obtain information about the different uses of automation within real-life criminal justice investigations. This was the primary question for discussion, as if there was a domain that was the one encountered by most participants, this would be a valid consideration for the focus of the framework.
- Are you aware of the use of volatile memory within DF?
  This question was designed to investigate a potential avenue that was being explored at the time of the focus group.
- Describe any concerns that you would have regarding using an automated tool for conducting analysis of DF sources.
  This question is intended to support the literature review investigation that is being conducted in this thesis and explore the challenges that are facing the implementation of automation within the DF field and obtain real-world information to support the secondary research that is showcased within Chapter 3 of this thesis.
- Describe what requirements you would place on such a tool, should you need to rely on it.
  This question is designed to begin the process of developing the framework that is the subject of this thesis and provide initial information to support the later findings of the in-depth survey and interviews that were conducted. This is discussed later in this thesis, within Chapters 2 and 4.
- Should such a tool be able to meet these requirements, would you be willing to rely on that tool for conducting analysis?
  This question is designed to investigate the hypothesis and research questions posed in Chapter 1.6 and 1.7 and begins to consider whether the framework developed would have any real-world impact.

As shown, several of the questions that were discussed within the focus group are related to other chapters of this thesis. The analysis of this will be contained in Chapter 1.3.1.3 below but will also be referenced in the relevant chapters later in this thesis.

1.3.1.3 Results, Analysis and Conclusion

The focus group that was conducted was held on 30th August 2023, the notes and information that were obtained at the time are shown as *Appendix 1* to this thesis.

From these results, the following key themes were identified through qualitative analysis of the responses that were given.

- *Question 1 – Use of Automation*
  Although there was one mention of analysis and identification of fingerprints, the overwhelming majority of participants within the focus group stated that the primary use of automation that they were aware of within DF was the analysis of image files.
  They provided several examples of tools that are currently used for this purpose, including CAID, ADF and Laseri X. These tools are discussed later in this research.
- *Question 2 – Volatile Memory use*
  The participants stated that they were aware of the use of volatile memory sources, but this is very limited, if it is used at all.
- *Question 3 – Concerns regarding automation*
  The participants within the work gave a variety of concerns regarding the use of automation in DF. These concerns ranged from accuracy, developer bias ion reporting and regulatory issues. These concerns are identified and discussed in more depth in *Chapter 5*.
- *Question 4 – What requirements would you have for an automated tool?*
  Several requirements were identified for the tool, including;
  - Accuracy (100%)
  - Proven to save time
  - Readable outputs
  - Legal compliance
  - Confirmatory Testing
  - Declaration from manufacturer

    These requirements are discussed in more depth in *Chapter 5*.

- *Question 5 – Would you be willing to rely on an automated tool if these requirements were met?*

  The participants in the focus group stated that unanimously that they would be happy to rely on an automated tool if all of the above requirements were met.

From these responses, the following conclusions were drawn;

- Objective 1 – In relation to objective 1 of the focus group, the overwhelming response from the participants was that the analysis of image files is the area that they are most commonly seeing automation being used within the definition that has been provided.
- Objective 2 – The challenges facing the use of automation within the DF field have been identified as above.

The primary conclusion to the focus group in relation to this specific section of the overall work is that the most effective domain for this thesis is analysis of image files.

This is further supported by the research that is conducted within *Chapter 3*.

1.4 Ethics Statement

The ethical concerns of this research have been considered by the author, with the relevant documentation found within *Appendix 2*.

1.5 Research Aim

To develop a novel framework to allow for the assessment of the suitability of wholly automated Digital Forensic image analysis tools to the Criminal Justice process within the United Kingdom.

1.6 Hypothesis

A reliable framework, based on the requirements of law enforcement professionals, would increase investigator confidence in the use of automation-based tools, and would allow for their use within criminal justice investigations in the UK.

This hypothesis has been tested by developing the described framework in collaboration with identified industry practitioners and stakeholders, as described later in this thesis, and deploying said framework to experts working in the field,

namely serving police officers and investigators of various forms from within an English Police Force and obtaining real-world expert opinion on the suitability and impact of the developed framework. The finding for this testing is contained within *Chapter 9.2*.

1.7 Research Questions

In the pursuit of this thesis, several questions were posed which were identified by the researcher following an initial review of the existing literature and uses of automation within the DF domain. The questions were selected to answer key questions around the research aim given in *Chapter 1.5*, whilst being actionable and achievable in the given timeframe and with the existing resource constraints of the project. From these considerations, the following research questions were identified:

1) What issues are facing the DF industry, which may lead to the implementation and development of new automated tools, which can be used to generate court-admissible evidence?
2) How has ML been integrated into DF?
3) What requirements would be placed upon a fully automated ML based solution, for it to be relied upon by the Criminal Justice sector in the UK?
4) Can a framework be developed by which ML-based tools could be judged to ascertain their applicability to Criminal Justice?

1.8 Research Objectives

To achieve the aim and research questions that have been posed in this thesis, the following objectives were identified.

1) To investigate state-of-the-art approaches, solutions, and frameworks to identify current issues facing the DF industry, how the issues are being addressed, and how automation/ML is being used within the Digital Forensic Industry.
2) To determine what requirements would be placed upon a ML solution to allow for that tool to be relied upon in a criminal justice environment and meet these requirements.

3) To design, develop and implement a framework by which ML tools could be judged to ascertain their applicability to Criminal Justice in conjunction with real-world practitioners.
4) To evaluate and validate the efficiency and performance of the proposed framework against state-of-the-art solutions.

1.9 Deliverables

The following are the deliverables for this thesis.

1) Results from primary research showing the requirements placed upon ML based tools from the Criminal Justice sector.
2) Analysis through primary & secondary research showing the current performance levels of common MLAs when applied to Digital forensic-based tasks.
3) Conceptual operational framework to judge the applicability of MLA-based solutions to Criminal Justice and provide evidence and reassurance of that applicability if a tool meets all criteria set out.
4) Analysis and testing, via primary means where possible, of current automated Digital Forensic tools to see if they meet the developed framework, once complete.

1.10 Framework Definition and Project Design

The artefact in this case is a novel framework for assessing the applicability of an automated tool to law enforcement investigations.

The concept behind this thesis is to develop the framework in conjunction with law enforcement, with the overall aim being that the framework will comprise a set of minimum standards and criteria against which a tool can be assessed in terms of function, results, and reliability. Once a tool has passed these criteria, law enforcement would have a rationale and a basis on which they could rely, to justify any reliance on the results of automated analysis via that tool.

The inputs and design of the artefact in this case have been summarised below in Figure 3:

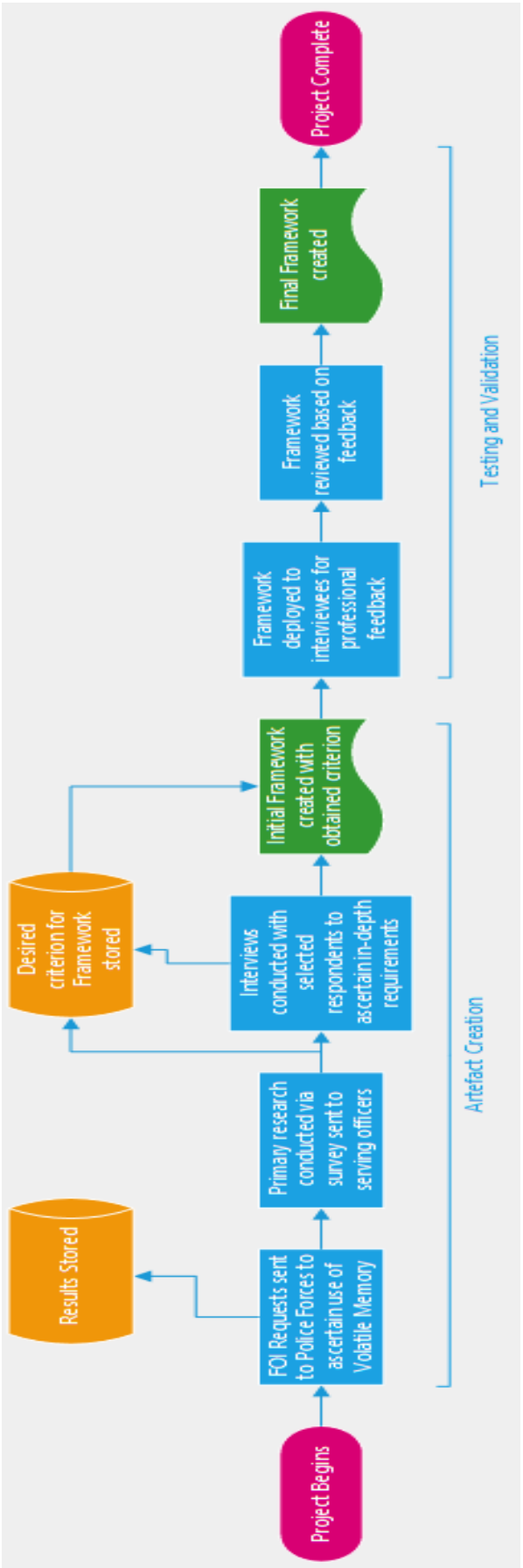


Figure 3 - Artefact Design and Project Process

Further to the design of the project itself, and how the information is going to be obtained, the artefact itself is defined.

A framework, in this context, is defined as the underlying structure behind a system (Cambridge Dictionary, 2022). In this case, the framework which constitutes the artefact is aimed at being a method by which law enforcement practitioners, and forensic investigators, can be assured that a tool, having met the criteria and standards set out, can be relied upon as a replacement for human manual review, and that the results of that analysis are as reliable, accurate and technologically sound as possible, meaning that their case / investigation is not jeopardised as a result of reliance on that tool. The framework additionally means that a tool need not be assessed on every case, if it has been deemed to have passed the framework, then for the purposes of image analysis within forensics, it can be relied upon.

It has been useful to conceptualise this idea as a series of 'gates' similar to logic gates such as AND, OR and XOR gates in computer function. The actual 'gates' themselves, as well as the metrics by which they are judged to have passed or not, are what will be determined with the assistance of practitioners in the surveys and interviews which are discussed later in this thesis – the framework is the collection of these gates concluding with an assessment of the tool.

1.11 Contributions to Knowledge

Within this thesis, there are six contributions to knowledge that have been identified, shown in Table 2, below:

Table 2 - Contributions to Knowledge

NUMBER	CONTRIBUTION	LOCATION	JUSTIFICATION
1	A detailed framework and assessment guidance, by which current and future ML based Digital Forensic solutions can be judged to ascertain whether they are applicable to the Criminal Justice sector and can be relied upon in Court situations. The framework will be validated by deployment to operational experts within the field, as described within the Research Methodology section of this thesis – and the obtaining of relevant feedback on the deployment suitability and suitability of the framework to the field, from said experts. This Framework would benefit both researchers in the field as well as developers in this industry, by providing the metrics, passing levels, and considerations that practitioners in the field would have for the tools that are being developed. Additionally, this Framework would benefit all individuals and organisations involved in this type of work by providing a proof of concept to show that assessment to the level of '*True Automation*' is possible, and that there is conceptually	Figure 40, *Chapter 5.4*	This contribution addresses the identified research gap – that being that there is no unified framework or assessment method against which *True Automation* powered DF tools can be assessed for suitability and reliance within CJ. Existing works do not provide a structured mechanism for assessment, nor is there a focus or serious consideration to how *human-in-the-loop* requirements could be removed. The framework therefore contributes new knowledge by introducing a practitioner-informed assessment mechanism specifically designed to bridge this gap.

	a point at which the removal of the *human-in-the-loop* component would be possible.		
2	A novel metric (Practical Reliability [PR]) to assist with the assessment of automated, ML and AI based tools within the DF domain, with the terms defined in *True Automation*. This will benefit all professionals involved in this domain as a key identified issue is the lack of a unified lexicon for automation technologies, which leads to confusion and a lack of understanding on the part of non-technical users and assessors when applied to a law enforcement domain. Additionally, this metric will assist assessors and practitioners by providing a clearer combinatory metric for the easier assessment of complex technical terms and tools.	*Chapter 5.3*	Existing assessment metrics for automation tools can provide evidence of the capability of such tools. However, these can be complex to calculate and require technical knowledge and skill to interpret, which may not always be available in the practitioner assessors tasked with assessing such tools within law enforcement and CJ. Accordingly, the PR Metric is designed to synthesise multiple technical metrics into a single, operationally effective and efficiency score which can be understood and applied rapidly by non-technical assessors.
3	Coining and definition of a novel term – '*True Automation*' – and updated definitions of technical terms in a novel lexicon for cross-domain working between AI, ML and DF. This will benefit both practitioners and developers in providing a definitive definition and objective to which both sides can work and will prevent confusion in documentation and discussions as the use of technology improves and increases.	*Chapter 1.2*	Within the current lexicon for commercial, academic and operational domains, the terms AI, ML and Automation are used in isolation. With particularly the technological domain using these terms to refer to specific functions. This can create ambiguity in relation to the level of human interaction required for the use of such systems.

			The novel term contributes to knowledge in providing a singular, defined term which can be used by all domains, ensuring understanding across the fields and knowledge levels involved. This is a contribution due to the current lack of such a cross-domain definition.
4	Findings on the lack of a defined and shared technical lexicon across technical and legal disciplines. This will benefit all professionals in this domain in the same manner as *3.*	*Chapter 7.0*	The thesis provides evidence showing the lack of a shared lexicon between technical and operational staff. From the research that has been conducted, this has not been a subject of significant research or analysis and as such and as such, the findings from within this thesis represents a contribution to knowledge which can be used within future research.
5	The results of targeted and comprehensive surveys / interviews to key stakeholders in the law enforcement sector (*See Chapter 2*) pertaining to the requirements and attitudes towards ML based solutions and the move towards the need to be able to rely on automated analysis. This will benefit professionals in this domain by providing a concept as to initial thoughts, requirements and blockers in	*Chapters 5.2 & 4.3*	This contribution provides empirical evidence regarding practitioner attitudes, concerns and requirements surrounding the adoption of AI and ML technologies within the operational DF domain. By capturing this information, the thesis contributes to knowledge by providing novel organisational,

	respect of the use of '*True Automation*' and will provide practitioners a clear assessment level, as well as providing developers with the information against which they will be assessed when tools are deployed.		legal and practical conditions which would be required by operational practitioners.
6	A comprehensive analysis of the current state of AI and ML use within a Digital Forensic setting. This will assist all professionals in this domain in the same manner as *5*.	*Chapters 3 & 4*	This thesis contributes a comprehensive assessment of the current state of AI and ML approaches within DF. This is a contribution as existing research on this topic is fragmented with various sources and often lacks a CJ focus. This thesis provides a clearer understanding of current technological use and capability and establishes an evidence base for future development and evaluation.

The primary contribution to knowledge in this thesis will be the framework that is outlined above. The concept behind the creation of this framework is to provide law enforcement, at strategic level, as well as the legal system, a basis on which they may be able to rely on an automated analysis tool as a replacement for human analysis.

The benefit of these contributions to the industry is a definitive move towards the acceptance of '*True Automation*' tools within the DF domain, enabling assessment of such tools to allow practitioners to reap the benefits of novel technologies, and showcase that there is a path of professionals in this highly sensitive field to make use of technology in a more effective way in the future.

Additionally, the framework seeks to show technical professionals this same pathway to acceptance so that work can be undertaken to ensure that tools are developed with the specific and unique criterion of the DF domain in mind, which will further enable professional access to the benefits that novel technologies can provide.

The intention is that following this research, professionals in the field both within DF and in software/application development will be able to work collaboratively to produce effective and usable technical solutions to ease the ongoing and increasing workload and burden upon law enforcement in respect of technical analysis of digital evidence sources.

1.12 Project Mapping

Table 3 contains a matrix showing where each objective is mapped to the aim, research questions and deliverables in this thesis:

Table 3 - Project Mapping Matrix

OBJECTIVE	AIM	QUESTION	DELIVERABLE
1	✓	✓ – Questions 1 & 2	✓ – Deliverables 2 & 4
2	✓	✓ – Question 3	✓ – Deliverable 1
3	✓	✓ – Question 4	✓ – Deliverable 3

4	✓	✓ – Question 4	✓ – Deliverable 4

1.13 Structure of Thesis

This thesis has been structured into 9 chapters, as demonstrated within Figure 4 *(below)*.

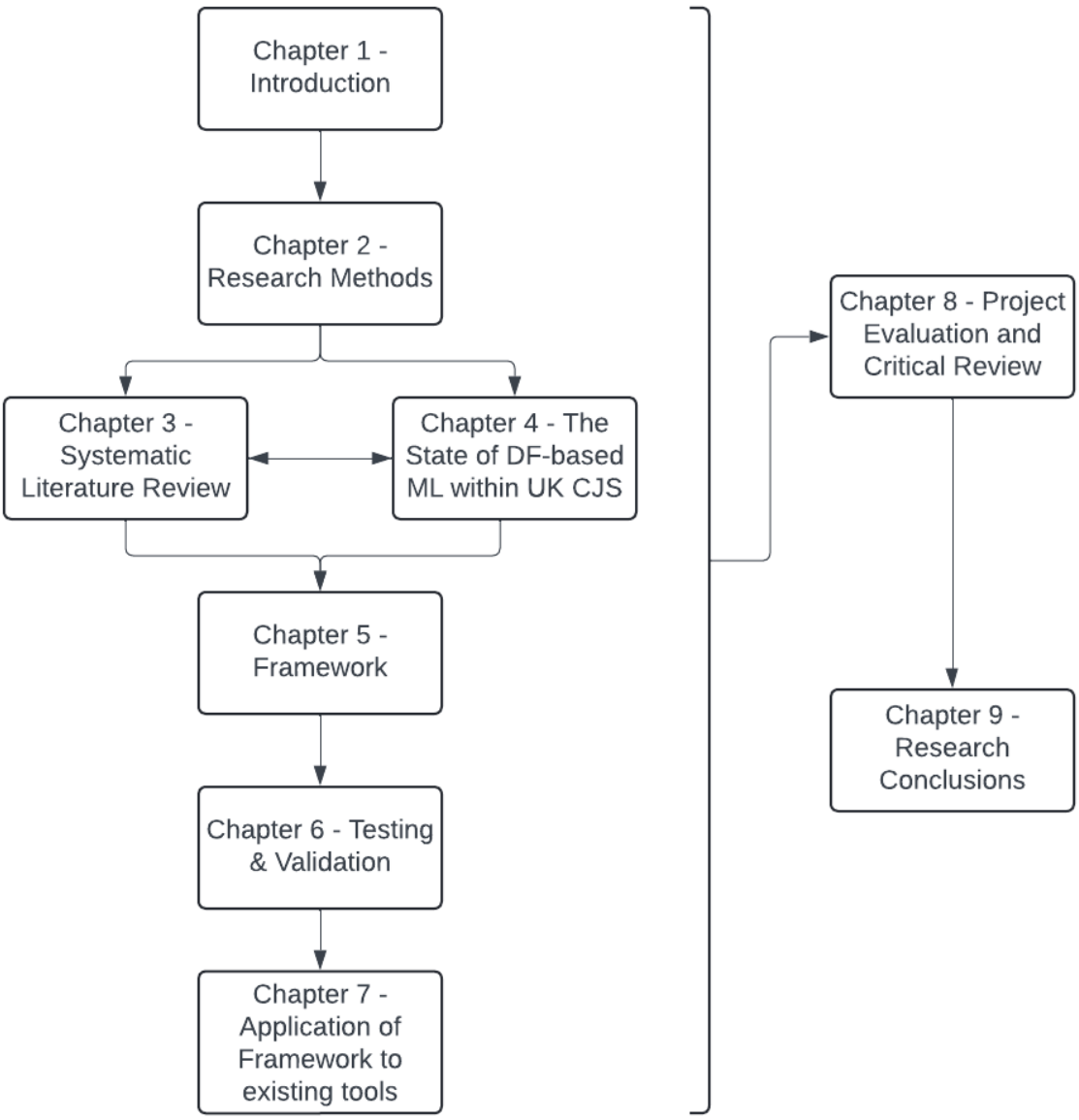


Figure 4 - Thesis Structure

The contents of these chapters are summarised, thus;

Chapter 1 – Introduction

This chapter introduces the project, and provides the problem domain and background, providing context for the remainder of the thesis. It further defines the aim of the project, along with the hypothesis that is being investigated and the objectives and deliverables of the project.
Additionally, this chapter provides the contributions to knowledge in the project and explains some of the key concepts.

Chapter 2 – Research Methods

This chapter explains and explores the research methods and methodologies that were employed during the project, as well as documents the decisions and plans that were made in the conducting of the project.

Chapter 3 – Systematic Literature Review

This chapter contains the literature review that was conducted as part of this thesis, exploring the key concepts that surround the problem under investigation, as well as the existing solutions and works that have been conducted within this problem domain, as well as the wider DF field. The chapter also includes a review of the existing frameworks and requirements placed on DF image analysis tools currently.

Additionally, this chapter looks at the ethical and legal considerations that exist around the use of automation technologies within the DF domain.

Finally, this chapter contains an analysis of this literature, leading to an explanation of the research gap that was located in this thesis, as well as how this relates to the later work in this thesis.

Chapter 4 – The State of Digital Forensic-based Machine Learning Within Criminal Justice in the UK's Criminal Justice Sector

This chapter examines the real-world state of DF automation within the UK Criminal Justice system, focusing on both organisational and practitioner perspectives.
It details the findings from Freedom of Information requests sent to UK Police Forces.

Additionally, the section analyses the responses to the disseminated survey discussed in *Chapter 2.4.2*.

Chapter 5 – A Framework to Implement Digital Forensic-based Machine Learning within the UK's Criminal Justice Sector

This chapter showcases the framework that has been created during the project, as well as the analysis of the primary research that has been conducted to provide justification for the criterion that that have been given within the framework.

Additionally, this chapter contains the supporting information and documentation that has been created to assist practitioners in investigating and understand the framework, and the criterion that have been included.

The chapter also includes a discussion of the metrics used to assess the accuracy of ML/AI tools, as well as a critique of these metrics, leading to the introduction of a novel assessment metric in the *PR Metric*.

Chapter 6 – Testing & Validation of Framework

This chapter contains the testing and validation that was conducted on the framework to ensure that it is fit for purpose, as well as an explanation as to the method of validation that was conducted.

Additionally, this chapter contains an analysis of the findings of this validation.

The chapter also contains the testing and evaluation of the proposed gate passing metrics for the developed framework.

Finally, this chapter also contains an explanation of how the framework was applied to existing tools from within the field to ascertain whether it was suitable in real-world applications, as well as whether there were any accessible tools that could meet the criteria that were given.

Chapter 7 – Testing & Validation of PR Metric

This chapter details the testing and evaluation that was conducted in both expert and practical terms of the proposed PR Metric in this thesis.

It discusses the method of testing, as well as the rationale for this method selection, as well as the results of that testing.

Finally, the chapter ends with an analysis of the testing that was conducted and the results in the context of the project as a whole.

Chapter 8 – Project Evaluation and Critical Review

This chapter contains an evaluation of the project, as well as a critical reflection on the project that has been conducted.

Chapter 9 – Research Conclusions and Recommendations

This chapter concludes the project and contains an analysis of the findings and results obtained to map these to the aim, hypothesis, objectives and deliverables of the project.

It also provides recommendations on how the fields of DF and Computing can work together to encourage further adoption of novel and developing technologies.

1.14 Chapter Summary

In summary, this chapter has established the motivation and context for the research by showcasing that the DF domain is facing an increase in the volume of devices involved in investigations, the evidential complexity of those devices is increasing and that these issues come together to present significant resource pressures for DF units across the world.

The findings from the initial focus group identified image analysis as the most appropriate domain for the thesis and highlighted key practitioner requirements for automated solutions including accuracy, validation, transparency and efficiency gains.

The chapter has also made an initial contribution to knowledge through the introduction and definition of the novel concept of *True Automation,* describing a paradigm in which AI and ML systems are relied upon to conduct analysis without human review.

Finally, the chapter has established the research aim, hypothesis, objectives, deliverables, and wider contributions to knowledge that underpin the remainder of the thesis.

In combination, all these elements have provided the conceptual foundation for the subsequent thesis and development of a framework to support the acceptance of *Ture Automation* within the DF domain.

2.0 Research Methods

This chapter explores the research methods and methodologies that were adopted for the project, as well as the strategies and approaches that were used, with appropriate rationale. Within the project, there are several research methods and methodologies that have been utilised, and which will be utilised, at various points. For the purposes of clarity in this thesis, the terms *research methods* and *research methodologies* are defined thus:

Research Method – The research method that is employed is the actual manner and process by which the research has been conducted (Goundar, 2012).

Research Methodology – The research methodology is the overarching strategies and concepts which explain the methods that were employed (Goundar, 2012).

These topics are explored further in the following sections.

2.1 Research Methodology

The "*Research Onion*" as provided by (Saunders, Lewis and Thornhill, 2019) has been adopted as a guide for the research methodological approach.

This model is shown below (Figure 5), with the selected methodologies highlighted.

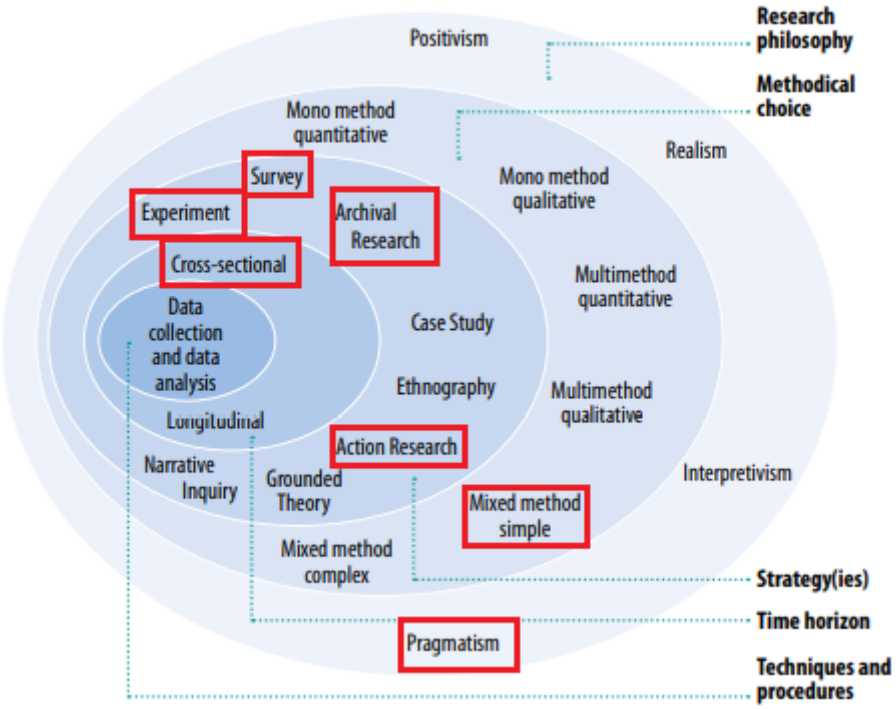


Figure 5 - Research Onion (Saunders, Lewis and Thornhill, 2019)

The utilised methodologies within this thesis have been so chosen based on the definitions from Saunders and Tosey (2013), and are explained as follows, other definitions and sources are highlighted where they provide additional justification:

Philosophy – Pragmatism

The problem which is being addressed within this thesis is a practical one, the primary issue being the lack of a structured assessment method. The focus of the thesis is producing an effective operational solution to the given problems, rather than exploring theoretical or philosophical questions. As such, pragmatism is the most appropriate methodological choice.

Accordingly, the study in this case does not utilise a specific ontology or epistemology, focussing instead on a practical approach to the choices made. The focus of the work is on the practical applicability of the given solutions, as well as their real-world effect on the given domains of study and work.

The researcher utilised different approaches as was required for the project and are explained below, rather than sticking to a singular approach.

Methodological Choice – Mixed Method Simple

The research contained within this thesis requires both qualitative and quantitative practices. This is due to the requirement for insight into various aspects of the DF process, assessment of practitioner sentiment and technological assessment.

Accordingly, the methodological choice in this thesis was *Mixed Method Simple* – the reason for this was that the work will cover qualitative methods in the literature review that was conducted, as well as the focus groups that have been held, following by quantitative methods that were conducted in the form of the survey and interviews that were conducted. As the sequence of these approaches was set, this is the most appropriate approach under the given definitions (Adeoye-Olatunde and Olenik, 2021). However, it was also appropriate based on the feedback from practitioners to include elements of experimentation and expert review later in the project, despite the fact that these were outside of this sequence, although this is not ideal, it was necessary for the project findings to be the most valid and acceptable as possible.

Strategies – Action Research

The development of the primary contribution in this thesis, that being the novel framework required collaboration with operational practitioners, with the developed gates being based on real-world requirements and refined through practitioner feedback.

Accordingly, the primary approach that was adopted in this case was Action Research. However, during the course of the project it was necessary to adopt several other approaches as the research developed and findings were analysed and included within the thesis. The different approaches that were utilised occurred in the sequence that they are given below, with some exceptions where the research required, such as the combination of archival research and experiment that was required in *Chapter 7.2* when it was discovered that the technical information required for appropriate testing of the *PR Metric* was not available.

These approaches are documented below.

Archival research was used for the review of existing works and applications of ML to DF, and Action Research / Survey as professionals in the field will be used to develop the framework, and their opinion and input will be obtained via the use of surveys and interviews.

Experiment, as well as archival research, will be used to test existing Digital Forensic implementations of ML to ascertain whether these already meet the created framework.

Time Horizon – Cross-Sectional

This thesis sought to obtain attitudes, requirements and concerns at a specific point in time to provide a practical solution for that time period. The objective of the thesis was to develop and validate a particular artefact, rather than examine changes over time. As such, this thesis is not longitudinal. This was considered as part of this research; however, this was discounted due to the time constraints of the participating organisations and the rapidly developing nature of the domains being explored. As a result, a cross-sectional study was determined to be the most effective within these constraints and allowed for a more effective study to take place with the artefact being developed being more streamlined and focussed on the developing problems that the domain is facing, and which are explored later in this thesis.

Techniques and Procedures – Data Collection and Data Analysis

The data collection conducted within this thesis, both quantitative and qualitative, will be structured to ensure that each of the specified aims are achieved, shown in *Chapter 2.4.1*. All data collected will be reviewed to ensure that it is scientifically valid and that therefore, any analysis of that data or conclusions drawn from it are valid.

2.2 Objective / Strategy Mapping

The following table (Table 4) shows the mapping of the objectives to the different research strategies:

Table 4 - Objective & Methodology Mapping

Objective / Aim	Method/Methodology	Specific Information
Aim	Archival Research, Action Research / Survey	Archival Research will be used to develop the requirements of the framework, supported, and enhanced by the input of real-world practitioners in the field of DF.
1	Archival Research	Archival Research shall be used to understand the existing issues and implementations as described.
2	Archival Research & Survey	Surveys and Questionnaires as well as interviews were appropriate will be used to develop the framework in conjunction with practitioners in the field.
3	All	All the given research methods will be used to develop the framework as this is the overall objective of the work – all actions are towards this point.
4	Action Research & Experiment	The framework will be deployed to real-world professionals in the field to test whether it meets their requirements and would function in a real-world scenario, hence action research. Further, Existing Digital Forensic solutions and MLAs will be tested against the requirements of the framework to ascertain whether they meet said requirements, and whether the framework is applicable to relevant existing solutions.

2.3 Research Approach

The research approach adopted in this thesis was phased and made use of a deductive approach. Deductive was selected as this is the most appropriate for this thesis, in which the researcher started with a predetermined theory, in this case the hypothesis that was discussed in *Chapter 1.6*, and then sought examples towards or

away from this theory rather than starting from nothing and building the theory based on the findings (Young *et al.*, 2020).

The project was broken down into three phases, to ensure that there was a structured and methodical approach.

These phases are explained below:

- **Phase 1 – Research**

The purpose of this phase of the research was to provide an academic basis for the project, explore existing problems and solutions in the field and to provide a practical basis for the problem domain selection, ensuring that the produced framework was suitable for the workplace, and was directed to the most appropriate domain.

- Phase 2 – Development

The purpose of this phase of the research was two-fold;

First, to ascertain the current state of the use of automation within the DF field, making use of primary research and reviewing existing systems and procedures from within the Police Service, to determine the current state of automation within the DF domain.

Secondly, to develop the actual framework and accompanying documentation, along with the novel PR Metric, that the project was intended to create. This was achieved in collaboration with serving officers from an English Home Office Police Force. The results of the consultations held with serving investigators were combined with the existing research and developments in the field of automation to construct a robust set of criteria that constitute a framework to move towards the acceptance of DF automation.

- Phase 3 – Testing

The testing phase of the project was intended to validate the artefacts that were developed within this thesis and takes two forms.

First, the framework is tested by deploying this to experts in the field – the officers that provided consultation on the criteria to include within the framework, to ensure that it is suitable for their needs and is appropriate to be deployed. In addition to this,

expert review of the gate passing metrics was conducted to ensure that the guidance values provided were suitable and appropriate.

Secondly, a series of existing automation tools were considered with the framework to ascertain whether these solutions passed the criterion provided, and whether the criterion themselves were applicable to those solutions.

The PR Metric was then tested in the same manner, to ensure consistency and validity.

The purpose of the testing phase of the research is to ensure that the product of the project is valid and is suitable for analysis and conclusions to be drawn.

2.3.1 Project Management

To ensure that the project was managed appropriately, several project management systems were researched, and an *Agile* management style was selected.

The rationale for this being the selected project management methodology, was that the project would need to be adaptable to change, rather than adhere to a rigid and defined process that moved from one aspect to another, as is required for a *Waterfall* method (Thesing, Feldmann and Burchardt, 2021).

This thesis relied heavily upon the input of professionals working in a changeable environment, that being the investigators and officers who would be responding to the surveys, and conducting the interviews and focus group required for the completion of this thesis. The nature of their work required a flexible approach as dates and times of actions may have to change at short notice, and the project could not always wait for them to become available again.

As a result of this flexibility, the agile approach was determined to be the appropriate choice in this case, as it allowed for concurrent working on objectives (Thesing, Feldmann and Burchardt, 2021).

To manage the project as it progressed, a Gantt chart was initially utilised to ensure that the relevant tasks were conducted in a timely manner, and that the overall project was kept relevant and was ran efficiently. A Gantt chart was selected as the

most relevant method of planning for this thesis as it is the best suited to a cross-sectional, time limited project with a linear development pathway (Bianconi, 2024).

An example of this initial Gantt chart can be seen in the below Figure 6:

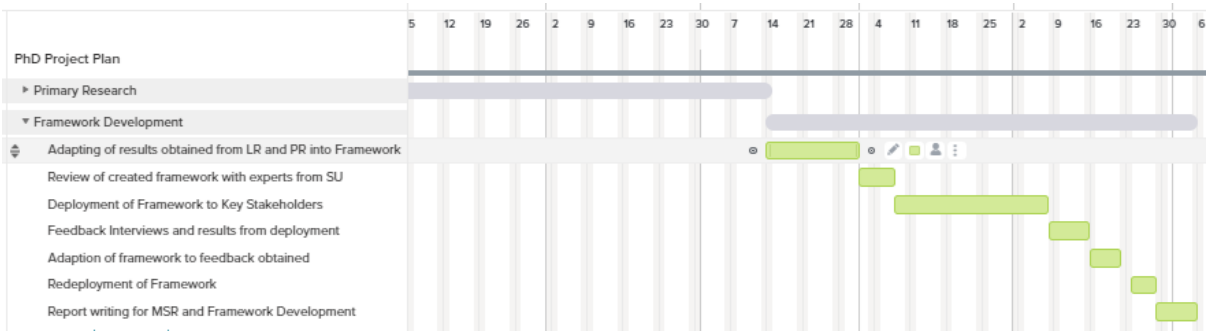


Figure 6 - Initial Gantt Chart

However, as the project developed, this was found to be of little effective use to the management of the project.

As such, the phased approach described above was adopted fully and visualised to assist the researcher in time management. This visualisation of each phase and how these phases interrelate is shown below in Figure 7:

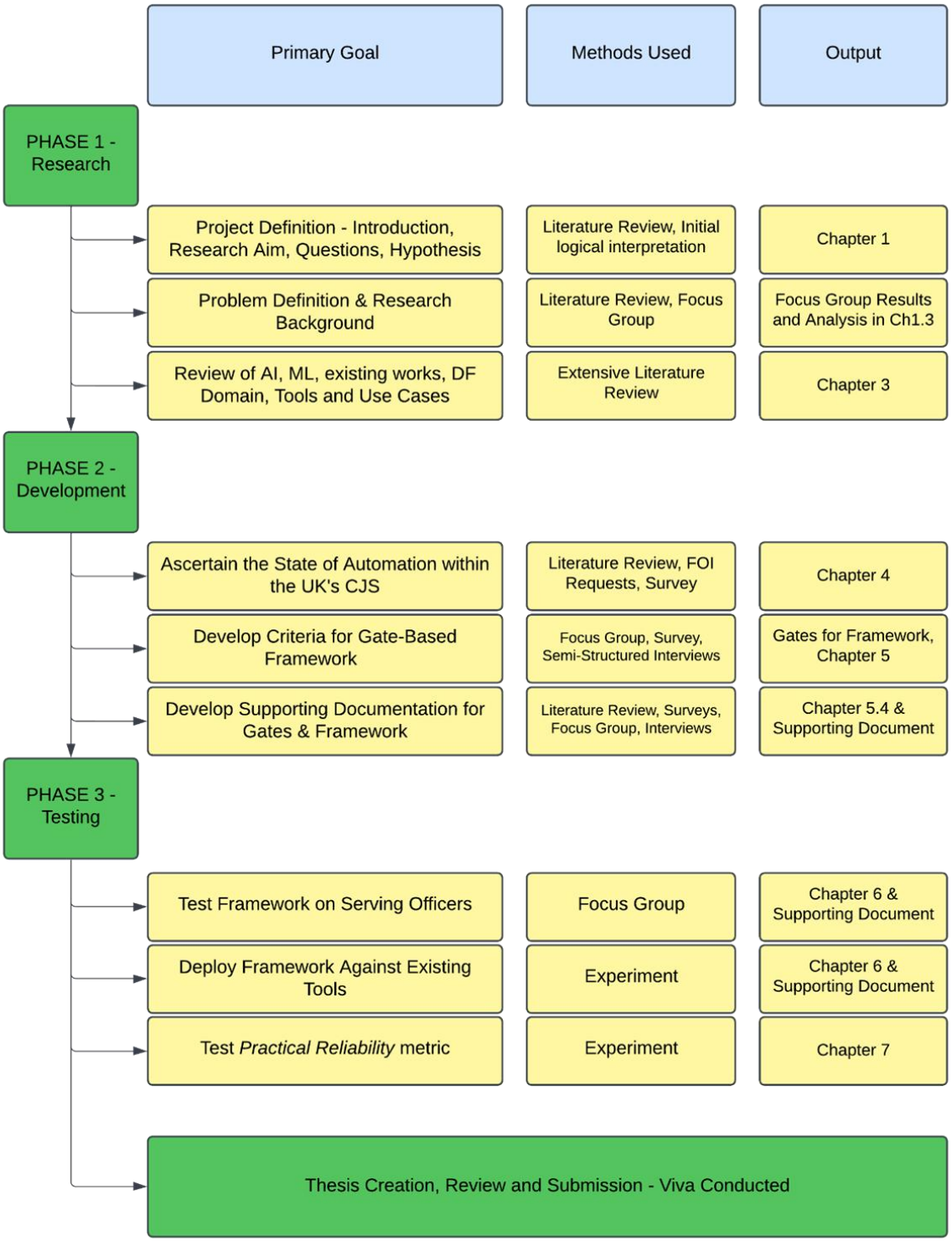


Figure 7 - Project Management

Following the completion of this phased approach, the thesis was written, checked and submitted.

2.4 Research Method & Research Design

Within this thesis, there are several research methods that have been adopted. This approach consists of a combination of methods – including experimental, archival and action research, designed to obtain a range of quantitative and qualitative data, as explained in each element, found below and in *Chapter 2.1*.

The approach for the research in this thesis, specifically in relation to the creation of the PR Metric and Framework artefacts, relies heavily on significant stakeholder engagement as the thesis places strong emphasis on co-creation and co-design. This is due to the thesis seeking to develop operational artefacts rather than purely academic constructs.

As a result of this approach, practitioners from the DF domain were actively engaged throughout the research through focus groups surveys, interview and panels. The input from these actions was used to identify requirements, refine developed criterion and gates, and evaluate the suitability of produced artefacts.

In addition to this, the adopted methods are explained per phase below:

2.4.1 Phase 1 – Research

This phase primarily encompassed the completion of the secondary research that was necessary to obtain the background and existing work in this area, as well as the existing legal and ethical constraints which would inform the later artefacts produced.

Specifically, this involved the conducting of a systematic literature review, that is found in *Chapter 3* to this thesis.

Additionally, a focus group was conducted, which is explained within *Chapter 1.3*.

The systematic literature review that was conducted in this thesis takes the form of *Archival Research* as defined in the Research Onion (Figure 5) and consists of a review of existing relevant and time-relevant research.

The process that was followed to conduct this review is shown below in Figure 8.

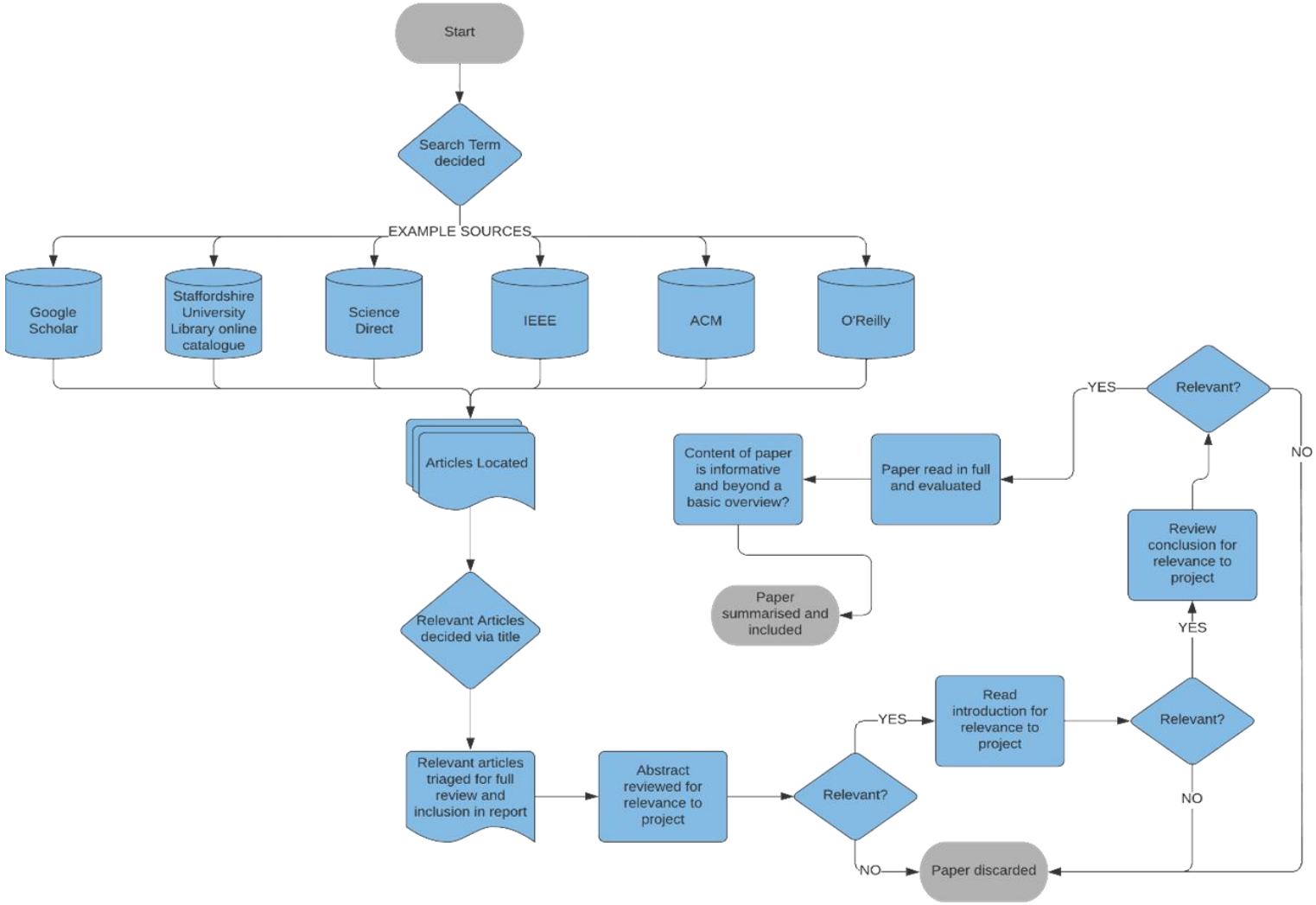


Figure 8 - Literature Review Process

When conducting this review, there were several criteria that were placed upon the articles that were reviewed to determine relevancy to this thesis, these are shown and explained below in Table 5.

Table 5 - Literature Review Criterion

Number	***Criterion***
1	*Peer-reviewed source – the material must be from an academically valid, peer reviewed, source.*
2	*Relevant topic – The topic of the material must be relevant to the topic of this thesis, which is judged by the title and abstract.*
3	*Relevant conclusion – The conclusion of the material must be valid and make a point or argument that is relevant to this thesis.*
4	*Relevant and informative content which goes beyond the basic – the content of the material should not be a basic overview of a topic and should have some in-depth analysis.*

Only material that passes all criterion was used in this thesis, material that did not meet all the requirements was discounted.

In respect to the age of the material, whilst this factor was considered, there were occasions wherein older material was required to demonstrate a particular issue; therefore, this was not counted as a set criterion. However, the age of the material was also considered when determining its relevance and whether it would be included in the work or not. In general, material that was over five years old was discounted, unless it was the only source of information or was particularly relevant.

2.4.2 Phase 2 – Development

In the second phase, the focus of the work moved to the creation of the framework itself, as well as the documentation for that Framework, and the PR Metric that was produced to work within it.

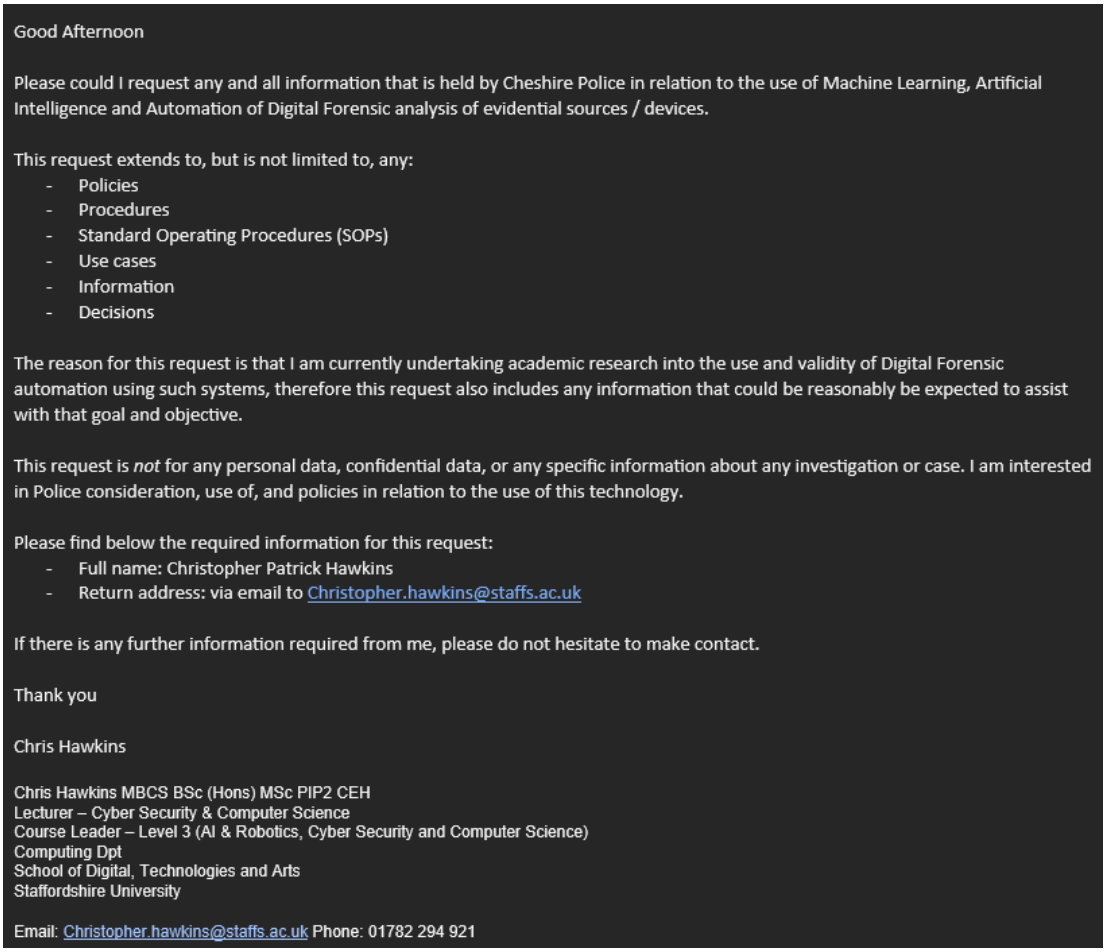
There were several primary research methods that were employed in this phase;

- FOI Requests

Freedom of Information (FOI) requests were sent to every police service in England and Wales (as defined by the Home Office) to ascertain if they have any specific information on the use of automation, ML or AI within Digital Forensic Investigations.

It was decided that this request should remain broad to capture any use of automation, rather than focusing on image analysis as it is considered likely that any image forensics, or analysis, will be contained within a larger Digital Forensic Standard Operating Procedure (SOP) and keeping this broad will allow for ease of access to material.

An example of the wording used in these requests has been included in the below Figure 9.



Good Afternoon

Please could I request any and all information that is held by Cheshire Police in relation to the use of Machine Learning, Artificial Intelligence and Automation of Digital Forensic analysis of evidential sources / devices.

This request extends to, but is not limited to, any:

- Policies
- Procedures
- Standard Operating Procedures (SOPs)
- Use cases
- Information
- Decisions

The reason for this request is that I am currently undertaking academic research into the use and validity of Digital Forensic automation using such systems, therefore this request also includes any information that could be reasonably be expected to assist with that goal and objective.

This request is *not* for any personal data, confidential data, or any specific information about any investigation or case. I am interested in Police consideration, use of, and policies in relation to the use of this technology.

Please find below the required information for this request:

- Full name: Christopher Patrick Hawkins
- Return address: via email to Christopher.hawkins@staffs.ac.uk

If there is any further information required from me, please do not hesitate to make contact.

Thank you

Chris Hawkins

Chris Hawkins MBCS BSc (Hons) MSc PIP2 CEH
Lecturer – Cyber Security & Computer Science
Course Leader – Level 3 (AI & Robotics, Cyber Security and Computer Science)
Computing Dpt
School of Digital, Technologies and Arts
Staffordshire University

Email: Christopher.hawkins@staffs.ac.uk Phone: 01782 294 921

Figure 9 - FOI Request Wording

In the main, this is considered to be a quantitative method, the rationale for this being that the main objective of sending this survey out is to ascertain which forces

will respond, which have the information requested and which – if any – actually have policies and procedures in place to consider the widespread use of automation within DF.

As such, the responses were quantitatively analysed to determine any statistical trends which would be of interest to this thesis. Such analysis of the provided data was conducted via SPSS Statistics, using appropriate statistical methods. The setup for the data entry via SPSS is shown below in Figure 10.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Force	String	50	0	Police Service	None	None	34	Left	Nominal	Input
2	Response	Numeric	8	0	Response	{1, No Res...	None	8	Right	Nominal	Input
3	RefusalRe...	Numeric	8	0	Code	{1, Cost T...	None	8	Right	Unknown	Input
4	Sent	Date	10	0	Date Sent	None	None	8	Right	Scale	Input
5	Notes	String	10000	0	Notes	None	None	53	Left	Nominal	Input

Figure 10 - SPSS Variables Setup

The explanation for each setting is as follows:

+ Police Service is listed as Nominal as there is no defined order or category for this data, it is simply a list of names. It is set as a string as it is simply alphanumeric characters to denote the name of the organisation.

+ The response is listed as a string as it is a response, there are set values given as follows:

1 – No Response – If the organisation does not respond to the request for information

2 – Refused – If the organisation responds but refuses to provide the information. Further notes will be recorded as the reason for the refusal as this can vary.

3 – Response, no information held – If the organisation responds but states that it does not hold any information that is relevant.

4 – Response, information provided – If the organisation responded and provided the information that is requested.

This setting is shown below in Figure 11.

Value Labels:

Value	Label
1	No Response
2	Refused
3	No information held
4	Information Provided

Figure 11 - Variable Setting within SPSS

+ Refusal Reason – This is listed as a nominal string variable as there are several possible reasons, with no particular order, that a force may refuse to answer the questions that are asked, this has been coded to assist in analysis of the reasons for refusal, as follows:

1 – Cost Too High – If the organisation states that the retrieval of the information requested, or if the identification of whether it exists, is too high and exceeds the statutory limits.

2 – Public Interest – If the organisation determines that the release of this information would be detrimental to law enforcement or is not in the public interest

0 – Not Applicable – If this code is not required

This setting is shown below in Figure 12:

Value Labels:

Value	Label
0	Not Applicable
1	Cost Too High
2	Public Interest

Figure 12 - Variable Setting within SPSS for Refusal Code

+ Sent – This is logged as a date variable, as the logical setting.

+ Notes – This is listed as a string as several notes will be recorded here for auditing and analysis purposes

[Sent and Notes had no analytical value; these were for monitoring and auditing purposes]

The findings of this method are explored and analysed within *Chapter 4*.

- Survey

Following the dissemination of the FOI requests, a targeted survey was disseminated to serving officers of an English Home Office Police Force, outside of the SFP which is discussed in *Chapter 1.3*. This force has not been named due to privacy concerns from the officers involved. The purpose of following the FOI requests up with a targeted survey was to obtain more specific and detailed information, from individuals who are dealing with live investigations in a real-world setting. The combination of this data, as well as the FOI request data, ensures that the analysis conducted on this data contains both organisational and individual input.

For the survey, it was decided that specific units would be contacted regarding this, to ensure both a large enough response rate as it is believed that a targeted survey to specific units will avoid respondent apathy and prompt a larger response rate. The specific units that were used have not been identified within this thesis in order to protect participant anonymity. However, all units which participated work in relevant areas to the project and conduct investigations which heavily rely on image analysis within the DF field or conduct forensic analysis and review of image files from seized suspect devices.

The survey that was sent was designed to be completed in a short order (around five minutes) and was comprised of yes/no or short form answers. Whilst it is accepted that this produces questions that are mostly closed in nature, limiting participant input in some respects, ensures, as much as possible, the highest possible response rate whilst ensuring that the responses can be analysed in an appropriate fashion and the minimise the time impact on the participants, with the aim of keeping a high response rate.

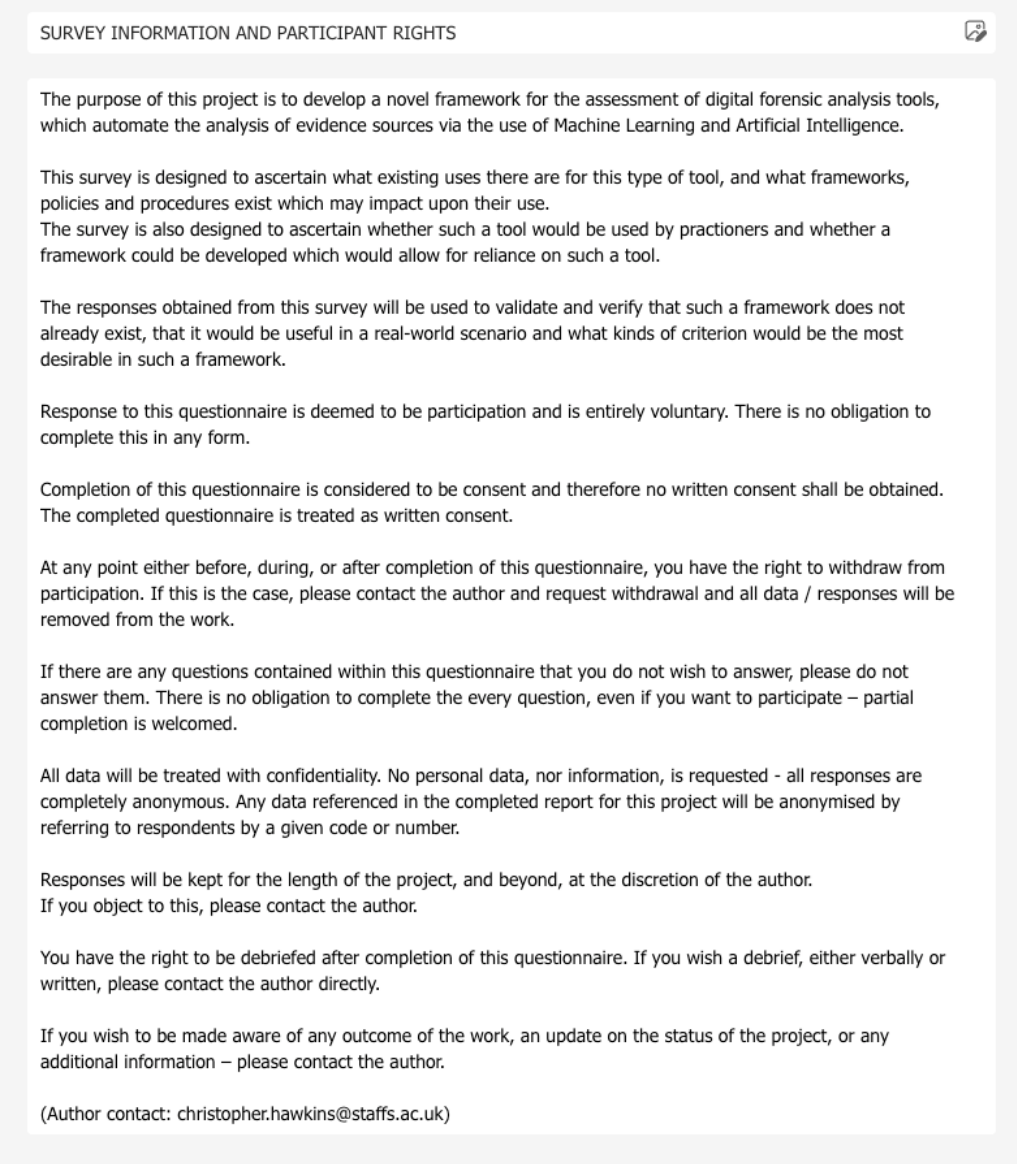
The survey was disseminated using Microsoft Forms so that it could be distributed online, ensuring that staff who are working from home or who are on leave when it is sent could access it when it suits them.

A four-week response period was given to ensure that there was sufficient time for

officers to respond to the information, whilst also ensuring that the research proceeded at an acceptable pace.

As before, the results of this survey were recorded and analysed using the SPSS software. All ethical considerations will be given to the dissemination of the survey in line with university regulations, participants will be fully anonymised to all parties, no personally identifiable information was recorded, and all participants were provided with a full information sheet at the time of signing up.

The information sheet that was presented to participants on accessing the survey is shown below, as Figure 13 *(Sample)* – The full copy of the information sent to participants is shown below in *Appendix 3*.



SURVEY INFORMATION AND PARTICIPANT RIGHTS

The purpose of this project is to develop a novel framework for the assessment of digital forensic analysis tools, which automate the analysis of evidence sources via the use of Machine Learning and Artificial Intelligence.

This survey is designed to ascertain what existing uses there are for this type of tool, and what frameworks, policies and procedures exist which may impact upon their use.
The survey is also designed to ascertain whether such a tool would be used by practioners and whether a framework could be developed which would allow for reliance on such a tool.

The responses obtained from this survey will be used to validate and verify that such a framework does not already exist, that it would be useful in a real-world scenario and what kinds of criterion would be the most desirable in such a framework.

Response to this questionnaire is deemed to be participation and is entirely voluntary. There is no obligation to complete this in any form.

Completion of this questionnaire is considered to be consent and therefore no written consent shall be obtained. The completed questionnaire is treated as written consent.

At any point either before, during, or after completion of this questionnaire, you have the right to withdraw from participation. If this is the case, please contact the author and request withdrawal and all data / responses will be removed from the work.

If there are any questions contained within this questionnaire that you do not wish to answer, please do not answer them. There is no obligation to complete the every question, even if you want to participate – partial completion is welcomed.

All data will be treated with confidentiality. No personal data, nor information, is requested - all responses are completely anonymous. Any data referenced in the completed report for this project will be anonymised by referring to respondents by a given code or number.

Responses will be kept for the length of the project, and beyond, at the discretion of the author.
If you object to this, please contact the author.

You have the right to be debriefed after completion of this questionnaire. If you wish a debrief, either verbally or written, please contact the author directly.

If you wish to be made aware of any outcome of the work, an update on the status of the project, or any additional information – please contact the author.

(Author contact: christopher.hawkins@staffs.ac.uk)

Figure 13 - Survey Information Sheet

The full question list of the survey that was disseminated is contained within *Appendix 3*, and what follows is an explanation as to the questions that were written, and their rationale:

Questions 1 – 6 encompass consent and participant information necessary for the administration of the survey and analysis of the results.

Questions 7 – 19 constitute the topic-relevant data collection questions.

Question 7 – Age

This question is designed to investigate whether there is any link between age and

the embracing of technology within investigations and ascertain whether there is any trend around developing investigative strategies.

Question 8 – Length of Service

This question is similar to *Question 7* – the intent with this question is to ascertain if there is any trend towards acceptance of technology or not when compared with experience and length of service.

Question 9 – Image Analysis

This question is designed to ascertain whether the participant had any experience in the domain that is under investigation. Participants that answered this question negatively would be discounted from inclusion in the study, as relevant input is required.

Question 10 – Awareness of ML and AI-based automation

This question is intended to ascertain whether investigators are aware of the use of automation within their investigative strategies, and organisational policies. This also covers whether investigators are even aware of the technological possibility of this type of automation, as well as the use of it in live investigations.

Question 11 – Tool awareness and Use

This yes/no question is designed to ascertain whether real-world investigators are aware of the application of the concept asked about in *Question 10*.

Question 12 – Use of Tools

This yes/no question is designed to assist the author in ascertaining whether investigators, who have answered whether they are aware of the existence of tools in *Question 11*, are aware of the actual implementation of these tools in real world investigation. This is designed to support any analysis of tools in later chapters in determining whether automation is actually used in the DF domain, as opposed to whether this is simply an academic awareness or whether the increase in technological capability is already resulting in real-world improvements.

Question 13 – Frameworks and Minimum Standards

This yes/no question is designed to support the identified research gap in this thesis, and the subsequent contribution to know that is defined in *Section 1.11* and *Table 2*.

Any positive responses to this question will be subject to analysis to ascertain what moves have been made in the field to define a set of minimum standards for the application of automation to the DF domain. The results of this question, and analysis of subsequent identified systems, will assist in the development of the resulting framework for this thesis.

Question 14 – Specific Tools

This question is intended to assist the author in identifying automation tools that are currently in use with the DF domain, to support any findings and analysis from the Systematic Literature Review (SLR) found in *Chapter 3*.

Additionally, the results of this question will allow the researcher – through investigation into the named tools – to ascertain the currently used definition of the term *automation* as held by the real-world investigators asked within this survey.

Question 15 – Willingness to make use of automation

This question is designed to in-part investigate the hypothesis found in *Chapter 1.6*. The purpose of this question is to ascertain whether investigators would be willing to make use of *'True Automation'* within their investigations.

Question 16 – Would a framework be of use in changing attitudes?

This question is designed to continue the investigation into the hypothesis given in *Chapter 1.6*.

The question is designed to ascertain whether real-world investigators are willing to consider adopting '*True Automation*', should the framework produce in this thesis become mainstream in application and use. It is also designed to ascertain whether a framework would be of practical use in changing the opinion of those who answered negatively to *Question 15*.

The results of this question, as well as *Question 15,* will also provide credibility and relevance to the findings of this thesis as a whole, as well as provide relevancy to the selection of the problem domain.

Question 17 – Willingness to rely on automated tools

This question differs from the previous ones in that it is asking whether investigators would be willing to *rely* on the results of automation, rather than simply make use of a tool as part of an investigative strategy.

This question is important in defining the distinction between the *use* of a tool and *reliance* on the results of such analysis.

The results of this question are designed to move towards answering several points from within *Chapter 1* of this thesis.

Question 18 – Concerns regarding reliance on automation

This question concerns the framework that is the titular element of this thesis and serves to assist in the identification of any concerns that are held by real-world officers, to ensure that these are addressed by the eventual end-product of this thesis.

This question is designed to support the findings of the later stage interviews, and provide a basis for testing and validation analysis of the provided work.

Question 19 – Framework Requirements

This question is designed to begin the process of considering the criteria that are required for the titular framework of this thesis.

The question is designed to support the later interviews conducted in this thesis and ensure a wide scope of responses including those that could not participate in those interviews.

Once the survey had been disseminated and results obtained, they were entered into *IBM SPSS Statistics* dataset, in the same manner as the previous FOI requests.

The codification and variable entry for this is shown below in Figure 14:

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	ImageAnaly...	String	8	0	Image Analysis?	{1, Yes}...	None	8	Center	Nominal	Input
2	Aware	String	8	0	Aware of AI or ...	{1, Yes}...	None	8	Center	Nominal	Input
3	Tools	String	8	0	Aware of Tools?	{1, Yes}...	None	8	Center	Nominal	Input
4	Use	String	8	0	Aware of tools ...	{1, Yes}...	None	8	Center	Nominal	Input
5	Frameworks	String	8	0	Any Framework...	{1, Yes}...	None	8	Center	Nominal	Input
6	Willingtouse	String	8	0	Willing to auto...	{1, Yes}...	None	8	Center	Nominal	Input
7	WillingFram...	String	8	0	Willing to use a...	{1, Yes}...	None	8	Center	Nominal	Input
8	Framework...	String	1000	0	Would you rely ...	{1, Partially ...	None	8	Center	Nominal	Input
9	Concerns	String	1000	0	What Concern...	None	None	8	Center	Nominal	Input
10	Framework...	String	1000	0	What requirem...	None	None	8	Center	Nominal	Input

Figure 14 - SPSS Survey Variables

Most questions, as explained above, in the survey are – by design – Yes/No. This was intended to maximise the number of responses by minimising the amount of time taken for the completion of the survey.

In relation to *SPSS* itself, the rationale for each variable selection in this case is as follows:

- Image Analysis through Framework Acceptance is listed as a nominal string with 1 – Yes and 2 – No coding, as these are – by design – Yes/No questions- they are recorded using nominal measures as there is no inherent order or ranking and Yes/No responses are categorical.
- The Concerns and Framework responses are also recorded in the same manner; however, these are longer strings as the themes are recorded here for later thematic analysis rather than statistical. These have been kept in the same document for ease of reference and analysis.

The analysis of the results of this survey can be found where it is relevant, and primarily in *Chapters 4, 5 and 6*.

- Interviews

Following completion of the surveys in relation to this thesis, the relevant information was extracted as shown in the results that are discussed later in this thesis.

It was then necessary to consider the selection of the criterion that would make up the titular framework for this thesis.

In relation to the survey, by the nature of the questions that were asked, by design, and explained above – the results of the disseminated surveys were quantitative in

nature, which provided a good basis for analysis as later described, however, for the determining of the criterion itself, and to allow for a deeper understanding of the particulars of investigations, the use of tools and the concerns and requirements of real-world officers, qualitative data was required as this provides opportunity for deeper-level thematic analysis of data and more thorough understanding of underlying issues around a topic than quantitative analysis alone will allow (Isselé, Heinze and Fenton, 2024). Additionally, it is known that making use of qualitative interview strategies to support and supplement quantitative analysis can yield a deeper understanding of a problem domain and the problem being investigated (Adeoye-Olatunde and Olenik, 2021).

To achieve this, it was determined that the most appropriate method would be to conduct semi-structured interviews. To ensure that these interviews were properly and effectively conducted, the "*Seven Steps"* approach found within Adeoye-Olatunde and Olenik (2021) was taken, and is shown below in Table 6:

Table 6 - Seven Steps Approach (Adeoye-Olatunde and Olenik, 2021)

Step	**Sub-Topic**
1 - Assess appropriateness of the semi-structured interview: Best method to address research objective(s)?	None
2 - Sampling and Participant Recruitment	2a – Sampling Approaches 2b - Recruitment
3 - Data Collection Design	3a – Developing the semi-structured interview guide 3b – Collecting participant demographic information
4 – Conducting the interview, transcription, and data transmission and storage	4a – Preparation and training 4b – Interview modality and recording considerations 4c – Transcription and Checking

	4d – Securely storing and transmitting data
5 – Data Analysis	5a – Coding and Theme identification 5b – Establishing rigour
6 – Drawing conclusions	None
7 – Reporting Results	7a – Reporting Guidelines 7b – Data Displaying

In accordance with the "*Seven Steps*" strategy, targeted interviews were designed and conducted under the following strategy which addresses the points raised in Table 6.

Participants were selected from units that had been involved in the completion of the survey, to ensure relevancy to the project, and had the appropriate experience and knowledge in the relevant areas so that they could comment in a realistic manner.

The interviews were conducted live and in-person to allow for more in-depth and personal conversations to take place, as this is preferable to online interactions wherein participants can feel less comfortable (Guest *et al.*, 2020).

The interviews were arranged with the unit supervisors and were conducted on the same day to ensure that there was no crossover of participants.
All participants in the interviews were anonymised via the implementation of a numbering system, with only the researcher having access to the participant mapping document. The sole purpose for the recording of participant names was to ensure that should a participant wish to withdraw; their data could be removed at any time without concern that data was erroneously included or removed.

All results were recorded on individual answer sheets, which were marked with the participant number only. The researcher recorded the responses directly onto the documents as they were given.
There was no audio, nor visual, recording that took place during the interviews – the rationale for this was that several participants had expressed a concern about this,

and so it was decided that the only record would be the paper document to ensure maximum participation.

Further, the written recording of the responses allowed for a more concise documentation process, to ease the act of extracting key themes from the responses, without the need to review several hours' worth of video documentation.

The ethical requirements of the process were addressed, with a specific information and rights sheet being given to every participant which clearly outlined the rights and entitlements of the individual as a participant.
The complete blank documentation for the interviews can be found under *Appendices 4, 5, 6 and 7*.

The actual conducting of the interviews took the form of a *semi-structured interview*. This approach entailed the preparation of an interview '*guide*' which included structured questions written in open-ended style, whilst these provide the researcher with a guide as to the main points that are needed for the interview to be successful, the intention is not that the guide is stuck to in verbatim fashion as the focus of the style is on the natural conversation flow of the interview (Adeoye-Olatunde and Olenik, 2021).
The guide that was prepared for this can be found under *Appendix 5*.
The rationale for this method being selected is that it is accepted as being favoured for the collection of data when the purpose is to obtain an individual viewpoint rather than a generalised comprehension of a particular domain or question (Adeoye-Olatunde and Olenik, 2021).

The actual questions that were asked during the interviews are shown in the above-mentioned appendices, however, the rationale for each question is shown below;

- *Question 1 – Would the investigator consider the use of 'True Automation' in investigations*

  The intention of this question is to ascertain more detail regarding current levels of confidence that investigators have in automated tools and their use in investigations. The intention is to make use of the answers from this question to support earlier assertions within *Chapter 4* of this thesis.

Further, the questions serve to direct the conversation flow of the semi-structured interview into the domain of automation and allows the researcher to ascertain the level of understanding on the part of the interviewee.

- *Question 2 – Concerns regarding automation in investigations*

This question serves 2 purposes; the first being that the answers will support earlier answers that were provided and secondly would allow the researcher to conduct further thematic analysis on the answers to ensure that the produced framework within this thesis addresses the primary, and most common, concerns that are held by investigators, increasing the validity of subsequent testing and validation of the produced framework.

- *Question 3 & 3a – Requirements of a framework for assessing automation-based DF Analysis tools*

These questions are the primary focus of the interviews that were undertaken. These topics allow the researcher to ascertain what investigators in real-world environments would need from the framework, and what they would expect to see from any set of minimum standards that would be included in that framework, to most appropriately move towards the objective of accepting the use of *'True Automation'* within DF analysis of image files.
Additionally, the responses to these questions would enable a confirmation that the interviewee had understood the purpose of the research, and the nature of the framework that was being designed by their responses.

Finally, the responses to these questions would allow an understanding of the format and style of the framework that the investigators would need to see to ensure that they understood the criterion and to ensure that the work was usable in a real-world environment.

In order to ensure that the potential for confusion over the nature of what was being said by both researcher and interviewee, question 3a was designed to provide opportunity for expansion and clarification of points that were made. The findings of this question were also designed to extract any additional detail that the interviewee could provide on the domain under investigation.

- *Question 4 – Would the existence of such a framework provide investigators with the confidence to rely on automated tools that have passed?*

  This question is designed to ascertain whether, at the earliest opportunity, investigators would be willing to engage with the technology, should this framework be successful at meeting their criteria. The purpose of asking this question is to investigate the hypothesis in this thesis, and to assist in determining real-world validity of the work, and the produced framework.

Following the completion of these interviews – the results were entered into *IBM SPSS* as with the previous research elements. Other software was considered for this such as NVIVO for qualitative elements to the survey, however, the lack of personal information being recorded in this case led to the conclusion that SPSS and a quantitative approach was the more suitable option in this case.

The following describes the codebook that was created in order to enter the data under the correct measure;

- Question 1 – This question relates to whether the investigator would be willing to make use of automation in their investigations in any way – this, by design was a Yes/No question and therefore is coded as having a nominal measure as there is no inherent order, and makes use of a 1 – Yes, 2 – No coding as previously used for ease of data entry.
- Question 2 – This question is concerned with recording the themes that arose from the interviewee being asked about the concerns that the interviewee would have around the use of automation in their investigations. This is recorded as having a nominal measure as there is no inherent order to the responses, which were subject to thematic analysis by the researcher and the themes identified recorded – hence a string was determined to be the most suitable type for the data.
- Question 3 & Question 3a – These questions revolved around the requirements that investigators would place upon the framework that was being discussed. As with the previous question, this was determined to be a nominal string value as there was no inherent order, and the themes were

being recorded based on the analysis of the researcher to the answers that were given and recorded as described above.

- Question 4 – This final question was asking the interviewee if the creation of the framework in this thesis would be a suitable solution to the concerns that they have around the use of automation in their field. This was also a Yes/No question and so, as with Question 1, a nominal measure was selected with a coded 1 – Yes, 2 – No string type.

The codebook for this is shown in the below Figure 15:

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Use	Numeric	8	0	Would Use Automation?	{1, Yes}...	None	8	Right	Nominal	Input
2	Concerns	String	10000	0	What Concerns Exist about Automation?	None	None	8	Right	Nominal	Input
3	Requireme...	String	10000	0	What Requirements on the Framework?	None	None	8	Right	Nominal	Input
4	Solved	Numeric	8	0	Would the Framework solve this?	{1, Yes}...	None	8	Right	Nominal	Input

Figure 15 - Interview Codebook

Appropriate coding for the answers obtained in these semi-structured interviews was vital for the proper implementation of the thematic analysis of said results (Adeoye-Olatunde and Olenik, 2021).

The results and analysis for the conducted interviews can be found primarily within *Chapter 5* although there are references to this within several chapters as they are found to be relevant for inclusion.

Additionally, to the above, additional thematic analysis was conducted on the survey and focus group stages of this thesis, and the relevant themes identified from the relevant questions included as described in *Chapter 5*.

2.4.3 Phase 3 – Testing

In the final phase of the work, the completed framework and proposed *PR Metric* were tested via focus groups and technical testing to determine both its real-world validity against the expectations of the interviewees, as well as against existing accessible tools.

The results of this testing were then subject to analysis to bring the work to a conclusion.

- Action Research – Focus Group and Expert Validation

In this phase of the testing – the produced framework was deployed to real-world officers and investigators. To ensure that the participants of this research were

appropriate in terms of their experience and knowledge, the same units and officers who were involved in the interviews that had been conducted earlier in the work were asked to conduct this testing.

The format of the testing was that the framework, and accompanying documentation created within *Chapter 5* were shown to the relevant officers of the units in question – the framework was showcased by the researcher, along with an explanation of the criteria and assessment of tools that were put through the framework.
Following this showcase, a focus group was conducted with the officers to ascertain their views and feedback on the produced framework, and the criteria contained.

The ethics of this focus group was considered, and the following actions were taken:

- A full information and rights sheet was prepared and disseminated to participants. This is shown in *Appendix 6*.
- No personal information was obtained nor recorded as part of the focus group – all responses were recorded as group discussion points to ensure that no single participant was singled out.
- All required University ethics forms were completed as found in *Appendix 2*.

The rationale for this method being undertaken, as opposed to other methods such as conducting of a survey or questionnaire for feedback was that the officers involved have limited time available, and substantial amounts of this had already been requested in the completion of the earlier aspects of this thesis. Therefore, the researcher decided that this approach allowed for the maximum possible retention of respondents by condensing the feedback session to a single event and attempted to ensure that the highest number of requested officers responded by having the feedback take place at the same event.
Additionally, this format allowed the officers to ask questions of the researcher and to gain a deeper and more accurate understanding of the produced framework, rather than the potentially more limited information that could be shared remotely. Additionally, this allowed the participants to provide the required feedback at the

time of the session, rather than risking either the officers running out of time to review before the focus group and either not participating or providing unreliable data.

By the same token, this strategy was designed to be short form in the requested answers, with minimal input required from the officers. The questions that were asked, as shown below, were designed to provide the most relevant information to the researcher, whilst taking the shortest amount of time for the respondent to discuss and input.

As a result of these considerations, it was decided that this focus group style of data gathering, making use of primarily qualitative responses, with opportunity for additional detail in provided responses, was the most appropriate and effective at ensuring that the most reliable, and widespread, responses were obtained.

In this phase of the research, personal information from the respondents was not required, as all participant profiles had been checked for relevance by the earlier methods. Therefore, no personal information was requested, required nor recorded as part of this phase of the research.

The specific questions that were asked during the Action Research testing phase of this thesis are found in *Appendix 7*, this semi-structured script is for the researcher to use to guide the conversation to the relevant domains. It is not designed for participants to read directly. The rationale for the domains that have been selected is as follows.

Domain 1 – Relevance and Applicability
The purpose of the inclusion of this domain is to ascertain the participants thoughts on whether the created framework is relevant to their field of work.

The domain inclusion safeguards against the possibility of the researcher having misunderstood, or misrepresented, the concerns that were found throughout the previous methods – creating a framework that does not map to real-world investigations, or which is not applicable to the scenarios that investigators would find in real-world cases.

Domain 2 – Practicability and Feasibility

The purpose of this domain's inclusion is to ascertain whether the developed framework is applicable to real-world investigations, and whether the format and selected criterion are usable and suitable for real-world scenarios.

Additionally, the domain allows for the investigation of whether there are any additional operational factors that would influence the framework and its suitability for deployment.

Domain 3 – Usability, Effectiveness and Confidence

The purpose of this domain's inclusion is to investigate the stated hypothesis within this thesis and ascertain whether the creation of the framework in this thesis has had an actual impact on investigator's confidence in making use of *'True Automation'* within their criminal investigations.

Additionally, the domain also allows for the investigation of the understanding of the participants, and how effective the framework is at explaining technical concepts and implementations to non-technical individuals who would have to make use of the framework in live investigations, and the subsequent court cases.

Domain 4 – Feedback and Improvements

This domain's exploration within the focus group allows for the improvement of the produced framework in response to the review and input of practitioners in the field and also increases the credibility of the framework by allowing for the researcher to respond to this feedback in correcting any errors or issues that are located.

In addition to the above, additional expert validation was conducted on both the gate passing metrics that were produced as part of the framework design, as well as the *PR Metric.*

The Gate passing metrics were assessed by a single expert reviewer, who had experience and knowledge in both the legal and technical domains, as explained in *Chapter 6.3.2*. The rationale for this is included in that chapter. The feedback from that review was then adopted into the metrics to ensure validity and applicability, as well as assessing whether the passing metrics could be understood and the levels agreed with.

Finally, the *PR Metric* itself was assessed by expert review before technical testing. This is explained in *Chapter 7.1*, with the rationale being in that chapter.

The purpose of this testing is to ascertain whether the metric was understandable and clear, but also whether it achieved the stated aim of making the existing complex technical metrics easier to understand and assess themselves. Additionally, the testing with the expert allowed for assessment of whether the new metric could be easily calculated and whether the actual output was understandable in and of itself.

- Technical Testing

The final phase of testing within the work was the *Technical Testing* method.

This involved the selection of several available tools under specific criteria;

+ Accessibility

The tool had to be accessible to the researcher with no paywall to the automated analysis that it could conduct.

+ Function

The tools had to have a function, which was relevant to this thesis, that being some form of image analysis.

+ Automation

The tool had to have some element of ML or AI-based automation in their function.

Once the tools had been selected, each one was independently run through the framework, either as a paper-exercise where statistical data was required, such as accuracy analysis, or practically where this was required.

This is explained and analysed in *Chapter 6.3* of this thesis.

The purpose of this testing was two-fold;

Firstly, the experiments ensured that the framework was applicable to real-world tools that are being developed. This meant that, in combination with the feedback from the professionals in previous steps, the framework was applicable to both the

user side of analysis in the law enforcement domain, and the industry side in that the criterion was applicable to the most recent developments in tools.

Secondly, this allowed for an investigation into whether some basic and accessible tools could actually pass the framework or not, which could provide some insight into any discrepancy, or congruence, between the current state of technology and the requirements of end-users.

2.4.4 – Testing Metric Summary

The testing that has been conducted in this thesis is covered in the relevant *Chapter 6*I for the framework and *Chapter 7* for the PR Metric.

Summaries of this testing can be found in *Chapters* *6.3.1.4*, *6.3.2.2*, *6.3.3.4*, *7.1.3* and *7.2.3*.

2.5 Chapter Summary

This chapter has established and justified the methodological approach adopted within the research conducted in this thesis. The chapter has demonstrated the suitability of a pragmatic, mixed-method design for investigating the acceptance of automation within DF.

The chapter has also defined the phased research strategy used to collect, analyse and validate data from both practitioners and existing literature, ensuring alignment with the research objectives and overall research aim of the thesis.

As such, the chapter provides the methodological foundation for the development and evaluation of the proposed framework and PR Metric, ensuring that the findings are supported by structured, transparent and defensible research processes.

3.0 Systematic Literature Review

The following section will present a systematic literature review, utilising the methodology and method described in *Chapter 2.4.1*.

The chapter seeks to answer questions surrounding the project, such as why there is a need for increased efficiency within the DF domain, and how automation technology is already being used within the DF domain. Additionally, the chapter explores the existing frameworks, regulations and structures within the domain.

The chapter then ends with an analysis of the research gap that this thesis seeks to address.

Specifically, this chapter seeks to answer Research Questions one and two, Research Objective one and Contribution to Knowledge six.

3.1 The Need for Increased Efficiency in Digital Forensic Analysis

The field of DF has seen a documented requirement for an increase in efficiency for almost 20 years, as shown by (Rogers and Seigfried, 2004), who found that at that time, there was a documented increase in public reliance on technology, and that this shift was straining the existing law enforcement processes, in particular human resources. Despite this fact being first identified in 2004, the issue has remained at the heart of Digital Forensic development since this time, to present day (Sunde and Dror, 2019; Hall, Sakzad and Choo, 2022; Patil *et al.*, 2022).

The specific issues that have led to the requirement for an increase in efficiency are the reliability and timeliness of analysis, and the growing number of devices that require such an analysis during a forensic investigation – among others (Gentry and Soltys, 2019; Jarrett and Choo, 2021; Hall, Sakzad and Choo, 2022). A key additional consideration is the changing face of criminality. Recent developments have shown deep concern amongst UK judiciary regarding the use of DeepFakes both as a source of criminal action in the creation of illicit pornographic content of children or non-consenting individuals, or as a means of disrupting evidential chains with the creation of fake evidence (Hawkins, 2025).

It must also be considered that, whilst at the initial developmental period of DF it was considered to be a peripheral problem (Rogers and Seigfried, 2004) for Law

Enforcement, this is not the case in the modern world. The requirement for Digital Forensic efficiency and improvement is a multi-agency problem which affects the entirety of the Criminal Justice sector (Dardick and Lau, 2005), a key problem with approaches taken since this first identification is that although this fact has been known for over seventeen years, there is still a lack of a truly cohesive and unified approach to DF. The focus still being on the technical side of analysis (Bhatt and Rughani, 2017; Mohammad and Alqahtani, 2019).

This is not to say that other aspects have been ignored, as there have also been attempts to improve the knowledge of cyber-investigators (Hunton, 2011), as well as research and proposals to move the view of the DF field to be process-driven, even cloud-based – even going as far as to change DF to be *DF as a Service (DfaaS)* (Du, Le-Khac and Scanlon, 2017). However, these attempts have been unsuccessful as shown by the fact that the same issues persist and are in fact increasing year-on-year (Butterfield *et al.*, 2018; Sunde and Dror, 2019).

This issue was investigated further in 2021 by the researcher in this thesis via a survey disseminated to serving Police Officers from within an English Police Force – this thesis is described in Appendix 8; however, this did show that these same concerns are still an issue with real-world practitioners.

As can be seen from the results below, of seven officers surveyed, 71.4% were concerned by the number of devices that are being submitted for analysis during investigations (Figure 16*)* and 100% were concerned by the amount of time taken for the analysis of Digital Forensic sources (Figure 17*)*.

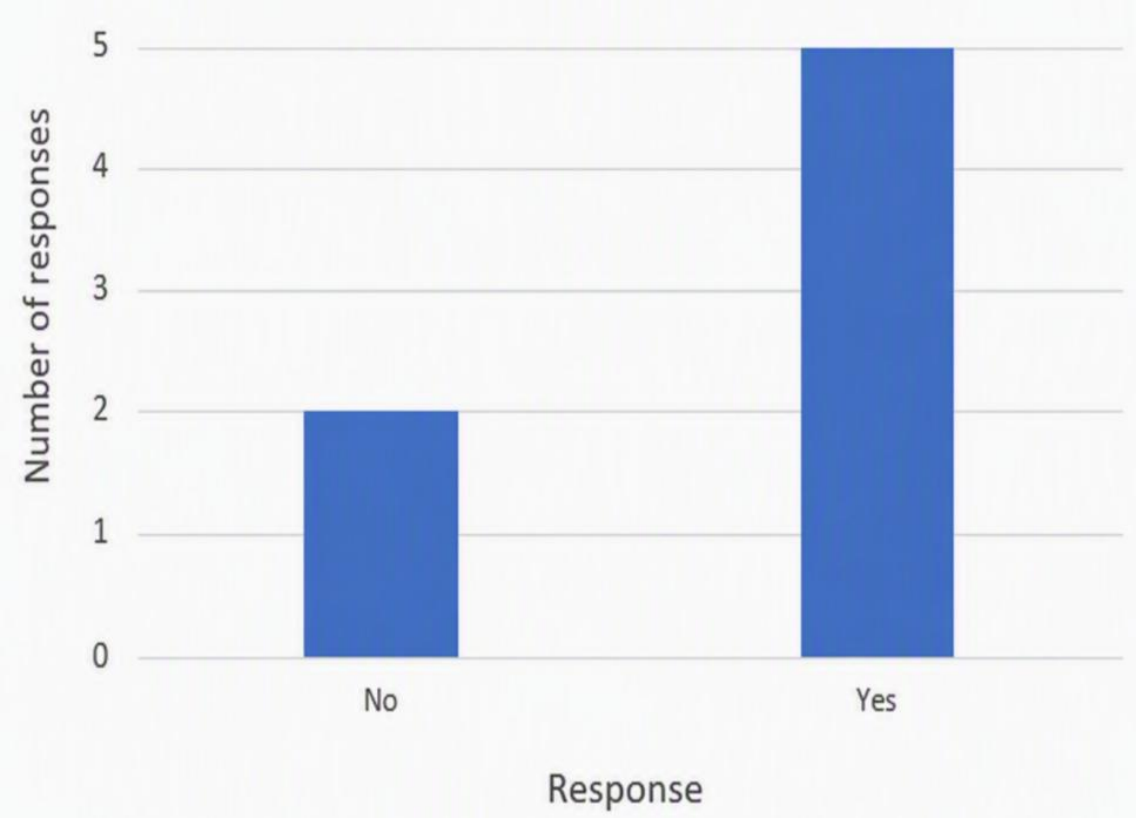


Figure 16 - Number of Officers Concerned about the number of devices being submitted for analysis

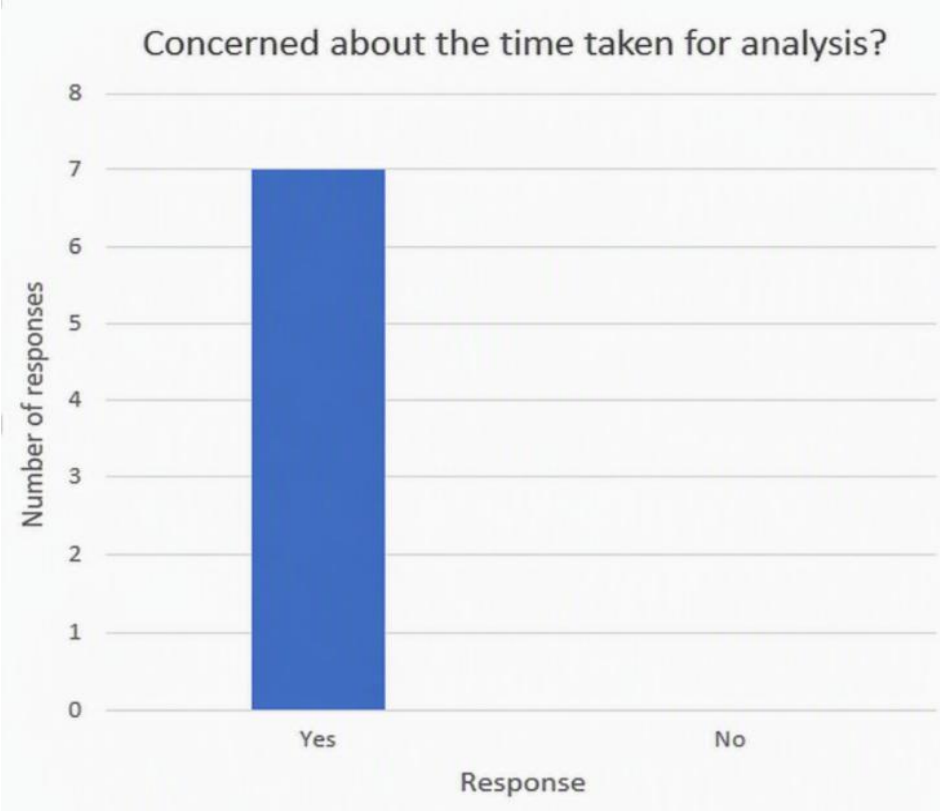


Figure 17 - Number of Officers concerned about the length of time being taken for analysis

3.2 Machine Learning Algorithms and Artificial Intelligence

In response to the above issues, there has been a recent shift in the focus of the Digital Forensic community towards the implementation of ML, AI and automation within Digital Forensic analysis tools (Jarrett and Choo, 2021), as well as across the Cyber Security landscape (Geetha and Thilagam, 2021).

This shift will be discussed in the next chapter to this thesis, however, before this is addressed the broad definitions of the technologies that are being discussed will be given, although an in-depth analysis is beyond the scope of this thesis.

The terms AI and ML will be used extensively in this thesis, and although they are often used within the same context, they refer to disparate and distinct technologies that perform different functions, and which operate in very different ways. An example of this, and how the different technologies interact is shown below in Figure 18.

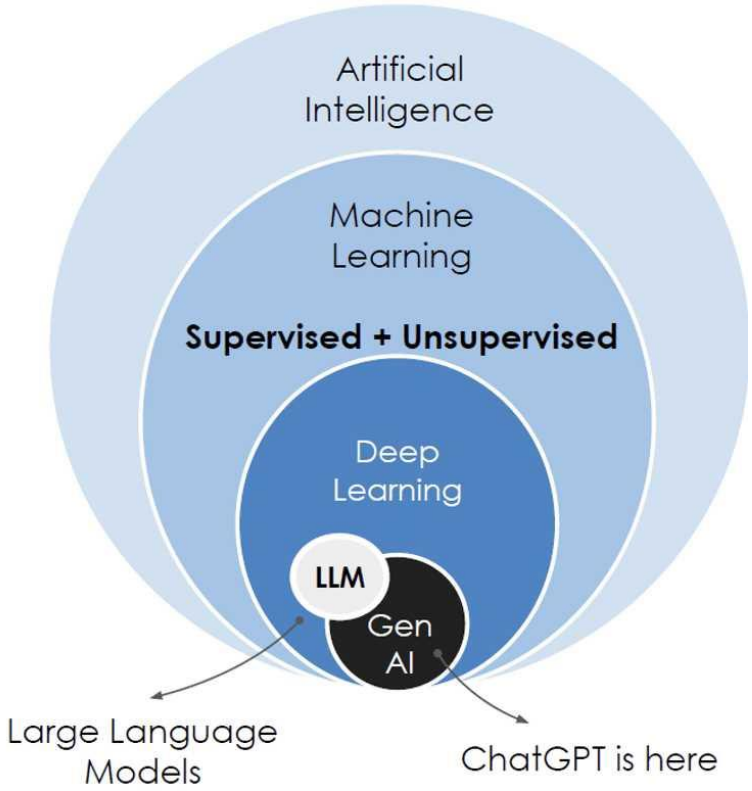


Figure 18 - AI and ML Venn Diagram (Malandruccolo, 2024)

For the purposes of this thesis, the definitions that are being used relate to:

Artificial Intelligence (AI): A system that is designed, as far as is technologically possible, to mimic human intelligence and perform tasks that would require analysis, and some decision making, rather than relying solely on prediction of future events based on previously seen data. Within a DF context, AI is generally used to refer to technological solutions that mimics any intelligent behaviour rather than repeating the same actions as would be seen in an RPA solution (Malandruccolo, 2024).

Machine Learning (ML): A subset of AI technologies which make use of Machine Learning Algorithms (MLA) that have been trained on specific datasets to predict outcomes from learned patterns within the data that has been used for training. Within DF context, ML is used to describe technologies that perform tasks that are beyond RPA solutions, but do not require analysis and perform based on previously established examples (Malandruccolo, 2024).

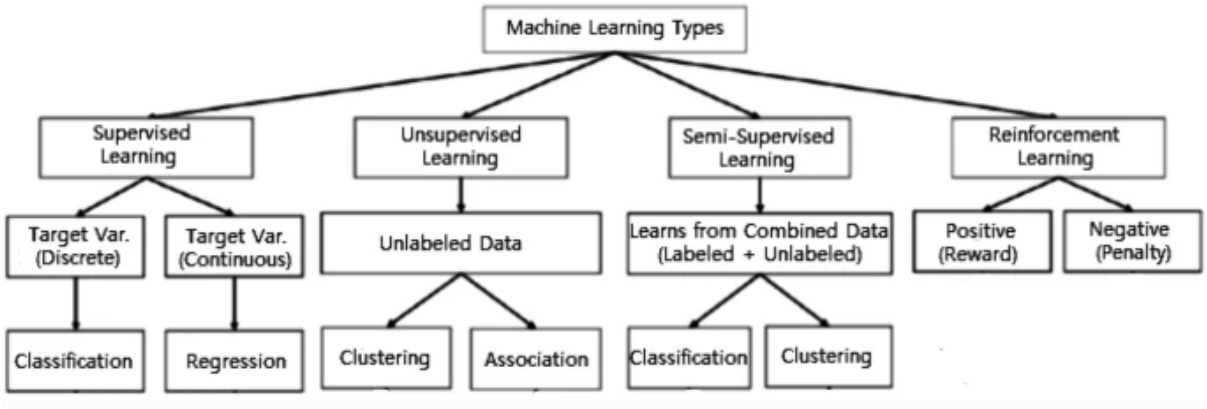


Figure 19 - Machine Types and Techniques (Sarker, 2021)

In general terms, the different types of ML techniques shown above (Figure 19) can be described thus: (Definitions used in this table (Table 7) are found in (Sarker, 2021)):

Table 7 - Machine Learning Types (Sarker, 2021)

Learning Type	**Model Building**
Supervised	Task-Driven – Learns from labelled data
Unsupervised	Data-Driven – Learns from unlabelled data
Semi-Supervised	Combined Approach – learns from both
Reinforcement	Environment-Driven approach – Learns from reward or penalty

ML and AI have seen a recent surge in use and popularity, as evidenced by the rise of ChatGPT, OpenAI and other generative-text systems (Khalil and Er, 2023). This has taken the field of AI and ML to higher levels of public awareness, however, even without this increase in popularity, ML was being used within some tools to assist in a range of areas of DF (Qadir and Varol, 2020).

The manner in which these tools are assessed is discussed in depth within *Chapter 5.3*.

3.3 The use of Machine Learning and Automation in Digital Forensics

The response to the issues that have been raised within DF as well as the wider criminal justice (CJ) sector, and shown in the previous chapters, was codified as far back as 2004, wherein Rogers (2004) said that the solution to the issues was the creation of tools. The authors in that work show that the primary method of increasing DF reliability and credibility, was the development of specific and tested tools for the assessment of digital evidence sources.

However, the work also pointed out that there was little focus on scientific rigour during the development process of the tools as well as the methods used to employ them in a real-world environment, this so-called "*ad-hoc*" (Rogers and Seigfried, 2004) approach led to concerns over the reliability of DF process and examination. Their solution to this being the increased judicial scrutiny of such procedures.

Despite the age of this initial research, it has been repeatedly stated within the field that the development of tools is the main solution to the problems – as shown in (Luciano *et al.*, 2018; Gentry and Soltys, 2019; Sunde and Dror, 2019), with the recent technological advances of automation and ML being highly desirable (Sikos, 2021). This continues to be a concern presently, with the 2025 ACE conference showing that a need for verification and testing of automated tools still remains (Hawkins, 2025).

Despite the length of time that these issues have been known, it is only recently that the research into this area has started to include bleeding edge automation technology.

Although the field of DF has not yet fully embraced the developing area of ML and AI in conducting forensic analysis of technical data sources, image files and evidence, there are limited advances in this area – Jarrett & Choo (2021) found that in response to the issues facing the DF landscape, there has been a recent move to include AI powered technology within the analysis that is conducted on Digital Forensic sources, a trend which has continued with more and more AI and ML based solutions being developed to solve a range of Digital Forensic solutions, freeing up investigative time and attempting to tackle the highlighted issues of time and volume (Shahzad *et al.*, 2022).

This shift has also been seen in the wider policing and law enforcement sector within the UK, with new platforms being developed and implemented at a rapid pace. This can be seen with large scale projects such as Project Nectar within Bedfordshire Police utilising Palantir Foundry's AI platform to provide both law enforcement specific gains in detecting criminality, as well as improving back office efficiency (EMSOU, 2025) and the development of application suites and platforms to improve the ease of access to technology from within policing through specific approaches such as Tackling Online Exploitation (TOEX) (Thacker, NO DATE). This approach is also codified within the 2026 government white paper on Policing, which creates the intention to develop nationalised AI policing capability, as well as improve operational policing through the use of novel technologies (Home Office, 2026).

Specific uses of ML in DF include in relation to the analysis of images to achieve specific objectives, such as the age analysis in the case of child abuse images (Qadir and Noor, 2021) and the analysis of images to detect nudity (Platzer, Stuetz and Lindorfer, 2014). There have also been studies which have sought to use ML to locate evidence from within an imaged systems filesystem (Mohammad and Alqahtani, 2019), and to locate malware and ransomware from within volatile memory and virtual machines (Cohen and Nissim, 2018) – some specific uses for ML have been reviewed and summarised in *Appendix 3*.

Despite the uses of ML being showcased, it is repeatedly stated that the uptake of ML in the field has been slow and that progress is often impeded (Jarrett and Choo,

2021), with much of the automation allowed being for menial data based tasks, leaving human review to undertake more complex and time-consuming measures (Al Fahdi *et al.*, 2016).

3.4 Existing Works and Tools

Following the above considerations, a review was conducted of existing tools within the DF domain which make use of ML or AI-based automation in their function, with special consideration given to tools which undertake image analysis as part of their function.

These tools are described below;

- SkinSheriff (Platzer, Stuetz and Lindorfer, 2014)
  This tool makes use of Support Vector Machine (SVM) analysis with a specific objective of detecting nudity within provided images.
- Child Abuse Image Database (CAID) (UK Gov, 2024)
  This tool is a centrally held and maintained databased of child abuse image HASH functions, which has been in operation since 2014. This database holds the MD5 functions of known Child Sexual Abuse Material (CSAM) and allows for Forces to upload hash values of located and graded images as well as comparative searching of hash values to locate images of offence.
  Whilst this is not an automated solution, it shows how technological solutions are currently being implemented.
  There are similar other databases in use worldwide, such as the S21 Global Alliance Database (GAD) (Semantics 21, 2025)*, however, CAID is the most commonly accessed amongst UK law enforcement agencies and therefore this is the entry that has been included here.*
- Magnet GriffEye (GriffEye, 2020)
  The GriffEye tool by Magnet Forensics makes use of AI and ML systems such as Brain and Thorn.AI to conduct the processing and analysis of large image and video collections. The tool makes use of AI classifiers to detect CSAM and other defined objects automatically.
- Magnet Axiom (Kim *et al.*, 2025)
  Axiom is described as a digital forensic examination platform which utilises the

Magnet.AI engine to categorise messages that have been obtained from a digital evidence source.

Although the tool is not used for the analysis of image files, it has been included here to show the breadth of tasks that can now be performed by automated solutions.

- EnCase (OpenText, 2026)

  The EnCase software is one of the most well-known DF solutions in use. The software provides a range of DF functions, including AI-powered image classification for the identification of CSAM as well as other defined features.
- Amped Five / Authenticate (Amped Software, 2026)

  The Amped Authenticate software is one of the most well-known software's in use for image analysis. The software is used commonly for the detection of deepfake imagery that has been generated using Generative Adversarial Network (GAN),
- Semantics 21 (S21)/Laseri (Stella Polaris, 2021)

  S21 is a tool for the mass categorisation of CSAM videos as well as victim identification algorithms.

The above research showcases the breadth of tools that are being developed using automation technologies, and how this is being used within the DF domain.

However, there are also concerns that arise from this use, which will be explored in the following section.

These tools, where available, are prescriptively analysed in *Chapter 6.3.3.2*.

3.5 Issues Faced by Automation Implementation into Digital Forensics

One of the primary issues that has presented in relation to the use of ML in Digital Forensic examinations has been the lack of a reliable dataset for use in the training and adaption of these algorithmic techniques. As shown in the work of Bhatt and Rughani (2017), when ML, in this case IBM's DeepQA, is applied to Digital Forensic processes, alternative datasets must be used for the training process, and these are not often available nor suitable for the training that is required to achievable workable accuracy levels.

The issue of lacking datasets is also supported by Qadir and Noor (2021), they show that there has been success in the uses of ML in DF for specific tasks such as image analysis and text identification, however, the lack of available and workable datasets in relation to the training of the algorithms has posed a problem.

It has been found that this can be a continuous issue, *Chapter 3.3* to this thesis shows that one of the most successful implementations of ML into DF is in the analysis of child abuse images. This is a sensitive topic, and the datasets that would be required to train the algorithms are often very highly controlled in terms of access, which would make the training of a specific dataset for a new tool difficult.

Another issue is also that historically it has been very difficult to explain to a human how any AI based system has come to a conclusion as a result of the analysis that it has undertaken (Hall, Sakzad and Choo, 2022), demonstrated in Figure 20:

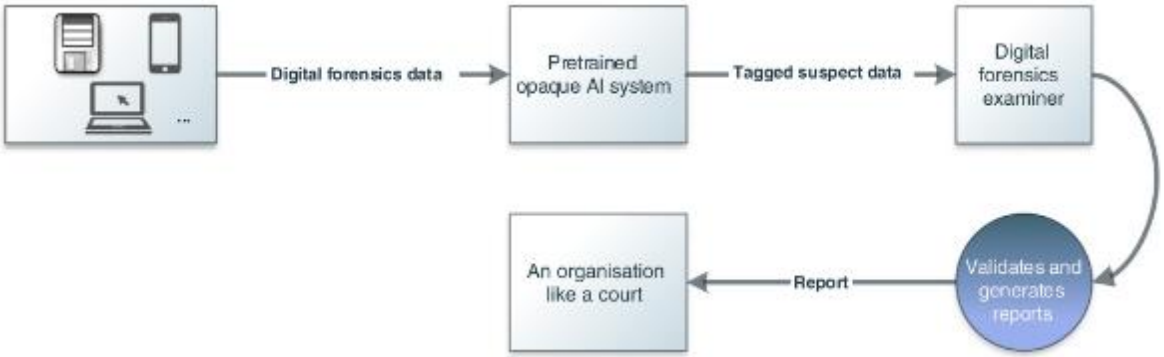


Figure 20 - Current AI-Based Digital Forensic System (Hall, Sakzad and Choo, 2022)

There are also known industry problems in the deployment of ML based systems. Despite significant advances in AI and ML technology, industry continues to battle problems in deploying such solutions in a real-world environment, with even single model deployments taking between eight and ninety days (or longer) to complete effectively. The reasons for this are described as myriad, with issues ranging from high costs, a lack of staff expertise and even data issues with bias (Paleyes, Urma and Lawrence, 2022).

Additionally, there are concerns within the DF domain regarding the function of AI models and how they function. (Khalid, Iqbal and Fung, 2024) point out that AI is often seen as a 'black box' with the function of the system often being unknown even to the

developer. This is additionally discussed within *Chapter 5.5* under Transparency. However, it must be noted that this is a significant concern due to potential conflicts with legislative procedures and requirements such as the Criminal Procedure and Investigations Act 1996 (CPIA) and disclosure legislation, should the tool be used within a court or criminal investigation setting. This is due to the fact that there is a considerable lack of transparency. Courts often will not be given access to, nor information about, the inner workings of AI technologies, nor the tools that make use of these technologies (Grossman and Grimm, 2025). Given the nature of the work that courts undertake, and the requirements under various current pieces of disclosure legislation, such as the CPIA, this is considered a definitive blocker to the use of automation technologies in a meaningful manner.

One important area of development, however, for this type of consideration is the implementation of XAI. This thesis seeks to explain, using multiple interlinked reporting systems, how ML Algorithms achieved their decisions and results (Hall, Sakzad and Choo, 2022). Such systems are also key in avoiding wide-ranging issues such as technical data and results being misinterpreted or misunderstood which can lead to significant legal issues (Horsman, 2022). (Khalid, Iqbal and Fung, 2024) also conducted work in this area, proposing an XAI framework predicated upon transparency and verification. This thesis seeks to address similar problems that are facing the DF domain to those explored by Khalid, Iqbal and Fung, emphasising the importance of these problems.

Finally for this section, practitioners remain concerned about issues being presented in the future if the governance and control of automation technologies is not maintained. There are currently multiple miscarriages of justice that have been identified, and which have been through the UK court system, which have hinged upon previous over reliance upon fingerprint evidence, when there was insufficient challenge and verification mechanisms in place to ensure that the then-novel technology was being used correctly (Campbell, 2011).

Given that the cases involved were from the 1990's when the technology was new, practitioners are rightly concerned about the parallels that could be drawn in respect

of the use of AI technology and so there is a remaining reticence to make full use of this technology until a proper system of governance is established.

A primary concern within the domain, however, would be the requirements for accreditation under ISO standards and UKAS. These are discussed further below in *Chapter 3.6*.

3.6 Existing Digital Forensic and Investigative Frameworks

After identifying that the field of DF is moving towards the adoption of AI and ML in respect to the analysis of digital evidence sources, a review was undertaken of the existing frameworks and investigative directives that are in place for such analysis.

A review of existing research did not find any significant material regarding an overall digital forensic framework for applicability to the court system or reliance on a tool without the need for a *human-in-the-loop* component, beyond the known guidance that is provided in the Association of Chief Police Officer (ACPO) Principles of Digital Evidence and associated specific guidance to specific Police Forces.

Therefore, for the purpose of this thesis, the following relevant documents were identified as being the main investigative guidelines for Law Enforcement, as well as the main national-level guidance in relation to the use of Cyber-related evidence and digital evidence (Table 8):

Table 8 - Existing Investigative Frameworks

Title	**Link to ML?**	**Citation**
Core Investigative Doctrine (2012)	No	(ACPO, 2012)
Murder Investigation Manual (2006)	No	(ACPO, 2006)
MCCT-2 (2015)	No	(College of Policing, 2015)
Major Crime Investigation Manual (2021)	No	(NPCC Homicide Working Group, 2021)
Digital Forensic Science Strategy (2020)	Yes – but as a *future* potential.	(Transforming Forensics *et al.*, 2020)

Electronic Devices Code of Practice (2022)	No	(Home Office, 2022)
Responsible AI Checklist	Yes	(NPCC, 2025b)
NPCC AI Strategy	Yes	(NPCC, 2024b)
Covenant for Using AI In Policing	Yes	(Science & Technology in Policing, 2023)
CPS and Police Joint Operating Procedure (JOP) for AI	Yes	(CPS, 2025)
NCSP Artificial Intelligence LLM Cyber Standards	Yes	(NPCC, 2024a)
ISO 17025:2017	No	(Forensic Science Regulator, 2025)
ISO 17020:2012	No	
ISO 42001:2023	Yes	
ISO 23894:2023	Yes	
AI Playbook for Policing	Yes	(NPCC, 2025a)
ICO AI Strategy	Yes	(Information Commissioner's Office, 2025)
ISO42001	Yes	(*ISO/IEC 42001:2023*, 2023)
EU AI Act 2024	Yes	(EU, 2024)

What has become apparent in the research that has been undertaken is that there is no definitive mention in any of these frameworks about the removal of *human-in-the-loop* from automated tools, and nor is there any specific guidance, or framework by which automated tools can be relied upon. Conversely, the EU AI act specifically requires human oversight measures, although it does not go as far as to define exactly by what means human oversight could be achieved.

The particular EU AI Act, as the world's first legally binding statute placing requirements onto organisations and member states in relation to the use of AI provides support for the approach that has been taken in this research. The framework that has been developed and proposed within *Chapter 5.4* provides a method of removing specific *human-in-the-loop* processes to gain benefits, whilst maintaining oversight in relation to the AI itself, and the tools that use it.

A key flaw in this research that must be pointed out however, is that there may be Force-specific guidance that is available within the Police Service, which is not readily available to the public. However, this has clearly not led to national level guidance and has not impacted on any form of work that would lead to the Sector's overall acceptance or use of *'True Automation'* in DF.

An additional consideration for the framework that is being produced as part of this thesis is the existing regulations and requirements that are placed upon forensic tools and processes in general.

The principal tenants of digital forensics can be found within the 4 ACPO principles from 2012, which have been mentioned in the above Table 8.
These principles require 4 elements in relation to the obtaining of digital evidence, and the analysis of that evidence. These 4 tenants are:

1. No action should change data that may later be relied upon in court
2. If original data must be accessed, the person doing so must be competent and able to explain their actions
3. An audit trail or record of all actions must be created and preserved
4. The person in charge of the investigation has overall responsibility for ensuring compliance

These tenants must be adhered to ensure that the collection of evidence is both reliable and acceptable. These tenants form a core component of the produced framework and are reflected in the gates that are provided as shown in Figure 40.

It should be noted within this section that there is increasing presence of the Forensic Science Regulator (FSR) in relation to the use of AI within UK policing. This

is especially apparent when considering the specific task of image and video analysis with DF as a use case for AI.

Although the FSR is a neutral entity and does not have a defined position on AI in its current form, there has been specific mention of the use of AI and automation technologies within deepfake investigations and manipulation/editing of video files for court purposes (Forensic Science Regulator, 2025). The requirements for this are that the methods that are used are defensible and validated, the same as requirements for other forensic science aspects. This supports the approach within this thesis, for proposing a system for validating and verifying a tool before use.

The biggest consideration that must be raised in respect of this section of the thesis, is that the current existing procedures and documentation around the use of automation technologies within DF explicitly embed requirements around governance and validation which are presently achieved through the use of *human-in-the-loop*.

The removal of this assurance and validation would appear on the face of it to put '*True Automation'* into conflict with this thesis and with accepted standards. However, that is the purpose of the *Section 5.4* and *Figure 40* Framework. The approach within this thesis seeks to achieve a satisfactory level of validation and verification without requiring a human input.

In to the above, and pointed out in *Chapter 3.5*, accreditation against the ISO standards via United Kingdom Accreditation Service (UKAS) is also required for the framework to be applicable to the DF domain.

This accreditation is required under the FSR's Code of Practice (Forensic Science Regulator, 2025) which in turn is a statutory requirement for UK Police DF laboratories under the Forensic Science Regulator Act 2021 (UK Gov, 2021). There are several ISO standards which may be applicable, however, the most appropriate to the forensic analysis of digital evidence sources would be ISO17025:2017. The purpose of this accreditation is to provide independent assurance of the competence and impartiality of a forensic unit and considers a wide variety of aspects, including the method and validity of methods, which would be the most

crucial in this case. It is also a key requirement under the FSR code of practice, and therefore would have a strong impact on the usability of the produced framework in this thesis, as well as the use of tools in *True Automation* sense.

The result of this accreditation requirement would be that a forensic unit employing the framework as an assessment method would also have to demonstrate that this framework was fit for purpose and appropriate, as well as where it is limited and as the point of *True Automation* and the framework assessment is to remove the human in the loop component, the unit must satisfy multiple aspects including ongoing quality assurance.
It is believed the framework, and the approach taken to that framework are met within this thesis, for example as shown in *Chapter 6.3.1.4* ongoing assessment is an expected requirement of the tool which is made clear within the documentation that has been prepared to accompany the framework (*Chapter 5.5*). However, a key limitation of this thesis is that this needs further consideration, this is discussed further in *Chapter 8.3.2*.

3.7 Legal and Ethical Considerations of *True Automation* Within Digital Forensics

In addition to the above framework and requirements that are placed upon any meaningful consideration of using automation technologies in the manner prescribed under the proposed '*True Automation*' definition, there are several legal and ethical considerations which must be considered.

The first consideration which must be met would be the right, under existing data protection law, for an individual to 'opt-out' of any processing which would result in a decision being made which results in an adverse legal effect (ICO, 2025). This naturally means that suspects may choose to opt-out of decisions being made by AI, which would include the framework produced in this thesis. There are means to deal with this requirement, as an individual has one month to enact this opt out, and they must be informed of this processing. There is an existing information sheet which is given to both suspects and victims whose devices are obtained and examined as part of a criminal investigation, and it is considered that the individuals could be informed of this processing change by a slight amendment to this thesis.

It is additionally considered that the potential gains that would be obtained from the automated analysis of the bulk of devices, would offset any extra time that would be needed to examine devices of those who opt out.
Additionally, it is considered that although this must be actively addressed as described, the framework in this case is not actively making decisions and therefore may be exempt from this, although this would require legal testing.
As a result, although this is a consideration, it is not seen to be fatal to the '*True Automation'* approach that is proposed in this thesis.

In addition to the above legal consideration, there is also a consideration around CPIA and disclosure, however, this has been addressed previously.

There are also ethical aspects that must be considered in addition to the legal considerations that have already been explored.

One of the most significant of these is the impact of bias. It can be tempting to assume that a machine cannot be biased towards any particular element or group, however, this has been found not be the case. ML and AI models are susceptible to cultural bias through the data that is used in the training (Mak and Luo, 2025). When this is applied to a policing context, the effects of this can be significant, and even more so when considering using such a technology without the need for a human to review the results.
Although this concern is significant, and there is certainly opportunity for bias to become a factor in an automated technology, it is considered that there are significant safeguards in place within the Framework around accuracy, reliability, and reporting which would enable this to be picked up without human intervention in the specific automated process. Additionally, this issue could be mitigated, if not removed, by ensuring appropriate data cleaning and review to minimise bias input into a chosen MLA.

The addressing of this issue is beyond the scope of this thesis however, this is something that the developers of models must be aware of during the course of development and model training.

Another issue presented, and one that has appeared within the primary research that has been conducted in this thesis, is one of accountability.
The question has been raised as to who is actually responsible for the decisions made, and the analysis conducted, when a ML or AI based system has performed an action without a human review. The argument arises that an officer conducting the analysis of a digital evidence source, may not be responsible for the findings of any automated process that they did not review, and therefore they would not be the one to present these findings to a court. This is an issue which has not yet been resolved in practical law enforcement, primarily due to the *human-in-the-loop* requirements ensuring that there is a responsible human officer.
This is a significant issue around the validity of evidence produced, and the chain of evidence when analysis is completed and is something that will need to be addressed organisationally in future work. It is considered that this is out of scope of this thesis as this is an organisation and legal matter that will have to be addressed and considered as technology advances, rather than it being within the remit of an assessment framework.

3.7.1 Responsibility, Trustworthiness, Acceptance and Sustained Accuracy

The considerations and findings in the above chapters have identified four connected requirements that must be clearly addressed if an automated tool is to be relied upon in terms set out in *True Automation*.
These four aspects are: responsibility, trustworthiness, acceptance and sustained accuracy.

3.7.1.1 Responsibility

Within this thesis, it must be clear that the removal of the need for human review does not absolve any involved party of responsibility for the use and deployment of the tools.

Instead, if *True Automation* were to be adopted, then the responsibility of the human component to the process would shift from verification of individual output to overall governance, compliance and deployment.

The main consideration in this being that deployment, use and assessment of tools by the produced framework in this thesis does not transfer legal or professional responsibility onto the automated tool itself.

There are several clear responsibilities that remain within the human purview – those being:

- Developers
  The developer of tools and models would be responsible for the providing of sufficient technical information about the tools produced, their intended function, model training, test results and any identified limitations.
  This would allow for the sufficient analysis and verification of the tool by forensic practitioner assessors, as well as allowing for the transparency required by the judiciary.
- Assessors
  The assessors of tools i.e. the main users of the produced framework in this thesis, remain responsible for the proper assessment of tools per the produced framework, and determining whether the provided information allows for the compliant assessment of those tools.
  They are also responsible for the recording of the assessment and results.
- Organisational
  The overall assessing and user organisation retains responsibility for the deployment decisions, local governance, staff training and competence, and incident management involved in the deployment of the tools assessed.
- Practitioners
  The practitioners using approved tools within their defined function remain responsible for appropriate deployment and use of the tool within the bounds and operating procedures set by their organisation and within the intended use and function of the developer.

3.7.1.2 Trustworthiness

An important distinction for the purposes of this thesis is that the trustworthiness of a tool does not imply or present an implicit presumption that the tool is fully correct.

The term is used to describe a demonstrable basis for reliance on the tool without human review of individual output.

There are several elements of the produced framework with would give this basis, these are the gates that are presented. Overall compliance with the framework then gives this trustworthiness.

This concept should also be kept distinct from elements such as accuracy. For example, an accurate tool may still be unsuitable for reliance in the DF domain if other requirements, such as explainability and transparency, are either missing or do not pass scrutiny.

As such, trustworthiness arises from a combination of technical, procedural and legal requirements and scrutiny, supported by independently verifiable evidence.
It is these elements that the produced framework, and this thesis, seeks to provide.

3.7.1.3 Acceptance

Within this thesis, the term acceptance is used to refer to the willingness of users, both individual and organisational, to make use of tools within the terms of *True Automation*.

The concept of acceptance is similar to trustworthiness as discussed above, however should be considered distinctly as a tool that is considered trustworthy under the bounds of the framework and assessment, may not be accepted by users, for example if they do not understand the function.

The proposed framework, and the surrounding interviews and surveys, is intended to ascertain use acceptance of *True Automation* as well as provide a mechanism by which users can assure themselves of the tool's function and performance and provide them with a structured approach to their own considerations.
The passing of the framework by a tool may not necessarily lead to acceptance, however, it provides organisations with a method for beginning the organisational task of gaining user acceptance for approved tools.

3.7.1.3.1 Overall Acceptance

In addition to the above Acceptance point, it must also be considered that in order for the framework, and the concept of *True Automation,* to be fully accepted and utilised by the DF domain, broader stakeholder engagement and acceptance is necessary.

Acceptance itself cannot be viewed as a technical-only problem, the legal, organisational and user aspects must also be part of the consideration. The literature that has been reviewed within this chapter support this, stating that barriers to adoption within the DF domain frequently arise because stakeholders lack confidence that automated systems are lawful, understandable and appropriate for operational deployment (Hall, Sakzad and Choo, 2022; Khalid, Iqbal and Fung, 2024; Grossman and Grimm, 2025).

This speaks to a limitation of the current thesis, which is discussed throughout *Chapter 8*.

3.7.1.4 Sustained Accuracy

A tool's accuracy, as well as other performance, may not be set and indefinite. Changes to the tool, updates to models and changes in functionality can have significant impact upon the performance metrics used in assessment.

As such, the term sustained accuracy is used to reference the requirement for ongoing levels of performance, rather than a one-time stationary claim.

A tool that is used within the bounds of *True Accuracy* must be able to demonstrate ongoing levels of performance throughout the period of operational use, rather than at the time of initial assessment alone. There should also be a system of documenting changes to the tools that are relied upon, to continue showing the required levels of transparency as well as function.

This is built into the framework documentation showing that continuous and repeated assessment of the tools approved should be conducted. The framework is not designed to be a single-use system. Additionally, if the scope of a tool changes, then the assessment should begin again as differing scope will necessarily provide different acceptance levels for performance and may change the function of a tool.

3.8 Gap Analysis

As shown in the previous chapters to this thesis, there has been significant work conducted towards the development of tools for the conducting of Digital Forensic analysis – although these are still lacking (Luciano *et al.*, 2018).

However, from both a review of the existing works in this area of study, and from the personal experience of the researcher, there is an apparent and identifiable research gap. That being that there is no unified and easy to access assessment system for the analysis and testing of automated DF tools, especially within the bounds that are defined in *True Automation.* Without this, the domain continues to rely on human intervention and review which constrains and limits the benefits that could be obtained by the use of appropriate and approved automated technologies.

Both the DF domain and the developers for tools within that domain, have focussed primarily on the introduction of automated tools, or the implementation of automated technologies into their tools, rather than looking at how these could be adopted in such a manner as required to solve the issues facing DF and how governance requirements can be adapted to allow for the potential gains to actually be obtained.

Whilst tool development work is important, limitations in the explainability and interpretability of AI-based forensic tools may impede trust and understanding among CJ practitioners (Khalid, Iqbal and Fung, 2024). Although a tool may be able to conduct analysis on an evidential source, a human is still required in most cases to conduct a manual review of the item before the resulting analysis is acceptable by the standards of the courts (Transforming Forensics *et al.*, 2020; Hall, Sakzad and Choo, 2022). There is no work that has been found in this research which would show development towards a universally accepted method by which the results of the tools that are being developed could be used in a court or criminal justice setting, in fact, the latest guidance on this point from the heads of UK Law Enforcement is only now recommending that the implementation of ML and AI into DF should be addressed, as a collaboration between academia and the industry. In this thesis, there is still no consideration as to how the Court System would react to such an analysis (Transforming Forensics *et al.*, 2020).

There are several Digital Forensic frameworks which have been suggested, covering a range of areas and scopes (Elsaesser and Tanner, 2001; Ribaux and Margot, 2003; Hasan, Mahmood and Hasan, 2011; Yeow, Mahmud and Raj, 2014; Mohammad and Alqahtani, 2019; Shayau *et al.*, 2019), however, all of these systems are for the dealing of incidents and tasks. They do not consider how this will be accepted or dealt with by the judiciary.

Outside of the judiciary, there have been some efforts made to create frameworks for the implementation of AI into forensic science (Stacey *et al.*, 2025), however, the focus of these others works has been on the overall use of AI rather than the applicability of the use of AI to the court system.

Additionally, previous identified works have looked at the use of AI, and how this can be implemented, but have remained steadfast in the assertion that human review and validation is required when looking at such tools, which can erode the gains that are made by using such tools in respect of time saving and efficiency.

This raises the question, when looking at the provided novel term of '*True Automation*' found in Chapter 1.2 of how forensic science can begin moving towards the use of appropriate tools without the need for said intervention.

In addition, this gap shows that there is an additional missing element in the combination of responsibility, demonstrable trustworthiness, stakeholder acceptance and continued maintenance of accuracy within a single unified technical assessment framework. Although the requirements for these within the assessment of DF tools are frequently discussed within literature that has been reviewed, this is often in isolation.

Therefore, the research gap that this thesis aims to address can be stated as:
The lack of a cohesive and unified framework to enable reliance upon 'True Automation' within forensic analysis of digital evidence sources by the Criminal Justice System.

This is an important gap that must be addressed as the conducted research has also shown that it has also been shown for a significant period of time that the issues being faced are not the responsibility of any singular agency, and that only a multi-

agency approach to the problem will suffice (Dardick and Lau, 2005). Although this point was made over 17 years ago, this same point is borne out in recent developments, with recent research showing that there is a move to include both the Computing and DF communities in work designed to address this problem (Bhatt and Rughani, 2017; Mohammad and Alqahtani, 2019).

Additionally, it was shown by Jones (2009) that this was still a developing problem, with a finding that between 1986 and 2009, DF processes for the assessment of technological evidence sources had not kept pace with development and that forensic science standards were still not being applied to DF by either the forensic community or the judiciary (Jones, 2009). Despite the fact that this research is almost fifteen years old, this still has not been addressed by the Digital Forensic community.

From a real-world perspective, there is a significant amount of time that is spent conducting analysis which could be done by an automated system (Al Fahdi *et al.*, 2016; Transforming Forensics *et al.*, 2020) – although it is yet to be seen whether the technology in its current state would meet with the requirements of any such framework.

A key issue that has presented in all of the research that has been reviewed as part of the Literature Review here is that when work is conducted as part of the development of new tools for the area, they are created and developed in isolation (Patil *et al.*, 2022), with the authors not appearing to give any clear credence or consideration to the legal profession and the legal side of the tools that they are creating, beyond the basic ones such as not altering the evidence that they are dealing with – although this is not a universal consideration.

As far back as Rogers (2004) it has been made clear that the legal side of the Criminal Justice system must be considered when tools are being created. However, as has been shown above, the frameworks that are in place to assist investigators do not make mention of the actual tools that are used, and there is no unified measure in place which would allow for an appropriate legal assessment of technical tools.

It is considered that such a system would allow the Criminal Justice system to consider the tools that are being created and assess as to whether they are acceptable for a court environment via robust testing and demonstration of accuracy and can thus be relied upon at such a level in the same manner that fingerprint and DNA evidence is accepted (Hughes and Karabiyik, 2020). In the absence of such an approach, the tools that are being created, and the new technologies that are beginning to be implemented, will continue to be underutilised and misunderstood by the Criminal Justice sector, and the desperately needed accuracy and time efficiency will continue to be missed.

The creation of a framework to address this gap would also have the secondary advantage of allowing the system more opportunity to understand and utilise the tools that are discussed in this thesis, which may mean that more datasets and information could be considered for release to allow for further training of the algorithms, and thus solve the main issue with implementing ML into the field.

Additionally, the proposed framework looks to address the multi-agency challenges that are identified above by providing a common assessment mechanism that can be constructed with input and buy-in from all the involved CJ stakeholder groups. The framework seeks to ingest multiple requirements sets from across these stakeholder groups and convert these requirements into a structured set of assessment criteria. In doing this, the framework provides a shared approach upon which various stakeholders can evaluate the suitability of tools that are produced and intended for *True Automation* use and addresses the fragmented approach that has been highlighted in the above chapters.

Finally, the produced framework would seek to address the combination of responsibility, trustworthiness, acceptance and sustained accuracy by converting these requirements into assessable criteria and treating approval as conditional upon the defined scope, evidence and continuing review necessitated by the proposed framework.

Having concluded this from the research conducted, it was finally necessary to narrow the proposed framework down to a specific area of forensics to ensure that it was as applicable and usable as humanly possible. The above considerations in

Chapter 1.3 have led the researcher in this case to conclude that the area of image forensics is the most viable for this research. The rationale for this being that it is an area of significant automation technological development, which has been shown in *Chapter 3*, and it was also mentioned within the *Chapter 1.3* focus group as a significant area of interest. These factors combine to mean that the opportunities for ML and AI involvement in analysis are significant and is therefore a viable area for the development for a framework designed to ensure that the results of said analysis, especially in cases where the human element may not be able to manually conduct the analysis, are legally and technologically valid.

3.9 Chapter Summary

The chapter has presented a comprehensive and structured literature review, which has demonstrated that the DF domain continues to face significant challenges relating to workload, evidential volume, and reliability.

These challenges have driven increasing interest in utilising AI, ML and automation technologies within DF processes and tools.

The review also identified that there are significant issues which remain regarding the use of automation technology within DF, such as legal challenges, ethical considerations and governance requirements.

Finally, the review has established the need for four explicit requirements for the proposed framework – responsibility, trustworthiness, acceptance and sustained accuracy.

Specifically, responsibility must remain with the human component, although that responsibility shifts from direct verification to governance and control, trustworthiness must be demonstrated through a combined technical and operational approach, acceptance must be conditional and evidence-based, and finally that accuracy must be sustained and monitored, rather than being snapshotted at a single point in time.

The chapter has further identified a clear research gap: the absence of cohesive framework by which AI and ML-based DF tools can be assessed for reliance within CJ without the need for a human review, as the human review requirement hinders development and has a detrimental, or at times preventative, effect on the efficiency gains coming with the use of such technology.

This finding provides justification for the thesis as a whole and establishes the conceptual basis for the framework developed in *Chapter 5*. The role of that framework being to provide a structured and reviewable basis for determining whether a tool can be adopted in the *True Automation* sense.

4.0 The Real-World Practical State of Digital Forensic-based Automation within Criminal Justice in the UK

To fully investigate the gap stated within *Chapter 3.7*, the current real-world state of automation and its use in live investigations, as well as its place in the investigative options open to law enforcement, needed to be established.

This has been addressed, in part, in *Chapter 3* in which various tools and use of automation are covered. This shows the reason for the development of such tools and techniques and showcases how this is being put into use from a developmental perspective.

On analysis of the data that has been obtained within *Chapter 3* it can be inferred that, with the range of tools that are available, the developing interest and consideration towards the use of automation within the field of DF is a global one, and hence one that affects far more jurisdictions than the UK alone. However, for the purposes of this research the focus shall remain on the UK, the rationale for this being that the project is based in the UK and that the primary research that is being conducted requires access and contact with law enforcement organisations in order to gain an accurate picture of the use and acceptance of AI and ML.
Additionally, with multiple jurisdictions, there will be multiple different legislative and procedural requirements placed upon tools and any produced framework.

This focus upon a singular, accessible, legislative area increases the validity and applicability of the research conducted in this thesis.

The previous chapters to this thesis have provided an overview of the current use cases, and uses, of AI and ML within the DF field, however, for the purposes of this thesis further work was required in order to ascertain the real-world state of automation within the DF domain in the UK's criminal justice system.

4.1 Organisational State of Automation

The first step was to obtain information about the adoption of such technologies from an organisational perspective. To achieve this, Freedom of Information requests were sent to all 48 Police Forces across the UK, asking questions about the adoption and use of AI and ML based automation within DF.

The exact wording of the requests, as well as a rationale for this wording, is explained and explored within *Chapter 2.4*.

Once obtained, the results were collated within the SPSS program, as described within *Chapter 2.4*.

The collated findings of the requests are shown below in Figure 21:

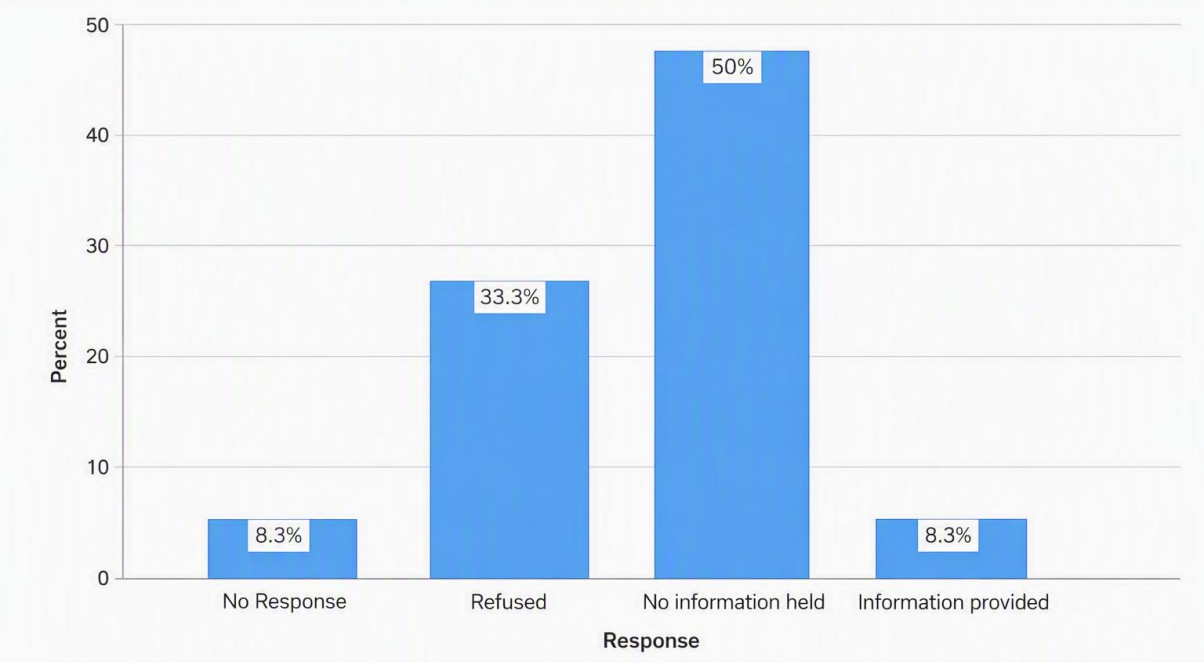


Figure 21 - Collated FOI Responses

As can be seen above, 50% of Police Forces across the UK (inclusive of England, Wales, Scotland and Northern Ireland) responded that they had no information relating to the use of AI and ML within their DF policies and departments. An additional 33.3% refused to provide information.

The results of this codification, summarising the refusal reasons are as follows in Figure 22:

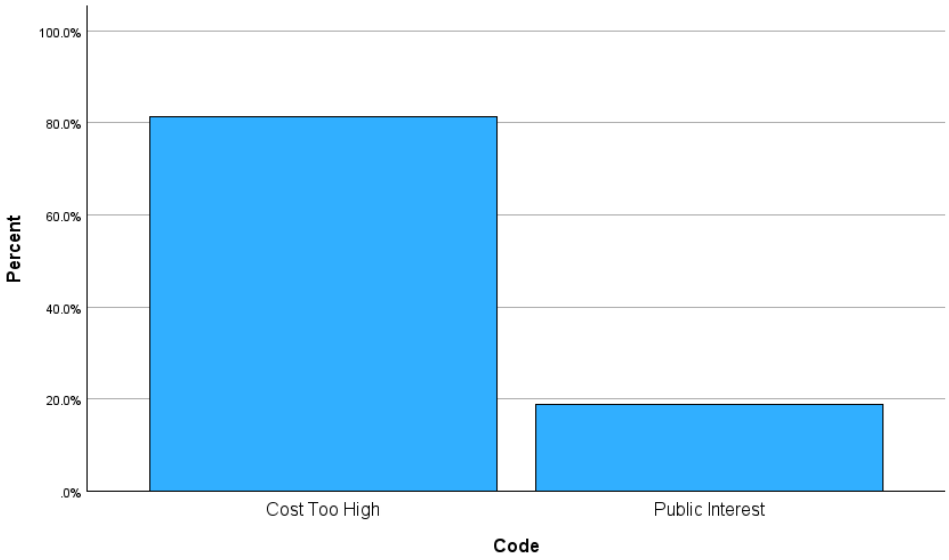


Figure 22 - Refusal Reason Breakdown

For the purposes of this analysis, results with the *Not Applicable (0)* code were excluded.

As can be seen from the above, 80% of refusals were based on the organisation reporting that finding the data, or lack thereof, would be too costly and breach the £450 limit on FoI requests under UK Law.

Following the codification of these results, an analysis was conducted to interpret these.
From the findings above, half of UK Police Forces actively state that they do not have any policies or procedures – nor do they have any research or development – ongoing in the realm of AI, ML or automation of DF analysis of devices (Figure 20). Over 33% refused to provide information, and 80% (13) of these were because the information was not readily accessible.

UK Freedom of Information Law requires an organisation to commit a maximum of £450, which translates to 18 hours of a full-time staff member at the given rate of £25 per hour (UK Gov, 2005). This being given as a rationale for refusal is demonstrated in the below Figure 23 showing a response from a given Police Force:

Reason for Decision:
Section 12(2) states that information need not be disclosed if the cost of locating/retrieving that information exceeds the 'appropriate level' of £450, or 18 hours based on a cost of £25 per hour.
Section 8(1)(c) Not a valid Freedom of Information Request.

Figure 23 - Cost Exclusion Response

Given the above difficulty in locating and retrieving the named data, which is a general policy document request, it is a reasonable inference that, if the information about automation is so obscure that it would not be accessible within such a timeframe, that it is unlikely to be a commonly used, referred to or required document. The natural consequence of this inference being that the use of automation in terms of AI or ML, is not widely used within these organisations in the same way as the organisations that have stated that they do not hold any information.

It is further noted from the above requests that there was no single organisation that responded stating openly that they make use of AI and ML, or automation, in their DF processes. Only 8% responded with information, and these referred to tools which have already been described above in *Chapter 3.4*.

The result of this analysis shows that there is a perceived organisational hostility, or reluctance, towards the use of AI and ML-based automation within the DF domain, coming from Police Forces across the UK. This supports the *NPCC* stance that a human-in-the-loop review is required for any meaningful use of AI or ML within Police processes (NPCC, 2026), as well as the anecdotal knowledge of the researcher from working in the Policing sector, and within the AI Enablement domain.

Additionally, this finding is borne out by recent developments in specific Police Forces, which publicly welcome the use of modern technology but also express significant caution. An example of this is in a recent news article involving an interview with a Cheshire counsellor wherein they say that the use of AI would be a decision for the Chief Constable, but which is something that he would be keen to see implemented (Ryan, 2024) – this coming in 2024 when technologies such as facial recognition have been available for several years, first coming into use in 2016 with the Metropolitan Police deploying it at the Notting Hill Carnival in 2016 (Robertson *et al.*, 2016) shows the historic and continuing reluctance for forces to engage with this technology.

The dichotomy between the Police's public-facing acceptance, and upcoming national policy of adoption and nationalisation of AI as seen in the 2026 UK Government White Paper (Home Office, 2026) and embracing of AI technology,

facing the more reluctant actual implementation of this is borne out by a recent order given out in Cheshire Police in 2024 – publicly, the organisation is taking the stance that AI technology is to be embraced where it is of benefit to the cause of preventing crime, however, internally they show such reticence that they have outright banned it's use for the creation of material destined for the Criminal Justice system.

This supports and complements the inferred findings of the FOI requests that are analysed above, and concurs with the finding of persisting organisational hostility, although there are the beginnings of a move towards acceptance of the technology. This result speaks to the suitability and requirement for the framework that forms the artefact and backbone of this thesis – organisationally and culturally law enforcement remains hostile to the use of such automation technologies, although practitioners are more amenable to the use of such technologies if it were to be approved by the organisation itself, but with these beginning moves to consider accepting it as a law enforcement tool, the constructed framework is an important inclusion to advance this movement by providing a set of structured and relevant criterion against which tools can be relied upon.

However, there are significant signs that this is changing. This has been discussed in *Chapter 3.4*, but it should be noted here that whilst the above discussed attitude remains the Police are changing. Cheshire Police have actively included AI in their most recent digital plan (Cheshire Constabulary, 2025) and have created an AI Enablement Team in which the author works. Through this knowledge, it is known that the use of AI is actively growing and is being deployed in a variety of ways which go beyond the documentation and analysis that have been discussed in this thesis, and the facial recognition technology that has dominated headlines in this area. An example of this is the implementation of a Virtual Officer for attendance at public reports as has been done in Northern Ireland (PSNI, 2025).

4.2 Practitioner State of Automation

The above research has shown the attitude and stance of law enforcement from an organisational perspective, however, in continuing this thesis and showing the full state of automation acceptance within the UK's criminal justice system, information

from the investigators was also required, in order to determine practitioners' opinion and ascertain if the organisational reluctance was shared with those 'on the ground'.

In obtaining this information, a targeted survey was sent to approximately 100 active and serving Police Officers from a Home Office Police Force – these officers were from a number of different units, investigative roles and ranks. The details of this are shown in *Chapter 2.4*.

From the ~100 officers contacted, 42 responses were obtained – representing a response rate of 42%.
The 42 responses were then subject to statistical and thematic analysis to locate any common themes and understand the statistical significance of the results.

The method of initial analysis and codification is explained in *Chapter 2.4*.

Once the survey results had been coded as described, the results were subject to statistical analysis as well as standard analysis. The first step of this was to ascertain the *Standard Deviation (SD)* of the relevant results, as shown below in Figure 24:

Descriptive Statistics

	N	Std. Deviation
Aware of AI or ML?	42	.497
Aware of Tools?	42	.497
Aware of tools used?	42	.505
Any Frameworks?	42	.397
Willing to automate analysis?	42	.354
Willing to use automation if framework was created?	42	.216
Valid N (listwise)	42	

Figure 24 - Standard Deviation of Survey Results

The data was additionally visualised to assist in analysis and interpretation of the data.
The results of this visualisation are shown in the below series of figures:

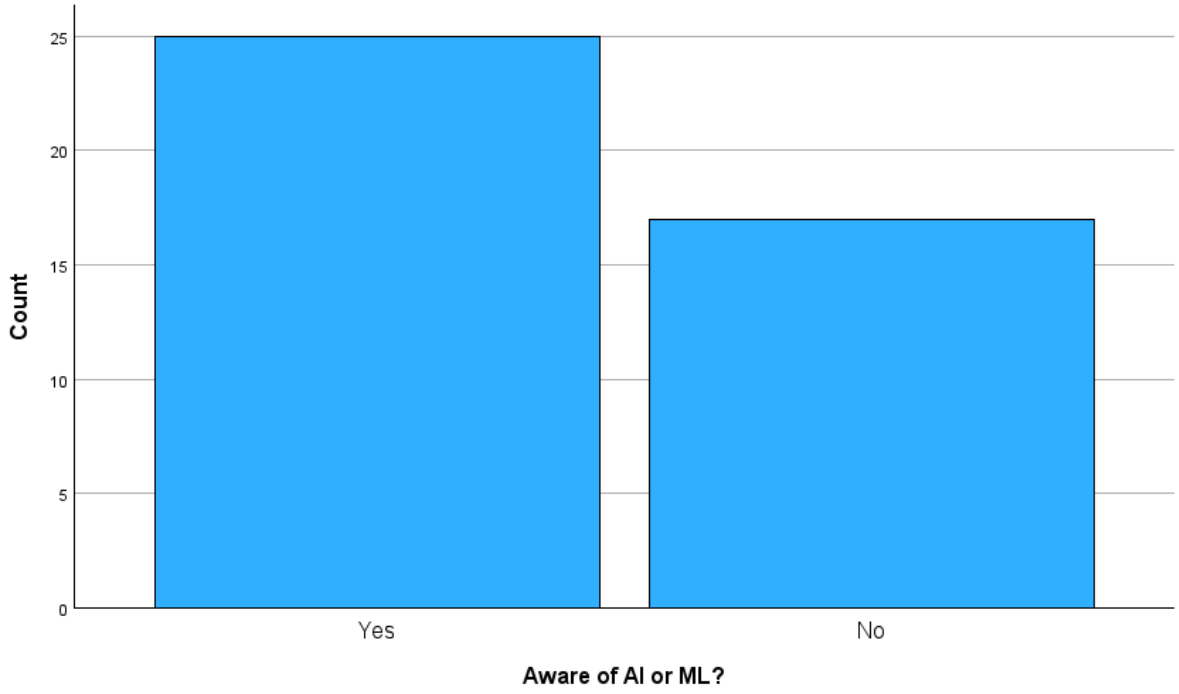


Figure 25 - Visualisation of Survey Awareness Question

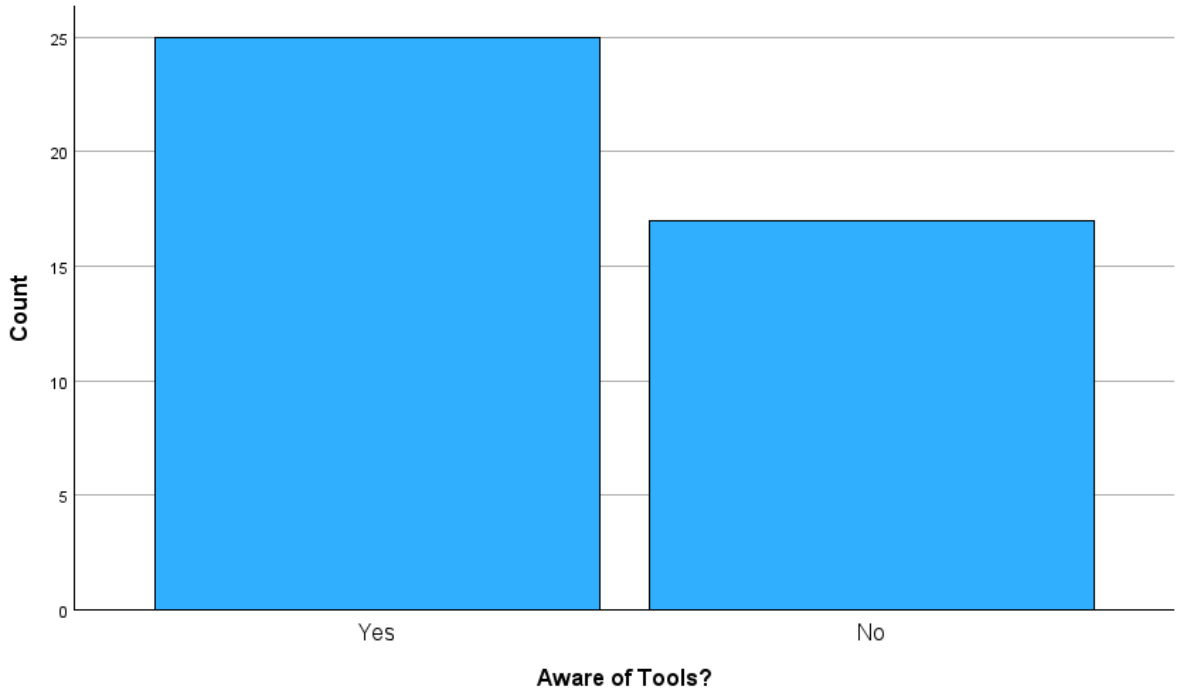


Figure 26 - Visualisation of Survey Tool Awareness Question

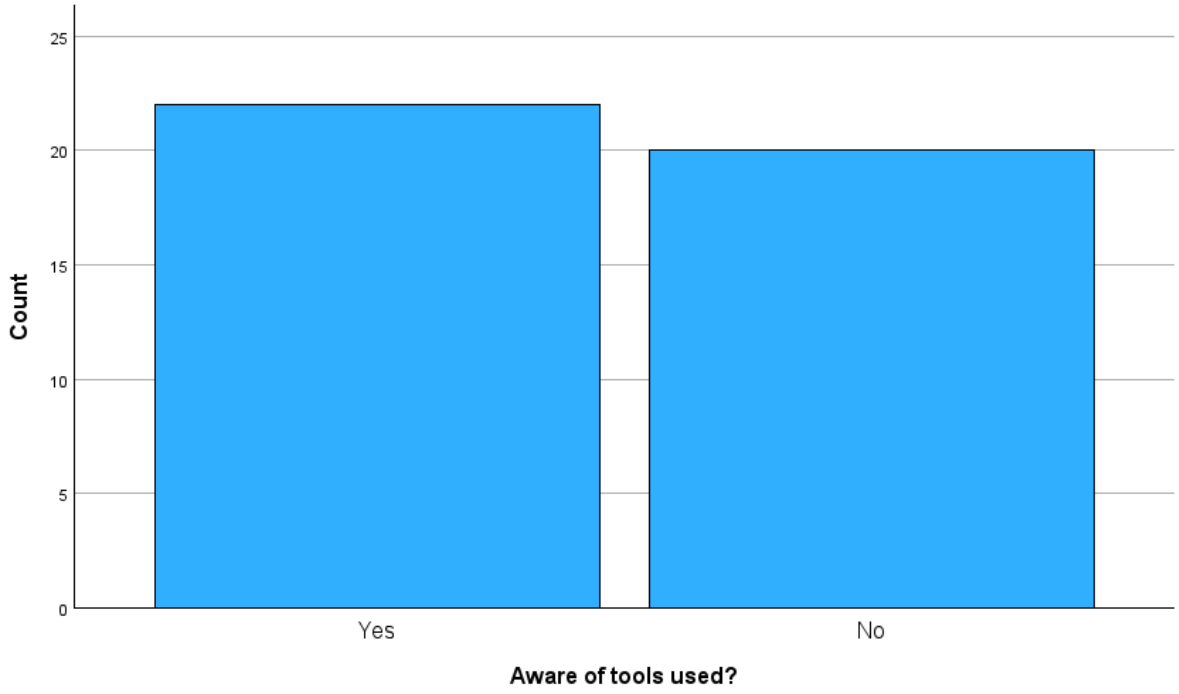


Figure 27 - Visualisation of Survey Tool Use Question

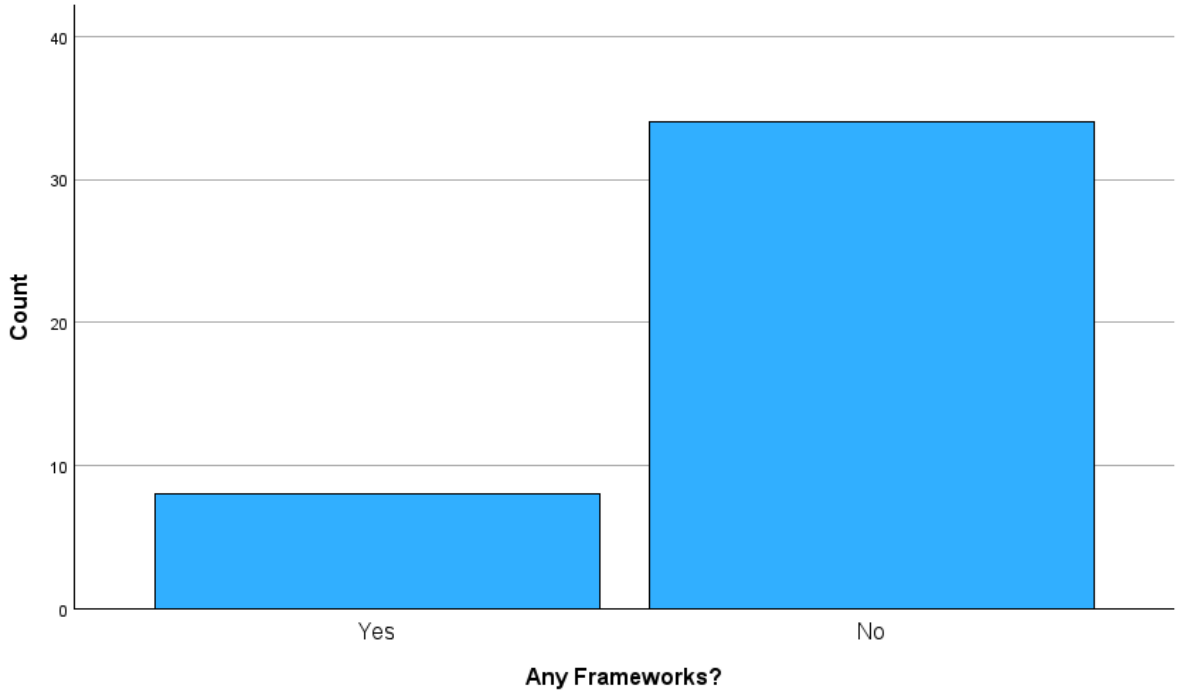


Figure 28 - Visualisation of Survey Frameworks Awareness Question

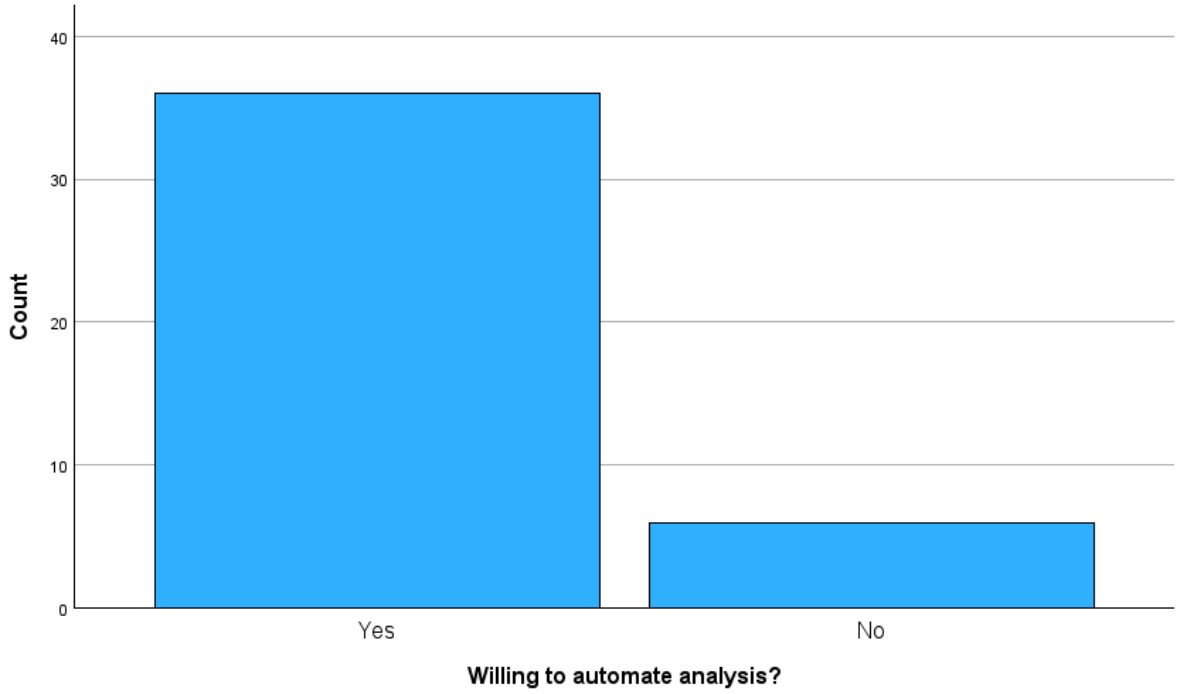


Figure 29 - Visualisation of Survey Automation Willingness Question

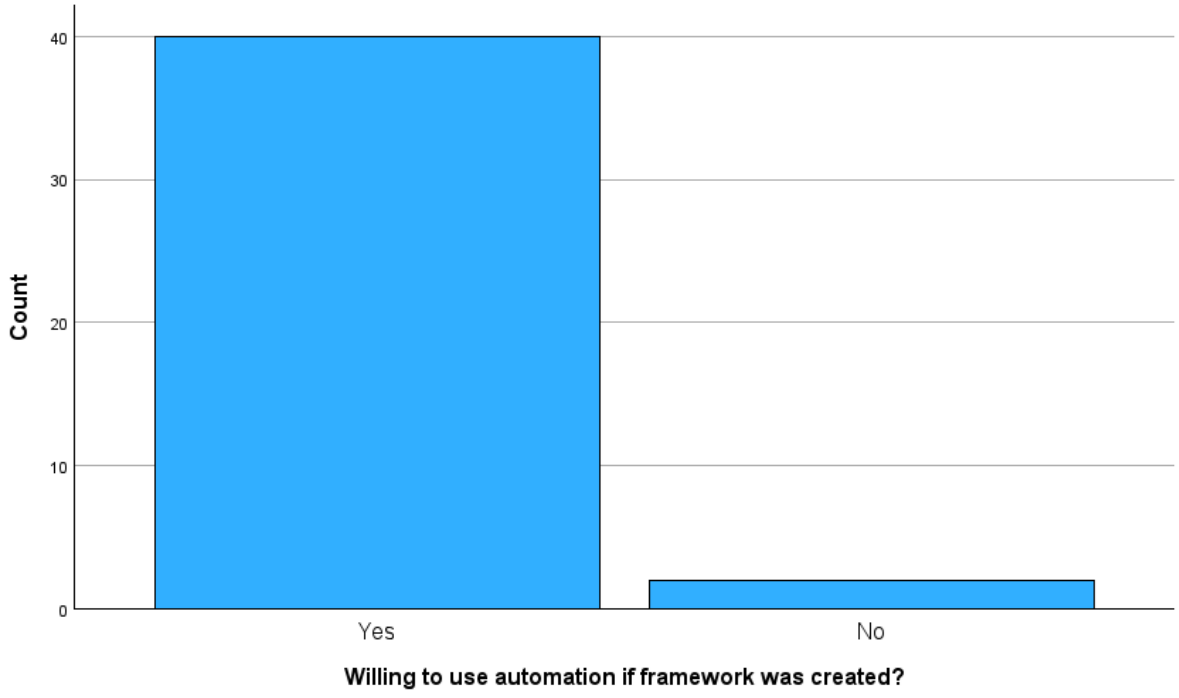


Figure 30 - Visualisation of Survey Willingness Question

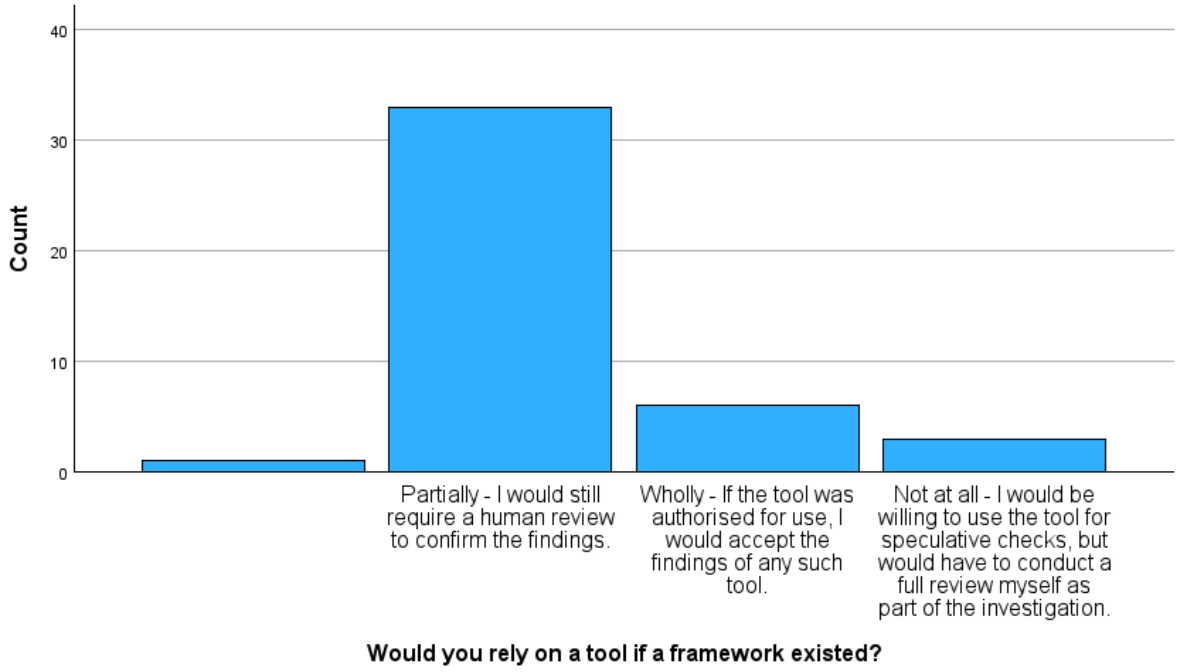


Figure 31 - Visualisation of Survey Reliance Question

When considering the standard deviation results that can be seen in Figure 24 it must be considered that these are in response to Yes/No questions and as a result a high result would indicate an even split of Yes/No responses, where as a low result would indicate a more consistent result – for example the participants responding further into Yes or No, respectively (Martinez and Bartholomew, 2017).
From the results shown, initial interpretation would show that on a per-question basis –

There was a relative split between Yes and No responses in relation to the awareness of automation and awareness of tools questions – the finding of the SD being supported by the results shown in the above figures.
In relation to these particular questions, 58% of participants reported that they were both aware of automation in the form of AI and ML and also were aware of tools that could be used for this purpose – this matching split shows that the participants were not 'randomly' selecting responses and lends credibility to the findings of the survey.
To further ensure that participants had not been 'randomly' selecting answers in relation to these first two pivotal questions, a *Pearson Chi-Square* test was conducted using these 2 variables, the results of which are shown below in Figure 32:

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	42.000[a]	1	<.001		
Continuity Correction[b]	37.952	1	<.001		
Likelihood Ratio	56.691	1	<.001		
Fisher's Exact Test				<.001	<.001
Linear-by-Linear Association	41.000	1	<.001		
N of Valid Cases	42				

Figure 32 - Pearson Chi-Square Test

The purpose of this test is to ascertain whether the data that has been collected is the result of random chance or is non-random which would mean that the data is valid for interpretation (Martinez and Bartholomew, 2017).

The *Pearson Chi-Square Test* was selected for this task as the analysis undertaken examined associations between categorical variables using frequency data. As such, this was the preferred test over other available statistical tests as they would assess continuous data, mean differences or linear relationships.

As can be seen from the above results, the *p-value* shown is <0.001 showing a strong indication that the results are not the result of random chance, and that the findings in these questions are statistically significant – hence the data can be relied upon for subsequent analysis and interpretation.

The specific results breakdown for these 2 questions, which are the results used in the above analysis, is shown below in Figure 33*:*

Aware of AI or ML?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	25	58.1	59.5	59.5
	No	17	39.5	40.5	100.0
	Total	42	97.7	100.0	
Missing	System	1	2.3		
Total		43	100.0		

Aware of Tools?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	25	58.1	59.5	59.5
	No	17	39.5	40.5	100.0
	Total	42	97.7	100.0	
Missing	System	1	2.3		
Total		43	100.0		

Figure 33 - Frequency Tables for Survey Question on Tool and Automation Awareness

These results also show that officers serving within law enforcement are aware of the technologies that exist which could reasonably be foreseen as being of benefit to their work, supporting the findings at organisational level.

The split shown lowers when officers are asked if they are aware of the use of automated tools actually being used within the organisation – with 51% of respondents stating that they are aware of the use of tools, as shown in Figure 34, below:

Aware of tools used?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	22	51.2	52.4	52.4
	No	20	46.5	47.6	100.0
	Total	42	97.7	100.0	
Missing	System	1	2.3		
Total		43	100.0		

Figure 34 - Frequency Table for Survey Tool Use Question

When this is considered in light of the previous work that has been done in this thesis, the finding is initially surprising and seems contradictory to the organisational stance that has been described above. However, further analysis of the comments

made by respondents in later questions, and covered later in this thesis, shows that this result is not as it would appear on face value – several of the respondents have stated that the tools that they are aware of include CAID, which is discussed within *Chapter 3.4* and is not a truly automated tool. As a result, these responses should be discounted when considering the impact of these results – on review, this would result in a much lower proportion of officers reporting that they knew of automated tools. All of the tools that were named in the responses were considered in *Chapter 3.4* and none of these would fulfil the coined '*True Automation*' definition that has been discussed throughout this thesis.

When this analysis is taken into account, the results do support the assertion of organisational reluctance that has been discussed earlier in this chapter, as there is no evidence from responding practitioners that there are any truly automated tools currently in use within the DF domain of the responding organisation.

Supporting this statement is the finding of the following question, which relates to whether the respondent is aware of any frameworks that exist in relation to the control and implementation of automation into the DF domain.

The results of this question are shown above in Figure 28, with the exact results shown below in Figure 35:

Any Frameworks?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	8	18.6	19.0	19.0
	No	34	79.1	81.0	100.0
	Total	42	97.7	100.0	
Missing	System	1	2.3		
Total		43	100.0		

Figure 35 - Frequency Table for Survey Frameworks Awareness Question

As can be seen from the above, 79% of respondents reported that they were not aware of any frameworks that exist which would govern or discuss the use of automation within the DF domain, which would have been contradictory to the finding of the previous question when taken at face value, as the nature of law

enforcement work would require that any tool that was used would be supported by an existing framework.

Based on this, there is a clear awareness of some automation systems in the DF domain, with both practitioners and law enforcement organisations showing an awareness of the potential benefits of automation and tool usage within the DF domain, however, the organisational reluctance to embrace this is clearly creating a barrier to use. This reluctance moves down to the practitioners with a seemingly innate distrust of the automated technologies that exist – this is additionally supported by the final questions of the survey, which asked officers if they would be willing to rely on automation on the basis that the automation was supported by a defined framework, as described in this thesis. The findings of this are shown below in Figure 36*:*

Willing to automate analysis?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	36	83.7	85.7	85.7
	No	6	14.0	14.3	100.0
	Total	42	97.7	100.0	
Missing	System	1	2.3		
Total		43	100.0		

Willing to use automation if framework was created?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	40	93.0	95.2	95.2
	No	2	4.7	4.8	100.0
	Total	42	97.7	100.0	
Missing	System	1	2.3		
Total		43	100.0		

Would you rely on a tool if a framework existed?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		1	2.3	2.3	2.3
	Partially - I would still require a human review to confirm the findings.	33	76.7	76.7	79.1
	Wholly - If the tool was authorised for use, I would accept the findings of any such tool.	6	14.0	14.0	93.0
	Not at all - I would be willing to use the tool for speculative checks, but would have to conduct a full review myself as part of the investigation.	3	7.0	7.0	100.0
	Total	43	100.0	100.0	

Figure 36 - Frequency Table for Survey Automation Acceptance Questions

In addition to the survey responses that have been obtained, the same question about whether officers would be willing to use automation as part of their investigative strategies was also asked within the semi-structured interviews which are described in *Chapter 2.4* and *Chapter 5*. The purpose of asking this question again was two-fold, first to support the findings of the survey above with in-person

discussion and ensure that there was a good understanding of what was being discussed in the survey, and secondly to ensure that the response rates were equivalent, on a pro-rata basis across the different sample size – the survey having 42 responses and the interviews being conducted with 23 officers.

The findings of this can be seen below in Figure 37:

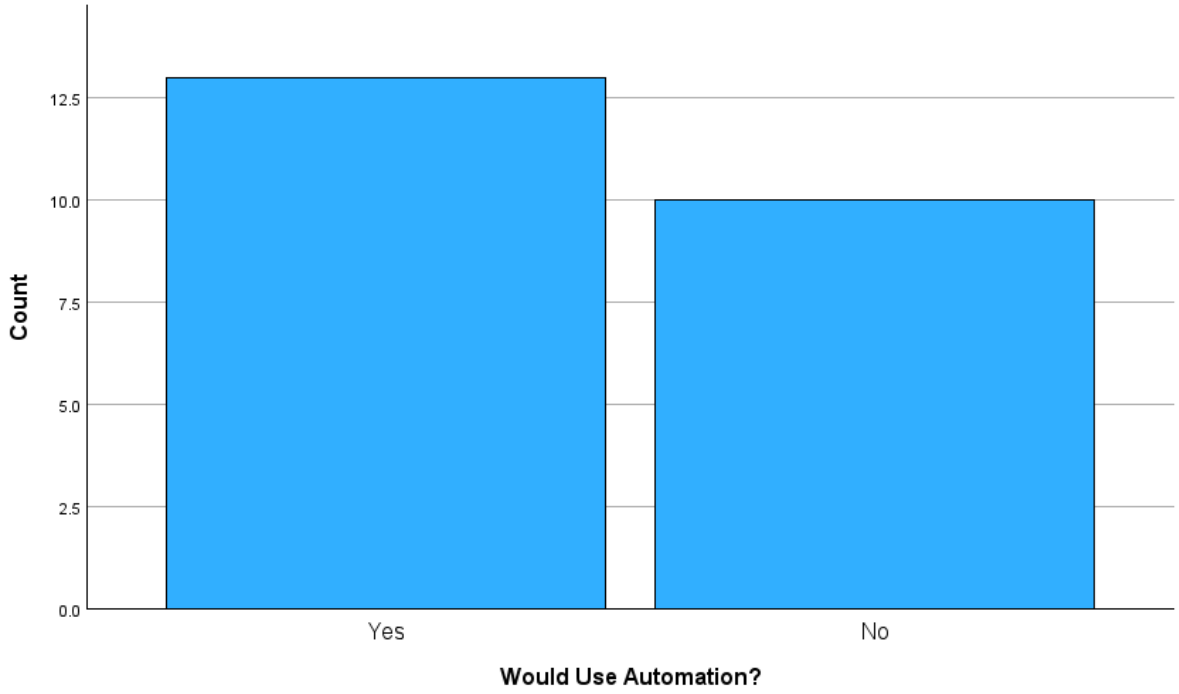


Figure 37 - Results of Interview Questioning on the use of Automation

As can be seen from the above, when answering online surveys officers report that they are willing to use automation as part of their role, echoing what was found in the organisational considerations above, but when they are asked if they would rely on this, the result is that officers report overwhelmingly that they would still require a human review to confirm analysis – with 77% of respondents stating that they would need this review to be comfortable with the tool use. However, again showing a dichotomy of approach, 93% of respondents stated that they would be willing to use automation if it was supported by a supporting framework.

Further supporting this dichotomy, the findings from the survey and the interview are significantly different from each other – opposed to the 84% of respondents on the online survey who stated that they would be willing to make use of analysis, only 57% of respondents stated that they would be willing to do this when asked face to

face.
This finding can be explained by the differing sample sizes but does reveal a disunion of thinking amongst officers which may be as a result of the organisational reluctance and paradoxical approach taken at organisational level.

Based on the findings of both the FOI requests and survey that have been disseminated, there appears to be a clear beginning of acceptance that automation is a factor that must be considered as part of the DF domain, echoing the moves from within industry, however, there is a clear and apparent reluctance to implement the technology fully as evidenced by Figure 35 in which 7% (3 responses) of respondents stated that they would be unwilling to use any form of automation whatsoever, and 77% (33 responses) stated that they would still require human review, and would not be willing to rely solely on automated analysis – with a mere 14% (6 responses) stating that they would be willing to solely rely on automation. This supports the purpose of this research in saying that this should be further explored and addressed in order to allow the domain to reap the benefits of technological advancement fully and address the problems that have been identified within the DF domain and explored in *Chapter 3.1*.

A separate finding that has come up within the analysis of the results that have been obtained from the conducted practitioner survey is that there appears to be a lack of technical understanding and knowledge on the part of practitioners in the field – this is evidenced above in the findings of the survey where the use of the terms automation, AI and ML resulted in confusion over what was meant and the nuances of those terms, and supports the coining of the *'True Automation'* phrase from within *Chapter 1.2* of this thesis. This is further explored in the conclusions to this research found in *Chapter 9*.
This surprising finding is also supported by the opposing findings of the survey and interviews that were conducted. When officers are faced with a direct face-to-face question on the issue, their responses are not always the same as when they are answering the points on their own, showing a potential confusion over exactly what they are dealing with and how this could be used. It is envisaged that the implementation of the framework designed in this thesis would assist in easing this

confusion and provide officers with the confidence to make use of the technologies available by resolving the confusion. This assertion is supported by the final element of the interview in which a large majority of officers stated that they would be willing to make use of the technology if an assessment framework of the type that is proposed in this thesis was created and accepted by the organisation, as shown below in Figure 38:

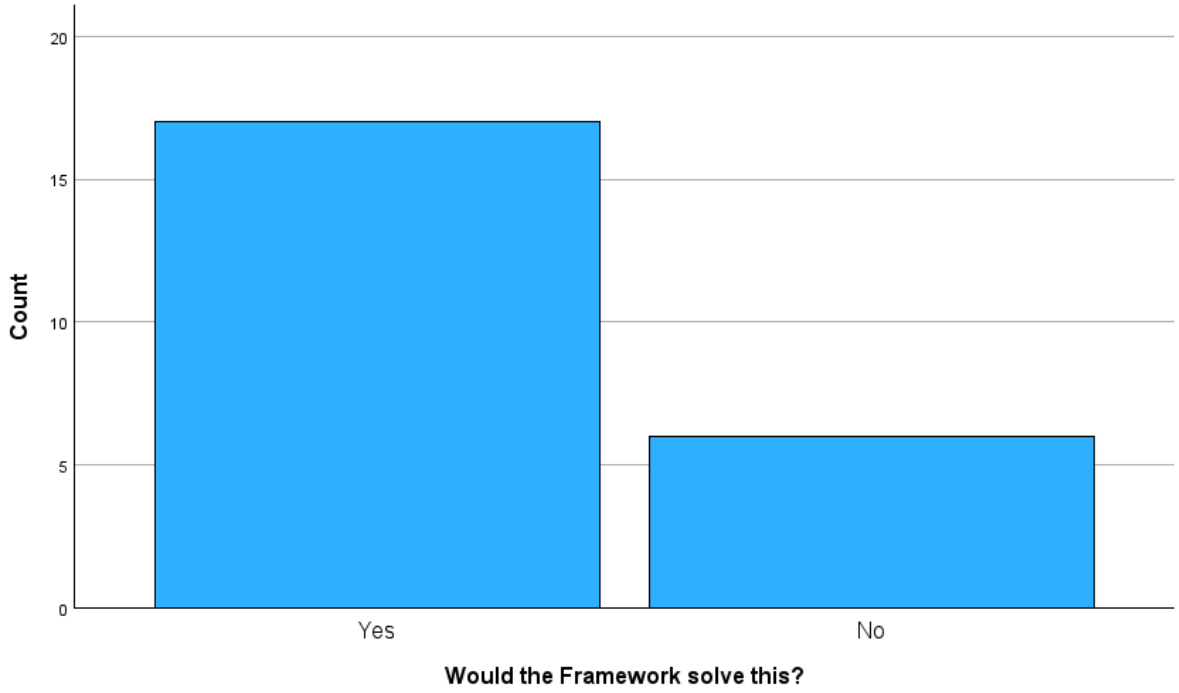


Figure 38 - Visualisation of Interview Question on whether a framework would solve concerns

4.3 Law Enforcement Concerns over Automation Use

The above sections have provided an analysis and research into both the organisational and practitioner state of AI and ML acceptance within Law Enforcement – however, this thesis has also brought several concerns that are held by the law enforcement community to the fore which should be outlined here.

This information was identified through thematic analysis of responses given to the focus group identified in *Chapter 1.3*, the survey described in the above *Chapter 4* sections and the below described *Chapter 5* interviews.

From within these sources, there are several common concerns and themes that have become apparent, these are described as follows:

- *Transparency*

  Within all the described sources, concerns over transparency were raised by officers in various ranks and roles. This concern involved multiple definitions of the term transparency, and when officers were asked about this, they gave two descriptions – the functioning of the tool in relation to its capabilities, and how the tool itself functioned.

  In relation to the first definition, one of the participants of the *Chapter 1.3* Focus Group stated:

  "*I would have issues with developer bias, they would not want to say that the tool does not do things.*"

  In relation to the second definition, multiple participants made reference to the way in which the tool had been trained, the way that the tool functions and how decisions made by tools are reached.

  In combination, these concerns are some of the most significant that have been found in the research that has been undertaken – with it featuring from the very first work that was done in the focus group being undertaken, it was also named as a concern by 12 respondents to the online survey constituting 29% of responses, and finally it was specifically named in the responses of 15 practitioners during the interviews that were conducted which constitutes 65% of responses.

- *Accuracy*

  The accuracy of automated tools and technologies was a primary concern throughout all of the stages of research that has been conducted in this thesis. It was a key theme that was identified in the first focus group, and has consistently come up as a concern and requirement in the survey and interviews that have been conducted as part of this process – it was the main, and at times only, concern raised in 29 responses from within the surveys disseminated, constituting 69% of respondents, and it again factored strongly in the interviews that took place – being a factor discussed by 22 out of 23 interviewees, a 96% respondent rate.

  Determining, measuring and mandating an accuracy level when using technologies such as AI and ML is not always a straightforward task (Räz,

2024) and is further complicated by the understanding concerns that have been explained in *Chapter 4.2*, however, this is addressed in subsequent chapters to this thesis, primarily *Chapter 5.4*.

- *Reliability*

  In addition to concerns around accuracy, reliability of tools and the technologies involved featured strongly in all stages of the research that has been conducted as part of this thesis.

  The issue of reliability has come up, as with accuracy, first in the focus group that was conducted, and in the survey that was disseminated to officers in which it was specifically mentioned as a concern by 20 respondents, accounting for 48% of respondents and it was subsequently mentioned in 100% of interviewees.

- *Verification and Validation*

  During the initial focus group, there were comments made regarding including a requirement that tools and technologies are subject to outside third-party verification, and that several of the in-use tools presently do not undertake this. The comments that were made were in relation to ISO accreditation. These specific requirements were not echoed in the subsequent survey or interviews, however, concerns over verifying the claims of the developers and validating the capabilities of the tool and technology did factor into 29% of survey respondents and 39% of interviewees.

Of note when considering these concerns, there was no inherent nor stated ranking for these concerns, the order that has been given in the above analysis is the order in which they appeared within the described results and has not factored into any form of analysis.

In addition to these concerns, additional elements were raised within the *Domain 4* testing of the Framework that can be found within *Chapter 6.3*.

These concerns included the above, but also added in a concern that there is a cultural block to the adoption of '*True Automation*' in that officers and staff within the Policing sector do not yet trust AI models to the extent that would be required, and they also show that in order for any meaningful adoption of automated technologies without the *human-in-the-loop* component that has been implemented so far, there

would need to be a significant cultural shift which is only starting to be seen as shown in *Chapter 4.1*.

These concerns are separate from the *requirements* of the framework as required by the respondents to the primary research that has been conducted in this thesis, and do not form a basis for gates as defined in *Chapter 5*, however, they do inform these requirements and the framework will be constructed to – as much as possible – address as many of these concerns as possible.

During the analysis that was conducted, it was identified that there was a strong correlation between the concerns that were raised by respondents and the requirements of the framework that they would impose. This would appear to make logical sense and lends credibility to the results of the survey and interview by showing that the respondents were logically responding to the questions that were asked, rather than providing minimal or non-sensical answers.

4.4 Chapter Summary

In summary, the chapter has shown the real-world state of automation in a practical sense through primary research that was undertaken with serving investigators in a Home Office Police Force.

The results of this are shown and show that although there are concerns that are present in respect of such tools, there is a pathway to acceptance on the part of practitioners. They have shown in the above chapter that they would be willing to make use of tools if they were accepted by the organisation.

They have also shown that in a real-world sense, the creation and implementation of an appropriate framework would ease concerns raised and would allow for an increase in acceptance of such a tools and would have a real-world effect in addressing the problems that have been raised.

5.0 A Framework to Implement Digital Forensic-based AI and ML Automation within the UK's Criminal Justice Sector

To address the above findings and increase the use of automation within the DF domain, this thesis creates and implements a dedicated test-based framework against which AI and ML-based automated tools can be assessed with the aim of increasing the viability and use of automation.

The assertion that the creation of such a framework would be supported by the findings of the survey discussed in *Chapter 4* and Figure 31 in which 93% of respondents reported that they would be willing to make use of an automated tool if there was a supporting framework behind it, and the necessity for such work is covered within *Chapter 3.1*.

As such, the following chapter introduces the proposed framework for this thesis, as well as a novel metric to assist in the assessment of the technical performance of tools within that framework. The chapter also shows and analyses the surveys and interviews that were conducted in this thesis, and how these inform the gates that are used within the framework.
Finally, the chapter explains the gates within the framework, how they can be assessed and presents suggested passing metrics for those gates along with an explanation as to why these levels are given.

5.1 Framework Premise and Operation

The premise of the created framework in this area is to create a gate-based system against which automated tools for use within the DF domain could be assessed. If a tool passes a gate, then it progresses to the second – a tool would only be deemed to have succeeded against the framework if it passes every gate.
This approach has seen success (Kapetanidou *et al.*, 2025) in previous deployments within practical applications of computing-based frameworks and seeks to ensure an easy to apply and easy to understand approach, which can be understood and utilised by practitioners from a range of technical backgrounds and knowledge bases.

Additionally, it is considered that this approach of explicit and clear validation and verification of tools ensures compliance with the FSR's codes of practice, giving the practice equal footing with existing scientific standards and procedures within the law enforcement domain (Forensic Science Regulator, 2025).

For the gates themselves, a combination of all prior primary research was utilised, as well as the targeted semi-structured interviews that have been discussed in *Chapter 2.4*.

A thematic analysis was conducted of all information sources, which directed the topic and requirements of each gate – the method for this is explained in *Chapter 2.4*. This was then combined with technical knowledge and archival research into the specific domains named to explain both the gate and how this should be effectively measured.

The flow of information into the framework is demonstrated in Figure 39 below:

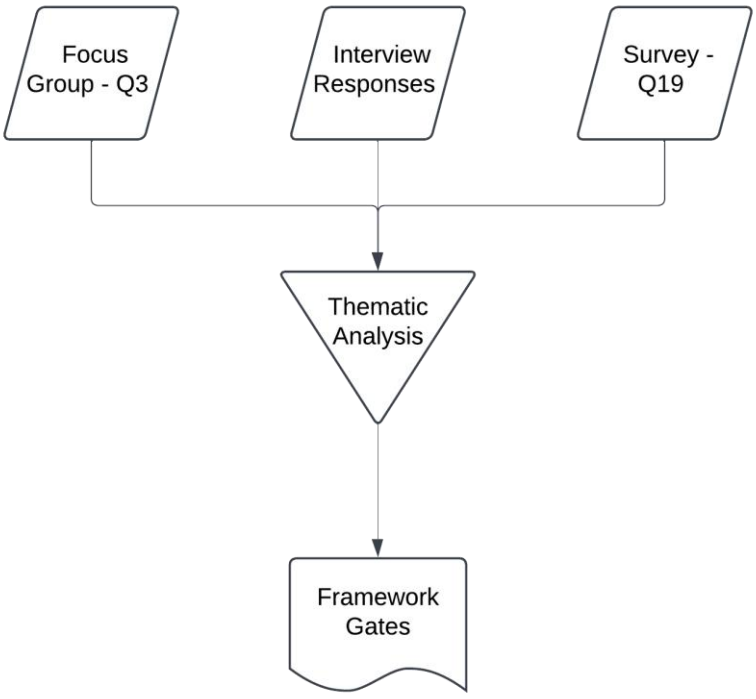


Figure 39- Framework Creation Information Flow

An explanatory series was then created to further explain and define the parameters of the framework, and how practitioners could make use of this, and understand it, in real-world settings, and which would also allow for transparency when making use of

the accepted tools in a court environment. The aim of this step was to reduce the uncertainty around the technical aspects of the analysis, as shown in the above chapters and analysis of the previously discussed survey and encourage officer engagement.

The design of the framework was centred around ease-of-understanding and plain language, utilising a clear flowchart-style approach which has seen success in other deployments in the field of technical analysis and the deployment of technical solutions to traditionally non-technical environments (Kapetanidou *et al.*, 2025).

5.2 Interview, Survey and Focus Group Analysis

As stated above, there are three defined sources of information which have been included in the creation and selection of the gates in the project framework.

The methods of capturing the data that is being used here has been described in the relevant sections of *Chapter 2.4*.

The primary source of data for the creation of the gates in this framework are the interviews that were conducted with a range of serving officers in an English Police Force.

The interviews consisted of a series of semi-structured questions and are described in *Chapter 2.4*.

23 officers agreed to participate in the interviews that were conducted, with the sample being from a range of units and roles all of which had engaged in the survey process that had been conducted.

As previously described, the responses provided were then subject to thematic analysis, the results of which are shown and described below.

During the thematic analysis that was conducted, it became apparent that as with the above, there were some respondents who had focussed upon very specific tool functions, which would not be suitable for a domain-wide framework, and as such when conducting this thematic analysis, it was necessary to filter out non-general requirements, as specific function requests had to be excluded as this is not the purpose of this framework, and it would not be possible to enforce individual requirements for tools across a spectrum and would mean that the assessment

devolved into a per-tool framework which would not be applicable to the overall DF domain.

From the thematic analysis that was conducted, the following requirements were located:

Accuracy

As described in *Chapter 4.3*, the accuracy of the tool was the most prominent requirement that was given across the range of primary methods that were implemented in this thesis.

This featured in 100% of the Focus Group comments, 29% of online survey responses and 65% of in-person interview comments.

It was found, within the thematic analysis of the comments and statements made throughout the research, that there was some confusion amongst respondents between *accuracy* and *reliability*. Many respondents would either confuse the terms or would conflate the two together when discussing the requirements that they would have of any framework or tool.

To deal with this potential problem, a novel method of defining and calculating accuracy has been conceptualised and is discussed below within *Chapter 5.4* to this thesis.

Reliability

The reliability of the tools that are used within the DF domain was a significant concern throughout the primary research that was undertaken in this thesis. It, along with accuracy of results, was the most prominently discussed aspect within the conducted Focus Group at the beginning of the project and continued to feature throughout – with 87% of survey respondents and 100% of interview participants identifying it as a key requirement of any produced framework.

Transparency

The transparency of software utilised by practitioners was another important aspect of a tool's performance and function that was identified with it factoring in all

feedback comments, 26% of survey responses as well as 65% of interview responses.

This was an interesting aspect of the findings, with a significant discrepancy being found in the responses of interview and survey. This can be explained by the fact that the survey had a more significant spread of respondents, whereas the interviews were more targeted towards technical professionals who had direct use of relevant software and so had more understanding of the function of the tools, as well as the limitations and requirements of current policies, procedures and functions.

The exact nature of transparency and what this means in practical terms was also discussed with interviewees, further thematic analysis was conducted on this with the common theme being how the models that are used are trained, on what dataset, and how the tool functions in terms of the technical ability. This is discussed further below in *Chapter 5.4* however, the respondents to the interview did acknowledge that this would be a difficult aspect to implement in real terms due to issues related to trade secrets and commercial interests, however, this was stated to be an aspect that organisations would have to accept when working in law enforcement. From the personal experience of the author, this is also something that is commonplace, with several companies enforcing Non-Disclosure Agreements (NDA) on forensic departments for some aspects – however, it is noted that such agreements cannot be enforced to prevent the Police or other entities from disclosing technical or internal information about a tool when dealing with court cases, as only a court can restrict disclosure under the *Public Interest Immunity* aspects of criminal justice cases (Parliament, 2018; Crown Prosecution Service, 2023). Therefore, it is considered that this requirement is suitable for inclusion within the produced framework as it is a currently accepted and known area of law and criminal justice within the UK.

Verification

The ability to verify the function of a tool, as well as the reported capabilities and compliance with the requirements outlined in this section was identified in all qualitative aspects of the feedback, interview and survey process. This specific

requirement required further exploration and analysis as the exact nature of the concern raised was unclear in the initial comments that were made. The clarification of this requirement is discussed below in *Chapter 5.4*.

Time Saving

The time that the tool takes to perform its function was a key element that was discussed in multiple respondents from the initial focus group through to the surveys and finally interviews showing that practitioners would only make use of a tool that performed at least as fast, or faster, than a human.
This finding was in line with the previously identified problem statement which showed that the time taken for analysis of devices is a primary contributor to the current state of the DF field.

Additionally, there were some specific functions that could be applied to the framework which were not specific to individual tools, and which specified common functions that could be applied to tools domain wide.

These functions were identified as:

Report Generation

The ability of a tool to generate a report of its findings was identified throughout all elements of the primary information gathering in this thesis, and also pertains to all other required aspects of a tool's performance, such as transparency, as the requirement from the respondents was that the tool would not only report the results of the analysis that was conducted, but also how this was achieved, as well as any flaws or issues that would affect the reliability of the tool.

Updating

The ability of a tool to keep itself up to date with the latest training methods, results and models is of vital importance to any software, and this is especially important when discussing a tool which is designed to run without human intervention as the underlying function of the tool is the sole arbiter of functionality and result.
Therefore, this was a strong requirement identified at every stage of the information gathering stage of the project.

Tool Tie-In

There are a wide range of tools that are used at various points within a digital forensic analysis of an evidence source, and no singular tool is expected to cover every element of an investigation – where possible, it is required that any solely relied-upon tool will have the capability to interact with multiple other tools to generate a complete evidential package of analysed images. However, it is accepted and discussed below that this may not always be possible.

Malleability

Digital forensic investigations cover a wide range of situations and circumstances, and investigations may change over time and as a result of the findings obtained during the course of analysis. Any tool that is relied upon in the manner described in this thesis must be able to respond to changing circumstances and findings, as well as be able adapt parameters that are used to conduct that analysis, even if it is that a human has to input those new parameters (depending on the function of the original tool).

These themes were reviewed and combined into a thematic gate-based framework for the assessment of '*True Automation*' tools in respect of DF analysis of image files.

This produced framework is shown below in *Chapter 5.4* and Figure 40.

5.3 A Novel DF Domain Assessment Metric – *Practical Reliability*

Within the unique use case of the DF Domain, a simple one-sided analysis of a tool on either *Accuracy* or *Reliability* is not possible. This is due to the nature of the analysis that is being conducted and the changing nature of images of offence, as well as active attempts at counter-forensics by offenders.

As such, it is key when assessing automated tools from within the DF domain that further steps are taken rather than relying on simple accuracy measures, research has shown that a combined approach to this problem is required including all factors contained in *Precision* (below) as well as *Accuracy* (Garfinkel *et al.*, 2009; Ayers, Brothers and Jansen, 2014).

As such, the author proposes a novel metric for the assessment of these elements, dubbed *Practical Reliability* as explained below.

Accuracy

The measurement of accuracy within the domains of ML and AI is not always simplistic (Räz, 2024).

Additionally, further analysis of the comments in relation to the mention of accuracy within the described primary methods was conducted in this thesis, this showed that the respondents were discussing specific elements such as the ability to detect particular actions or elements such as nudity or age, as well as how well the tool achieved the stated objective, for example detecting pornographic images, or images involved nudity. This evidences a further complication to the assessment of accuracy in the DF domain is the confusion that surrounds the term *accuracy* – with law enforcement practitioners being confused with what the term would mean for any tool that was assessed when using the everyday definition of *Accuracy* as opposed to the ML-based technical definition - in technical terms, the traditional metric of accuracy is defined as shown below in Equation 1 (Sokolova and Lapalme, 2009):

Equation 1 – Accuracy Calculation

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Within this calculation, the abbreviations used and their impact are described below in Table 9:

Table 9 - Automation Accuracy Confusion Matrix

TP – TRUE POSITIVE *Correctly identified positive cases*	**TN – TRUE NEGATIVE** *Correctly identified negative cases*
FP – FALSE POSITIVE *Cases incorrectly identified as positive*	**FN – FALSE NEGATIVE** *Cases incorrectly identified as negative*

However, it must be considered that within the DF domain, a tool having high accuracy may still be missing critical detections, for example Child Sexual Abuse Material (CSAM) that displays material constituting an offence in a novel manner, which would be very difficult for a computer-based system to detect and classify may still have a high accuracy rating as this metric would show that the tool was still able to detect other already known images, but would not accurately reflect that the tool was not able to detect novel or newer images. This is a vital element of image analysis within the DF domain, as offenders are frequently seeking out new material or are frequently altering material to avoid DF detection methods, and so for this reason, the metric of *Accuracy* cannot be relied upon solely when considering whether a tool can be relied upon without the need for human intervention.

Reliability

To deal with the issues of reliance upon *Accuracy, reliability* is another metric that is used to assess automated systems.

For clarity, *Accuracy* measures how correct the overall results of an automated tool are, whereas *Reliability* would show how trustworthy the results of **positive** detections are (Sokolova and Lapalme, 2009).

Within the context of AI and ML systems, the metric of reliability is determined by the *Precision* score and calculation, along with the *Recall* score and calculation.

The precision score of a tool measures what would be known classically as *Reliability*.
The calculation for *reliability* makes use of the same terms as shown in Table 9, and can be summarised as shown below in Equation 2 (Hafdi and El Kafhali, 2025):

Equation 2 - Precision Calculation

$$Precision = \frac{TP}{TP+FP}$$

However, again this measurement does not provide a complete picture as this *Precision* metric would only provide information on how often the analysis correctly and positively identified an image, but would not provide information on how much of the suspect content was actually detected in relation to the overall amount of

illegal material that was present within a digital evidence source – for example, this assessment would show how reliable a positive identification of illegality was, but would not show effective the analysis was on the whole.

For this additional element, the *Recall* metric is used (Hafdi and El Kafhali, 2025). This metric can be calculated as shown below in Equation 3:

Equation 3 - Recall Calculation

$$Recall = \frac{TP}{TP + FN}$$

This would allow for the evaluation of the overall effectiveness of the assessment in relation to the detection of illegal material in the context of the entire data source.

To practically make use of these metrics, one of the traditional approaches has been to combine these metrics together, with the resulting combination being known as the *F1-Score* (Terven *et al.*, 2025). This combined metric ensures that the issues associated with high accuracy, as discussed above, are addressed and produces a much more usable and effective assessment metric. This score works by showcasing disparity between the metrics of *Precision* and *Recall* meaning that results provide a balanced view of a tool's real-world performance.
The DF domain is a key area that the flaws associated with ill-fitting metrics would be prominent, for example when assessing and analysing a set of images extracted from an evidence source, the number of illegal images would, usually, be far lower than the overall number of images contained and therefore a tool could report an extremely high *Accuracy* score by marking every single image as legal, which would not be useful nor actually accurate. The F1 score ensures that this is captured and represented within the result and so enables a full and proper assessment of the tool that is being reviewed.

The F1-Score is calculated as shown in Equation 4:

Equation 4 - F1 Score Calculation

$$F1 = 2x \frac{Precision x Recall}{Precision + Recall}$$

This is a useful metric as it seeks to address the concerns around reliance on *Accuracy* and *Reliability*, which were both metrics that are required by practitioners in

the field, as shown in Chapter 5.2 – however, this is further flawed due to the fact that the F1-Score only focuses upon the positive results, whilst ignoring the negative ones. This is problematic in the DF domain as datasets are likely to be imbalanced with there being many more negative results than positive ones.

To address this, the *Matthews Correlation Coefficient (MCC)* is added. The MCC is considered as one of the most effective measures for binary classification and when datasets are imbalanced, this is very effective in DF analysis tasks as the results are binary in nature, being illegal or legal, and the sets are very often imbalanced as has already been discussed.

The MCC is calculated using the same elements as shown in Table 9, and is worked out as shown in Equation 5 (Chicco and Jurman, 2023):

Equation 5 - MCC Calculation

$$MCC = \frac{(TP x TN) - (FP x FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Once these have been calculated, the results are rescaled to 0 – 1 as shown in Equation 6:

Equation 6 - MCC¹ Calculation

$$MCC^1 = \frac{MCC+1}{2}$$

The resulting values from this range from -1 to +1, with +1 meaning perfect prediction and -1 being completely inaccurate. 0 is included to show that a tool meeting this score would be scoring no better than random (Chicco and Jurman, 2023).

MCC provides improvements over F1 and other metrics that have been discussed as it addresses imbalanced datasets by combining the previously discussed sub metrics and provides a single score for interpretation, unlike F1 which can be more challenging to assess due to the nature of the score and interpretation that is required.

For the purposes of the created framework, all of the metrics discussed above are required. This is because *Accuracy* would measure how universally 'correct' the

analysis conducted by an automated tool would be, whilst the *Precision* and *Recall* metrics, along with the F1-Score, would produce information on how trustworthy the positive detections would be, which is vital consideration that false negatives and false positives can have legal consequences for suspects or result in missed victims.

Whilst MCC is a good addition to this, however, does contain flaws which mean that it cannot be relied upon (Chicco and Jurman, 2023). For the purposes of this thesis, the chief issue here is that the MCC can be hard to interpret for non-technical assessors, and the titular framework for this thesis is designed for non-technical assessors. Therefore, a novel approach is required to obtain the benefits of F1 and MCC, whilst addressing the inherent complexity that would otherwise be a barrier to inclusion in a formal assessment structure.

Proposed Combined Metric – *Practical Reliability (PR)*

In conducting this thesis, the existing *accuracy* and *reliability* calculations above (F1 and MCC[1]) were reviewed. It was identified that these can be difficult to calculate and even more so complex to understand and interpret by non-technical assessors and practitioners. It is also possible that, in the absence of specialist knowledge, assessors would misunderstand the difference between the existing metrics and the limitations that these come with, which are explained in the above chapter.

This potential for misunderstanding is emphasised by the existing misunderstandings which have occurred in discussing the use of technical frameworks in survey questions 15 and 16, and which is discussed in depth in *Chapter 8.3.4*. The existence of these misunderstandings underscores this potential.

As such, it was considered that the existing metrics needed to be condensed into a new metric which would emphasise the positive aspects of F1 and MCC[1] whilst using each metric to address the weaknesses of the other. From the research that was conducted in the completion of this thesis, such a metric does not currently exist, and so the *PR Metric* was conceived. The development of this metric was not undertaken in isolation. Throughout the literary research that was conducted within this thesis, existing performance and classification measures were extensively

examined. Although existing measures provide a range of benefits and insight into the performance of tools, models and algorithms, none of the found metrics address the interpretability requirements of non-technical practitioners whilst retaining the technical validity required for robust and reliable assessment.
The PR Metric is therefore designed to be simple to implement and assess for ease of assessor use and to ensure that non-technical practitioners, or practitioners of limited technical knowledge and ability, can undertake this assessment without this placing an undue burden upon the organisation – for example, requiring the use of an external assessment organisation as has been seen in UKAS accreditation requirements.

The combined metric seeks to address the unique DF domain use case, which will almost always produce unbalanced datasets as has already been discussed, whilst attempting to provide a single average score that non-technical assessors can use for the judgement of automated tools. The designed PR metric is similar to the work that is found within (Cao, Chicco and Hoffman, 2020), however, differs in that the previous work operates across different thresholds, whereas the proposed PR metric in this thesis is a simplified alternative that provides a simple average at one threshold.

The focus of the metric is to provide a simple to use and interpret number that a non-technical assessor can use to judge a tool in a real-world environment.

It is not designed to be a singular test, but rather to form a key element to the surrounding framework, allowing for rapid but informative assessment of *'True Automation'* tools as they develop and seek integration into DF workflows.

The PR metric is calculated using the standard data that is obtained with any assessment of MLA or AI based systems – this is shown in Table 9. Following the obtaining of these, the F1 and MCC scores would be calculated as has already been described.

The MCC^1 result would then be calculated and used with the PR calculation shown in Equation 7:

Equation 7 - PR Metric Calculation

$$PR = \frac{F1 + MCC^1}{2}$$

This produces a single average score including the overall performance on positive detection in the form of the F1 score, as well as overall correctness with the MCC^1score.

The rationale behind the averaging of these elements is to obtain a single, easy to interpret score between 0 and 1 that an assessor can use to quickly determine whether a tool passes this particular element.

It is considered that this new metric, whilst being similar to previous work that has been identified, achieves the stated objectives and is both easy to use and provides an accurate assessment tool for '*True Automation'* tools that can be calculated based on readily available, or procurable data.

An additional identified benefit of this metric and the scoring system that it produces, is that it addresses the findings of the surveys, interviews and focus groups. Within this and has been described within *Chapter 5.2* there was a surprising finding in that practitioners would be willing to accept a tool, for automated tasks, that had a performance level of below 100%, which was the intuitive assumption when beginning this research.

The scoring system that has been created in this metric allows for variance of down to 0.90 whilst remaining within the top scoring bracket. This assists in accepting some level of error, as shown by the practitioners, but also providing a robust level for assessment, as shown below in *Chapter 5.4*.

5.4 A Novel Assessment Framework to Implement *True Automation* into the UK's DF Domain

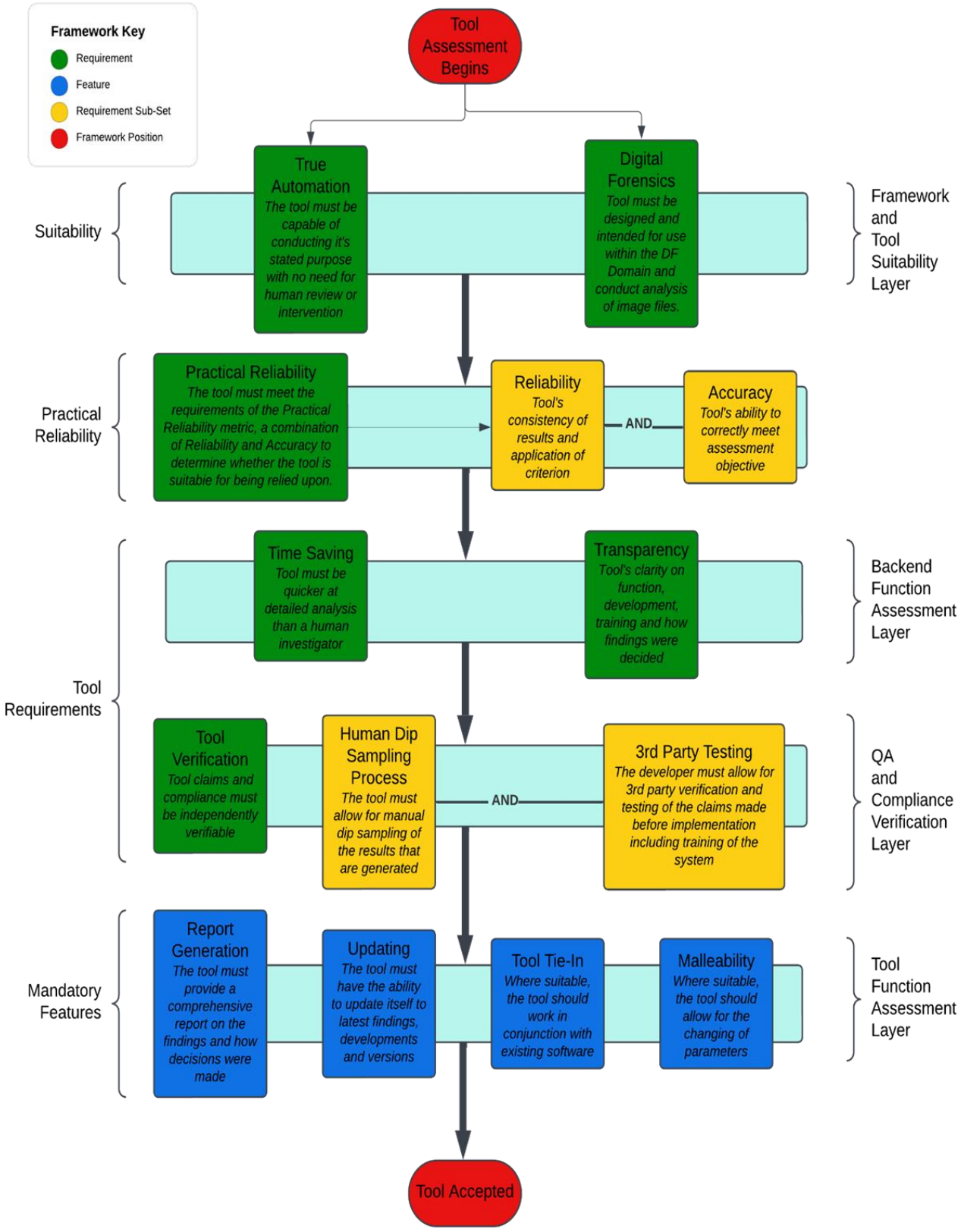


Figure 40 - A Novel DF Tool Assessment Framework

5.5 Gate Explanations & Framework Function

The design of the framework is deliberate in the stages that are undertaken and is designed to work in compliance with ISO 42001:2023 requirements for governance, accountability, and continuous improvement (*ISO/IEC 42001:2023*, 2023). The framework is not intended to be a complete Artificial Intelligence Management System (AIMS) as defined in 42001:2023, but form part of the overall assessment and management system.

The produced framework (Figure 40) aligns with these requirements by providing a structured and repeatable process for assessing DF tool.
Specifically, governance is addressed through the presence of defined criteria and decision points for the tool assessments, accountability is addressed by the framework requiring assessors to evidence whether each gate has been passed and how, and continuous improvement is addressed by the framework, and associated guidance, acknowledging that a tool must be repeatedly assessed and that the framework is not intended to be a one-off assessment.

The framework is designed to be used as a gate-based system with gates organised into assessment layers which group gates for ease of assessment and analysis. A tool must pass each individual stage before moving to the next one, and only when a tool successfully passes each gate is it to be relied upon as a '*True Automation*' tool within the definition discussed in this thesis, and so can be used within the digital forensic environment without the need for human intervention or review.

Each gate is discussed below, with guidance provided as to how the gate metric is to be assessed in a practical sense. Where a specific matrix has been included, there is a suggested value that must be met in order for the tool to be considered as having met the required standard for that gate within the framework, at which point assessment can continue to the next gate, or it can assessed as having failed and either assessment can cease, or the assessor can choose to continue the assessment to obtain further information. However, it should be noted that for a tool to be approved via the framework, it must pass every gate. This value has been reached via consultation with real-world investigators in the same manner and system as was used in previous chapters, however, this should not be taken to be a

static value, it has been set based on generalised operational parameters and can be varied based on organisational risk appetite and decision making, these are not static values, rather living assessment mechanisms designed to ensure consistency across organisations and assessors. There is a section on this also included in the subsequent *Chapter 6.2* testing element of the framework to ensure that the values set met the requirements of investigators in real-world situations.
The method for calculating the metric levels should be conducted in line with the guidance below, which contains suggested methods and criteria which can be used in assessment. It is accepted that the calculation of several elements will be subjective to the assessor, however, clear articulation of the criterion and the metric description should be provided to enable subsequent review of decisions made.

Within the framework, each gate is weighted equally. Each gate is considered a pass/fail, based on the metrics that are discussed below in subsequent chapters. The rationale for this is that each gate represents a distinct minimum requirement for establishing whether a tool can be used within the bounds of *True Automation*. This means that a positive result in relation to one gate does not directly address or mitigate a failing level in another. For example, a tool with a strong result in relation to the *PR Metric* may still fail in relation to transparency or reporting, and the presence of strong *PR* results does not account for nor mitigate this issue.

The strong suggestion to organisations adopting the Framework as an assessment system for their DF department/unit would be that the Framework is a living assessment system. A single one-off assessment is not suggested as this would present the risk of assessment results being skewed or invalidated by updates during the software's lifecycle. This follows standard practice in the assessment of tools and is an organisational decision in respect of tool assessment and as such is not directly included within the Framework as seen in Figure 40.

STAGE 1 – Suitability

The following metrics are designed to ensure that, in the first instance and prior to any further testing or analysis, that the submitted / considered tool is suitable for analysis via this framework.

They additionally seek to ensure that the submitted tool is suitably advanced in the areas of analysis to be suitable for the framework, as basic analytical or review tools would not be suitable, meaning that several later metrics do not make sense or are not applicable.

True Automation

The term '*True Automation*' has been coined for this thesis and has been discussed in *Chapter 1.2* to this thesis.

In order for the developed framework to be applicable to a tool, the tool must first fall into the provided definition of *'True Automation'*.

The method of assessment for this will be varied based upon the function of the tool, and how the tool is designed, however, for the purposes of this framework as well as the '*True Automation*' definition, any tool that is considered for assessment via the produced framework must have differential factors from being simply a repetitive machine. Many tools exist which would be capable of repeatedly doing an action or set of actions without the need for human intervention in the absence of problems or errors, these exist across the world in factories and management systems (Rahman *et al.*, 2025). For assessment via this framework, the tool must be capable of going beyond this and having the ability to pass the later gates contained within the framework. Such tools will require the inclusion of learning elements either in the form of ML or AI as discussed in *Chapter 3*.

Digital Forensics

The DF term is well known and is a key aspect of the computing industry. The term is split into 2 sub-terms, *Digital* meaning:

> *"data, information, or evidence that is stored, transmitted, or processed in binary form by electronic devices or systems, and which can be collected, preserved, analysed and presented using scientifically validated forensic methods"* (Sibe and Kaunert, 2024)

And *forensic*, meaning:

> *"Denoting the application of scientifically validated methods and principles to the recognition, collection, preservation, examination and interpretation of evidence with the purpose of establishing facts that are legally permissible in a court of law, whether in criminal, civil or regulatory contexts."* (Breitinger *et al.*, 2024)

The key aspect of both of these definitions is *scientifically validated*; it is this requirement that the titular framework is seeking to ensure compliance with.

In order for a tool to be suitable for assessment using the framework that has been produced, said tool must meet these definitions, or the purpose of the use of the tool would have to meet these definitions, as tools can be used for a variety of purposes that may not always align with their designed purpose. If the user is using a tool for an applicable purpose, designed or not, then it can be said to be applicable.

STAGE 2 – Tool Requirements

The below metrics pertain to the specific benefits, functions and requirements that are placed on the tools that are being assessed. They provide the method by which the tools can be assessed for both their benefit to the DF domain, and also their suitability for reliance.

When assessing this set of criterions, there is no defined number of attempts or datasets that are required by the framework, this decision is left to the assessor to determine within the specific requirements of their individual unit or organisation. This follows recent studies showing that validation of tools should use sufficient, representative, ground-truth data over prescribed counts (Göbel, Breitinger and Baier, 2025).

Practical Reliability

The PR metric has been discussed in full within *Chapter 5.3* to this thesis, with an explanation as to its inception and use.

For the practical purpose of being contained within this framework, the score is calculated as it is shown in the relevant chapter. Once calculated, the score can be entered into the below Table 10 for assessment:

Table 10 - PR Assessment Matrix

PR SCORE	ASSESSMENT SCORE	DESCRIPTION
< 0.50	0	None – The tool is random or misleading.
0.50 – 0.59	1	Minimal – The tool is unreliable.
0.60 – 0.69	2	Basic – Tool is likely to contain significant error.
0.70 – 0.79	3	Moderate – Tool may contain too many errors.
0.80 – 0.89	4	High – Tool performs well with a small risk of minor error.
0.90 – 1.00	5	Full – Tool very reliable with very low risk.

For a tool to be said to have 'passed' this gate, an ASSESSMENT SCORE of **5** is required.

The rationale for this number being given is that reliability of the tool is of paramount importance to law enforcement, the court system and accreditation under ISO standards (Veksler, Akkaya and Uluagac, 2025). Whilst the work that has been undertaken shows that a 100% reliability or accuracy score is not required, there is a variance built into the PR score as shown above, therefore a score of 5 is the required standard.

Time Saving

The time saving metric is designed to assess whether a tool has an actual real-world benefit against traditional manual review that is currently conducted. A tool that passes this gate would be able to evidence a reduction in examination hours as well as in the overall reporting time for digital examinations, whilst not sacrificing evidential value or quality of output.

The assessment of this metric requires organisational data capture and recording, which on the face of it imposes an organisational demand which would seem to be counterintuitive to the goal of saving time, however, it is considered that the vast majority, if not all, of this required data would have already been captured in order to ensure compliance with the 4 ACPO Principles which have already been discussed.

In addition to a defined procedure for the manual examination of a device in the same manner as would be expected of a tool, for example the grading of indecent images located on an evidential device, the specific data requirements for recording in order for the assessment in this metric to be undertaken are:

- Analyst Time
- Machine Time
- Overall Lead Time

  In relation to the tool's particular activity only [see below].

Particular care will need to be taken when assessing a tool via this metric, as it is considered to be easy to mistake the time saving elements of an entire examination or investigation and compare this to the singular function that a tool may be intended to undertake, an example of this would be that when assessing a tool that is designed to grade indecent images, it is this element for time that should be assessed and not the overall investigation time as there may be other elements, such as the extraction time for the device, which is not impacted by the tool being assessed. The inclusion of these may result in skewed findings of time saving.

The key metrics that can be used for the assessment of the *Time Saving* criterion are as follows:

- Effect on overall lead time for investigations or examinations
- Effect on analyst or examiner time per applicable device
- Number of reworks that are required

  This is designed to include times when the tool either fails in actions or requires restarting, for example a device connection loss.
- Time for setup of tool ahead of analysis

  Including time for examiners to enter parameters or start up particular processes or equipment, for example starting a Cellebrite Premium TurboLink.

In order to assess these in a real-world setting, the following matrix has been developed that an assessor will use to conduct the assessment of this criterion, this is shown below in Table 11:

Table 11 - Time Saving Assessment Matrix

SCORE	DESCRIPTION
0	None – The tool's time-saving benefits are either unproven, or the net effect on the lead time is an increase.
1	Minimal – There is a <10% reduction in either lead time or time taken per action. The time savings that are generated may be offset by the time that is required to either rework or setup the tool.
2	Basic – 10 – 20% reduction in either lead time or taken per action. There is a noticeable level of rework that is required when the tool has completed actions. Minimal setup will still have an effect on the time that is saved.
3	Moderate – 20 – 40% reduction in lead or action time, with minimal reworks required.
4	High – 40 – 60% lead or action time reduction, which is not offset by rework or setup requirements. For example, saving elements are included such as auto-field completion within report generation.
5	Full - >60% reduction in either lead or action time, not offset by rework or setup requirements.

For a tool to be determined as having passed this gate within the framework, it must achieve a minimum score of **3** in this element.

The rationale for this score being given in relation to this metric is that whilst time-saving is a critical component to the use of a tool, there are multiple factors which can impact on the actual amount of time that is saved from the use of a tool, and the actual amount of time that is directly saved can be difficult to quantify and measure (Wen *et al.*, 2025). Therefore, although there is a requirement for this for a tool to be used as there must be a benefit derived, this can be more flexible in relation to the minimum standard that would be accepted.

Transparency

Transparency is a vital element of DF work, as there are specific requirements imposed by various statutes and procedures, which have been explained within *Chapter 3.7* to this thesis.

It is accepted that some of the requirements here are difficult to achieve, and that there are significant considerations around trade secrets and the reticence of some companies to disclose information about the function of their tools and processes – however, it must be accepted that this requirement for transparency is a legal standard and would already be applied to court cases should a tool be used in any event. This has been discussed in *Chapter 5.2* to this thesis.

The *Transparency* metric itself within this framework assesses how clearly a tool's manufacturer will reveal key elements of tool development, such as how the model was trained, the datasets that were used and how the tool makes the decisions that it does when conducting the analysis for which it is used or designed. This requirement is a vital feature underpinning the key requirements of reproducibility, and documentation as required by the previously mentioned 4 ACPO Principles.

There are several key elements that must be present for a tool to be said to contain *Transparency* elements, as shown below:

- Documentation
  The tool should come with, or offer on request, documentation on how the tool functions, how the model was trained, how decisions were made as well as relevant information such as Accuracy and Reliability information as defined in Chapter 5.3.
- Plain Language
  A tool cannot be said to be transparent in function if the documentation or information that is provided with it cannot be interpreted or deciphered by a non-technical assessor or individual.
- Ready Availability
  The required documentation for the tool may be required quickly in response to court or defence legal requests, and therefore although there is no requirement

that the necessary documentation be provided with the tool at point of use, it must be accessible in readable format in a reasonable timeframe as agreed with the relevant client at the time of use.

- Operationally Useful

  The provided documentation must be operationally useful – this means that the documentation should be versioned so that it can later be checked for the relevant version that was in use at the time of the in-question analysis, it should also be searchable and in readable format i.e. not be a raw text file but be formatted and usable in a real-world manner.

To assist with the assessment of this metric, an assessment matrix has been created to assist practitioners. This is shown below in Table 12:

Table 12 - Transparency Assessment Matrix

SCORE	DESCRIPTION
0	None – No meaningful documentation, operation is akin to a 'Black Box', version information is opaque, decision rationale is not recorded or available. Technical information such as training is missing and will not be provided.
1	Minimal – Marketing documentation only, little to no decision rationale, most available information is in the form of text logs although these are often absent or incomplete. Limited technical information is available on request only.
2	Basic – Documentation is present but is largely High-Level and lacks detail. The actual training information is missing or is opaque, requiring expert interpretation which may not be available. Some technical information is provided, with additional key elements only available on request.
3	Moderate – Clear documentation is provided with some guidance accompanying. There are structured logs, but they fall short of being meaningful documentation. Exportable run

	configurations are present. Limited decisions information or validation information. Technical information is available, although remains limited.
4	High – Clear documentation is provided with appropriate and clear version information tied to each tool release. Logs are useful, explained and clear. Exportable decision rationale is provided with some basic explainability tools provided.
5	Full – Comprehensive and full documentation, which is presented without overly relying on technical jargon, detailed dataset information with clear training explanation. Decision logs are apparent and technical information is freely available.

For a tool to be considered as having 'passed' this gate, a minimum score of **3** is required for this element.

The rationale for this level being provided for the passing of this gate is that whilst transparency in AI functionality is extremely important for the use of any tool within the DF domain, there are inherent 'black box' elements to AI, such as chain-of-thought which are not always fully understood by the developers themselves, especially in more complex systems (Lewis-Lim *et al.*, 2025). As a result, there must be some flexibility in the transparency of function that is provided along with documentation, additionally it is considered that detailed technical logs will not always be necessary, nor understood, within the legal profession and so requiring detailed comprehensive datasets or information would not be prudent, it is considered that if there were significant questions raised about the tool after an analysis had been undertaken, that developers would be the ones that would be called to interpret this data in their own evidence. This approach is already being seen in implementations of novel AI systems across British policing, such as with Project Söze in Avon & Somerset Police, with strong success and support.

Verification – Human Dip Sampling and Third-Party Verification

Verification is a necessary feature which pertains to the tool's ability to allow for either practitioners or third parties to validate the claims that are made by the tool's manufacturer.

The method by which the verification can be completed is malleable and may change dependant on the tool, the method of use and the origins of the tool itself, for example crossing international borders when deployed to another country.

There are two elements to the Verification step of the produced framework – Human Dip Sample and Third-Party Testing. Both steps are required for the assessed tool to be considered.

A common theme across both elements to the verification requirement is that this is conducted independently. The tool cannot be said to have been verified by a third party if that third party was contracted by the manufacturer in the first place, as this would leave the tool open to criticism and scrutiny when used in legal fields.

The Human Dip Sample step requires that the tool provide the ability for investigators or examiners, as well as any other party such as defence experts, to review a meaningful subset of the tool's results. It does not mean that the tool should allow for unfettered access to the entirety of a dataset, as this may not be technically possible depending on the function of the tool in question, and may also open avenues for challenge such as claims of undue interference. However, it should allow for the identification of a meaningful sample of the data that has been analysed and provide the ability for edge cases to be checked and re-analysed. The specific requirements on the tool for the information being presented cannot be codified in this framework as this would change depending on the function of the tool and it is therefore the determination of the individual assessor and organisation to determine whether the provided capability of the tool meets the below outlined requirements.

In addition to the above requirement, there is a secondary requirement for this stage of the framework – that being the requirement for third-party verification of claims that are made by tool developers.
In practical terms, this means that the developer of a tool should provide

reproducible evaluation items – for example, the providing of known results and inputs, and the use of APIs for scripted re-runs of ground-truth datasets. Technical information such as readable logs of decisions and function, as well as versioned change logs for updates and training should be made available for third-party expert review.

A good assessment point for this would be a review of End User Licence Agreements (EULA) in relation to the assessed tool – in some cases, developers may include clauses and elements that would prevent verification of the tools that they produce and would prevent the investigation of those claims. This is a strong indicator that a tool produced by that developer would not meet this requirement.

To assist with the assessment of this metric, an assessment matrix has been created as shown below in Table 13:

Table 13 - Verification Assessment Matrix

SCORE	DESCRIPTION
0	None – There is no dip-sampling support provided, outputs are vague or not provided, and agreements such as EULA prevent verification.
1	Minimal – Manual spot checks of function only, limited decision-making visibility, vendor claims cannot be assessed by third parties.
2	Basic – Tool permits some review and simple exporting function. Logs are simplistic and incomplete in terms of technical information; third party checking is possible but is heavily monitored and controlled.
3	Moderate – There is a documented dip-sample process built into the tool, with key exportable elements such as decision logs. Reproducibility is limited and restricted third-party testing is permitted.

4	High – Structured dip-sampling is built into the tool as standard; there are signed released provided with comprehensive logging for the decisions that are made which are readable and can be reviewed by experts in the field. Independent lab verification of claims made is possible with minimal intervention by the developer.
5	Full – End-to-End verification is built into the tool as standard, automatic production of verification datasets is provided, no resistance or interference on third party testing from the developer, logs and documentation provide court level verifiability and defensibility.

For a tool to be said to have passed this gate, a score of at least **3** would be required.

The rationale for this level being set for this metric is that the dip-sampling requirement that was described by the expert participants in the Focus Group for this Framework identified a set level of verification, that being that results are reproduceable and that third party testing is permitted.
The above definitions go beyond this, and the level that is set is given as a minimum standard rather than an idealised optimum. Within the bounds of the field, there is no requirement for automatic verification nor that the tool itself provide court level defensibility as this could be the responsibility of the officers involved, or the organisation conducting the assessment. The requirement from the participants required only that the opportunity for verification was provided by the tool and the developer, this is met by the above level.

STAGE 3 – Mandatory Features

The following gates are designed to be in relation to specific functions that a tool should be able to complete or demonstrate in order for it to be acceptable for use within a DF environment. This requirement is emphasised when a tool is being considered for use without any need for human intervention.

Report Generation

Each tool that is designed, intended, or adapted for use within the DF domain must have the ability to output a readable and formatted report. A generated report within the DF Domain not only evidences the running of the tool and provides key information and audit trails, which are required for compliance with the 2012 ACPO principles, but also serves as the evidentiary document that will be submitted to court, either as evidence in its own right or as unused relevant material, retention of which would be required under the Criminal Procedures and Investigations Act 1996.

As such, any report that is generated by a tool, especially one that does not require human review, should be properly formatted, readable and contain all the information required to reproduce the same results if the analysis were repeated which is a key requirement of the aforementioned 4 Principles.

In order to achieve this, the produced report is assessed based on the inclusion of several elements;

- The time, date and length of the analysis that was conducted
- Investigation details that were entered e.g. exhibit number, reference number and versions of the software that were used
- Any parameters that were entered by the investigator
- A summary of the findings, for example the summary of numbers of indecent image categorisations
- Description of findings such as the type of image and what was shown
- Accuracy levels, as determined by the above section on *Accuracy* within this framework
- A description of any dip sampling that was conducted in accordance with the above section in this framework

To assist in the assessment of these criteria, the below matrix has been developed (Table 14) for assessors and practitioners to apply to any tool:

Table 14 - Report Generation Assessment Matrix

SCORE	DESCRIPTION
0	None – The tool does not produce a report and only provides raw data following completion of the analysis. Output is raw data only with no formatting.
1	Minimal – The tool only produces a basic log, for example timestamped events, with no relevant detail beyond this. Basic formatting may be present, but this is simply a log rather than a presentable report.
2	Basic – The tool produces a report with some details provided but allows for no flexibility, changing of report details, nor does it include key elements such as type of analysis or entered parameters. Report is a basic one but does include formatting to showcase key areas and increase readability.
3	Moderate – The tool produces a report that includes key details and is largely complete, although there are missing elements from the above list. The report is largely a basic findings summary. Report is somewhat formatted and does contain key headings, however, would require further formatting before being presentable.
4	High – The tool produces a comprehensive report which covers the majority of elements but lacks customisation and intricate detail. The produced report is readable and presentable with only minor omissions or formatting faults. Some minor input required prior to court submission.
5	Full – The tool produces a complete and fully defensible report which includes all required elements. Report is fully presentable and is ready for court ‘out of the box’.

For a tool to be said to have passed this gate and move to the next assessment point, it should score at least a **4** on this metric.

The rationale for this level being set for this metric is that reporting is a key element of the assessment system, given that the tool is intended to be used without any human intervention. As a result, the level of detail required from the reports provided by the tool is higher than most other elements, as this would be what would be relied on for many other components such as transparency of function.

In order to achieve other elements, such as time saving, the formatting of the report is thought to be important to avoid further human work being required changing the output of the tool.

Updating

Updating is a critical factor when assessing the practicality and usability of any software tool (Vinayagam, 2025), as the constantly evolving nature of digital technologies means that forensic tools must be able to keep up with the latest developments across a range of target technologies from mobile devices to more advanced computing platforms. These requirements are even more the case when considering and assessing a tool designed to work without the need for human intervention, as the nature of such pipelines make it even more likely that compatibility issues will either cause errors in evidential analyses which may go unnoticed, or would cause failures across the board, resulting in unacceptable delays.

Tools that are assessed via this framework must have the ability to update themselves and keep themselves up to date. However, this process will necessitate the inclusion of some level of user interaction as the tool should not automatically update itself without prompting as different versions of software may have other conflicts, vulnerabilities, issues or other concerns that would require either delayed updating or skipped versions in some cases (Souppaya and Scarfone, 2022). Simply put, a tool should not automatically update without at least prompting for user confirmation before doing so.

In terms of practical guidance, the framework recognises that having some form of update mechanism is preferable to having no update mechanism at all. A tool lacking update capability is likely to lead to investigators relying on outdated and insecure software. A tool that does support updating is more likely to remain within the strict DF domain guidelines and standards, even if that updating method is not the ideal.

These considerations led to the creation and implementation of the below assessment matrix to assist practitioners in the assessment of tools against this metric (Table 15);

Table 15 - Updating Assessment Matrix

SCORE	DESCRIPTION
0	None – The tools are not able to update and is a standalone single release. No patches or versions can be installed.
1	Minimal – Updating is not possible, new versions may be released but this would require a complete uninstall and reinstall of the tool itself every time a new version or patch is released.
2	Basic – Tool has some updating features, but this is irregular and requires significant downtime and manual work to implement.
3	Moderate – Tool supports updating but this is automatically applied with no user interaction.
4	High – Tool supports user interaction on updates with notifications and the option to schedule updates but enforces updating after some time.
5	Full – Tool provides user-controlled updating with additional functionality such as retaining previous versions and allowing for easy user or organisation-initiated rollback.

For the assessment of this matrix, a tool can be judged to have ‘passed’ this gate if it achieves a score of at least **3**.

The rationale for this level being set for this gate is that the only aspect required in relation to this from the participants was that the tool had the ability to update, and this is met by the given level under the definitions given above. The given definition is far from ideal, as automatic updating without human intervention is not preferred due to security and compatibility concerns (Mohayeji *et al.*, 2025). However, this would be sufficient as a minimum standard for the purposes of this assessment framework.

Tool Tie-In

Tying in with other forensic tools enables an assessed tool to work within the existing and extensive digital forensic ecosystem. This involves evaluating integration with established forensic suites such as EnCase, FTK, Cellebrite, MSAB XRY and Autopsy and the ability to exchange data in standard formats.

A tool meeting this metric would be shown to have a strong ability to work within broader workflows instead of operating in isolation. This is additionally vital as the digital forensic ecosystem increasingly works with API-driven architectures to connect tools together in automated pipeline-esque systems.
The import of this kind of function, and therefore the assessment of this, is highlighted by the fact that comparative assessments of forensic tools increasingly make use of this measure already (Vinayagam, 2025).

Additionally, this metric is intended to assess and ensure a level of versatility in the assessed tool across the variety of operating systems and file formats that are encountered in a real digital forensic laboratory. Examples of this include ensuring that the tool is able to function across, or understand data from, iOS, Android Windows, Mac, and Linux-based systems, as well as across the most commonly encountered and utilised file formats such as E01, AFF, RAW, and HEIC as well as the most common standard formats such as PNG and JPG.

This metric also assists, as a result of the above, in assessing how useful and practical the tool is in a real-world context, as the assessment criteria that are given ensure that the tool will fit into existing structures, policies and procedures.

Finally, the metric is designed to assist in assessing other areas as well, for example a tool with strong tool tie-in will naturally feed into timesaving as a versatile tool which can make use of other tool results and outputs will also assist in saving time and making workflows smoother.

The criterion developed for the assessment of this metric are as follows;

- Integration with other forensic tools
  The tool should connect with established forensic platforms and suites such as EnCase and FTK.
- Cross-Platform Support
  The tool should operate reliably on either Windows, Mac or Linux and should be able to take in data from these OS, for example reading images in HEIC format which have been taken on different systems from the assessing systems.
- Data Format Compatibility
  The tool should be able to take in evidence from a variety of evidence sources and file formats, including direct data from the range of common image and video formats such as MP4, JPG and PNG as well as evidential data formats from other tools such as E01, ADD and RAW.
- Stability of Evidence Hand-Off
  Metadata and continuity details must remain intact when evidence is transferred between tools.

In order to assist in the assessment of this metric in a practical sense, a matrix has been developed and is shown below in Table 16 against which a tool can be assessed.

Table 16 - Tool Tie-In Assessment Matrix

SCORE	DESCRIPTION
0	No interoperability – the tool is completely standalone, with proprietary data formats and no ability to integrate or exchange data with other forensic tools or systems.
1	Minimal – Very limited compatibility. Supports only basic import/export and cannot interface with forensic suites or case management systems without extensive manual work.
2	Limited – Supports some evidence formats or partial integration features, but overall workflow tie-in is weak. Might run only on one OS or lack support for common reporting outputs, requiring significant extra steps to use in a broader investigation.
3	Moderate – Provides moderate interoperability. The tool can handle multiple data formats and possibly be used with some additional platforms or has an API for this but still has notable gaps. Cross-platform support or integration is incomplete and only relates to major tools.
4	High – Broadly compatible. Runs on multiple OS and supports most common forensic data formats. Can integrate or exchange data with popular forensic suites and reporting systems with minor limitations.
5	Full – Complete interoperability. The tool seamlessly ties in with other forensic tools and platforms and adheres to open standards for data exchange and is fully cross-platform. It produces outputs and reports that can be fed into reporting systems.

For a tool to be judged as having *passed* this gate, a score of at least **3** on this metric has been agreed.

The rationale for this level being set is that whilst the ability for any tool to work with other formats and file types is important, there is ample precedent for tools making use of proprietary file types in their function (Schatz, 2019). Therefore, although this is an important element that was discussed by the participants of the interviews and the focus group for the Framework, it is not considered that complete interoperability is a key feature that must be focussed on for the purposes of assessment.

As a result, the limited version of interoperability and compatibility that is given by the mid-level set above is considered to be sufficient to allow for a tool to be properly assessed.

Malleability

Malleability is a key element of the framework, however, defining this for a technical assessment can be challenging in real-world situations as many tools will have a defined function, and set of capabilities which may limited their operational scope due to their architecture or intended use case. The real-world challenge that the Malleability metric seeks to address is how adaptable a tool can be to meet the evolving needs of an investigation whilst maintaining the integrity and admissibility of the evidence that is produced as a result of its analysis.

The malleability metric is intended to ensure that a tool that is assessed via this framework can tailor either the given parameters, or defined operational parameters, for any analysis that is conducted. The rationale for this is that investigations can be extremely complex and have varying parameters which change over the course of the investigation and in response to the findings of analysis that is undertaken. A rigid and single-use tool would not be able to adapt to such changes, which would inherently risk missing vital and relevant artefacts or evidence, or risks producing an output which is less valuable to the developing investigation. An example of this would be that during the initial stages of a child sexual exploitation investigation, the focus may be on the identification of images of offence, but as the investigation continues the focus of this may shift to identification of either scenes, victims or

commonalities in elements such as background or clothing, which would become more of an investigative element that numbers of images once the large collection threshold has been passed.

The method of achieving this can vary between tools and interfaces (for example, CLI or GUI), however, there are key elements which should be present for any input to be considered sufficient for malleability, which include;

- User-Defined input for specific search elements
  The tool should allow investigators to specify exactly what is being searched for, whether that is a set of characteristics, hash values or metadata as the investigation requires and the tool function allows.
- Refinement of Analysis Results
  It should be possible to filter or narrow results in line with the requirements of a particular investigation, enabling the exclusion of irrelevant findings without re-running a complete analysis.
- Dynamic Parameter Modification
  The tool should permit changes to parameters during or after analysis, allowing the investigation to respond in real-time to new findings or intelligence.
- Report Amendment Capability
  Generated report should be updatable to incorporate newly analysed data, ensuring that case documentation remains current and reflective of investigative developments. The report should also document and changes made by the investigator, as well as any automated changes to parameters or analysis to ensure compliance with disclosure requirements.

To assess how effective the implementation of the malleability metric has been, and to allow for assessment as to whether the gate has been ‘passed’ in respect of the overall framework, an assessment matrix has been developed for this, which is shown below in Table 17;

Table 17 - Malleability Assessment Matrix

SCORE	DESCRIPTION
0	No malleability – The parameters that the tool uses are fixed, allowing for no user input or modifications.
1	Minimal – limited user input, changes require restarting the analysis from the beginning.
2	Basic – Supports some parameter changes pre-analysis, but no live adjustments during execution.
3	Moderate – Allows for parameter adjustments between runs and supports partial result filtering.
4	High – Supports dynamic parameter modification during analysis, filtering and targeted result refinement.
5	Full – All parameters are fully customisable at all stages of analysis, generated reports are fully amendable and dynamically linked to updated findings.

The agreed level for this element to be considered *passed* has been set at **5.**

The rationale for this level being set for this element is that the DF domain, and the work that is conducted within it, is by definition malleable in the scope, scale and tasks that are undertaken (Bérubé *et al.*, 2025). The parameters of any tool that would work in an automated fashion, without the human-in-the-loop component, would therefore have to be as malleable as the work that is being undertaken, as there would be no review part way through the analysis to ensure that the parameters used were still valid (Andersen, Sunde and Porter, 2025). Additionally, the parameters and considerations being used for the analysis of any image would also, at points, change based on the findings up to a given point, meaning that further malleability is required (Hargreaves, van Beek and Casey, 2025). Finally, the analysis would also need to be able to change based on the data that is seen, for example if there were any information found which would identify the victim of a CSE image.

As a result of these considerations, it was determined that the level required to pass this element was higher than other elements for the Framework, due to the importance that this would have in relation to the analysis that was being conducted, and the loss of the human-in-the-loops component.

5.6 Framework Assessment Matrix

The below Table 18 shows a summary of the assessment levels suggested as a minimum passing standard for a tool to be considered as having passed each relevant gate and therefore the framework as a whole. This matrix is designed as a quick reference guide for assessors and developers to quickly review and understand the required level for each gate.

Complete explanations of each are provided above in *Chapter 5.5*.

Table 18 - Framework Metric Matrix

METRIC	**0**	**1**	**2**	**3**	**4**	**5**
TRUE AUTOMATION	MUST MEET DEFINED CRITERIA					
DIGITAL FORENSICS	MUST MEET DEFINED CRITERIA					
PRACTICAL RELIABILITY						X
TIME SAVING				X		
TRANSPARENCY				X		
VERIFICATION				X		
REPORT GENERATION					X	
UPDATING				X		
TOOL TIE-IN				X		

MALLEABILITY						X

5.7 Chapter Summary

In summary, this chapter has presented the primary contribution in this thesis in the form of the proposed framework.

It has also presented the combined findings of the primary research conducted in the form of gates and metrics based on the comments and concerns made by practitioners.

The framework is presented in visual flow form, with subsequent documentation giving a description of each gate and how this can be assessed, as well as a guide as to what scores would be required for the tool to have passed the respective gate.

6.0 Testing & Validation of Framework

The testing strategy for the framework shown in Figure 40 was muti-staged to ensure that the framework was as applicable to the real-world DF field as possible.

The following sections detail the decision-making process for the testing of the framework, including the definitions and methods that were considered, as well as the rationale for each decision, before moving on to how the framework was assessed at both Expert Validation and Technical Testing stages, before moving to an analysis and interpretation of the obtained results.

The Expert Validation testing that was conducted includes testing of the framework, the PR metric and the matrices scores that were required and is contained within *Chapter 6.3.1* and the Technical Testing in *Chapter 6.3.3* discusses the technical application of the produced framework to real-world applications and software that are currently in use within the DFU of a Home Office Police Force.

6.1 Testing Methodologies

The testing methodologies that were considered in relation in relation to the produced framework in this thesis were as follows:

Expert Validation

Expert validation in this context refers to the deployment of the produced framework to experts and practitioners in the field of DF, and for them to provide their feedback on the provided framework for the purposes of validating that it is applicable to their real-world environment (Amarullah, Ramll and Regina, 2025).

Pilot Testing

Pilot testing occurs wherein there is a small scale, high controlled, deployment of the framework to a real-world operational environment, with a subsequent analysis of the performance undertaking real-world tasks (Stacey *et al.*, 2025).

Delphi Study

A *Delphi Study* occurs when an expert panel made up of various members of a domain, including investigators, lab managers and other experts, respond to anonymised questionnaires about the framework (Fadill and Rochmadi, 2025)

Technical Testing

Technical Testing, in this context, refers to the systematic evaluation of the framework's performance and technical behaviour against defined metrics and requirements (Escalante-Viteri *et al.*, 2025).

6.2 Testing Method Selection

Following the above considerations, the decision that was made in respect of the testing of the framework was to employ a 2-tier approach of *Expert Validation* followed by followed by technical testing of in-use and available DF Image Analysis software – this supports the Research Method approach demonstrated in *Chapter 2.3*.

The rationale for this selection is as follows;

Expert Validation

In this case, the framework is designed to be used by practitioners in the field of law enforcement and DF to conduct analysis of tools that purport to be ready for use within that domain, without the need for human intervention.

As such, it is logical and appropriate that the primary assessment of the framework itself is conducted by said practitioners, as this is the most effective and efficient method by which the real-world applicability of the framework could be achieved.

Additionally, as this method allows for direct communication with the experts in the area, it is considered that this could additionally result in a reduced incidence of bias on the part of the researcher, increasing the reliability and credibility of the results that were obtained.

Finally, the framework that has been designed is intended to bridge the gap between technical developers of tools and the less technical practitioners in law enforcement and related fields who would make use of those tools in the real world. As such, it is

important to obtain the views of those users in addition to undertaking more technical analysis.

Technical Testing

The framework that has been produced in this thesis is designed to be used in a technical area and allow for the assessment of technical tools, albeit by less technical users.

As such, technical testing of the framework to ensure its applicability, and suitability is required.

This will allow for an analysis as to the gates that have been designed and whether it is possible to undertake the analysis that has been intended, as although the framework has been created with the input of practitioners, the technical aspects must also be confirmed as viable.

This testing will also have the secondary benefit to providing an initial assessment of tools that are used within the DFU undertaking the testing, and as such will allow for an initial insight into whether the forensic world even has the capability for *'True Automation'*, or whether this is in line with the larger Policing world as seen in *Chapter 3*.

The rationale for the rejection of the other methodologies given is as follows;

Pilot Testing

Pilot Testing was rejected from the study for the reason that it requires deployment in the real world, which was not possible given the sensitive nature of the work that is undertaken in the DFU space, and the data concerns that arise from the use of AI tools as discussed within *Chapter 3*.

Additionally, the department was able to offer support in relation to the expert validation focus group that was conducted and is discussed below in *Chapter 6.3*; however, the more longitudinal nature of a *Pilot Test* would not be possible given the work requirements of the department.

Delphi Study

A *Delphi Study* was rejected from use within this thesis for the reason that it requires the cooperation of additional members of staff outside of the DFU area that had agreed to help with this study. As a result, the panel would not be sufficiently varied to allow for a *Delphi Study* to take place.

Additionally, such a study requires that the data is obtained via the use of anonymised questionnaires, which would not have been possible in this case due to the time requirements that this would require as well as the need to then disseminate the framework out to those staff members, and wait for them to review it fully, which was another time requirement that they could not provide.

6.3 Testing and Validation of Framework

The testing and validation of the framework was undertaken with experts in the field, specifically the current digital forensic examiners, investigators and managers of the Digital Forensic Unit of a Home Office Police Force. The two stages to this testing are described and discussed in the below sections.

6.3.1 Expert Validation

The *Expert Validation* of the framework was undertaken in line with the above via the conducting of a focus group to present and discuss the framework that had been produced, identify any changes to be made and validate that when the changes were made, that the framework was fit for purpose in the view of the experts that were involved.

The method of undertaking this focus group took 4 stages:

1. Identify the scope of the Focus Group
2. Identify the participants for the focus group
3. Produce the specific questions to be asked of each participant, including Likert scale recording
4. Conduct Focus Group

The method for undertaking the Focus Group was the same as has been described within *Chapter 1.3* in relation to participant anonymisation and information retrieval.

The script that was read out to the Focus Group participants can be found in *Appendix 6* to this thesis, with the specific questions being found in *Appendix 7*.

The below sections discuss the Focus Group itself, as well as an analysis of the results. The completed survey result document with associated comments and Likert scores can be found in *Appendix 9*.

6.3.1.1 Focus Group Scope

The Focus Group scope was to discuss with the participants the framework that has been created and obtain from them their views on this from a practical perspective with a view to adapting and modifying the framework to ensure that participant requirements were met, before moving to an assessment of the framework via Likert scale based closed questions.

6.3.1.2 Participants

The participants for the testing focus group were selected from the available staff within the pool of officers who had undertaken an interview prior to the framework's creation.

Consideration was given to the wider dissemination of the Focus Group invite and for running multiple sessions to obtain wider feedback, however, there were several reasons that this was not undertaken;

1. It was unknown how many officers would be able to undertake this extra session in addition to their normal duties
2. From the results of the survey, it was clear there was a lack of understanding amongst the wider officer pool around the technology which could mean that there were skewed results in relation to the focus group results.
3. It was considered that a smaller, more streamlined and relevant participant pool would lead to results that were reflective of the actual practitioners that would be anticipated to use the framework in the real world

As a result of these considerations, it was decided to include officers from the interview pool i.e. officers who had undertaken interviews, and who also worked in the technical DFU as well as officers who worked in technical related areas such as OCAIT.

The Focus Group invitation was sent to all relevant participants on the above. However, several potential participants were unable to attend due to operational requirement. As a result, the Focus Group was undertaken with the below participants (Table 19);

Table 19 - Framework Validation Participant Information

Participant Number	Unit	Experience Level
1	OCAIT	10 Years
2	DFU	14 Years
3	DFU	11 Years
4	DFU	4 Years
5	DFU	10 Years
6	DFU	6 Years

6.3.1.3 Focus Group Description and Conduct

The method of completing the Focus Group was via group discussion, as described in *Chapter 2.4.3* rather than individual direct questioning. The rationale for this is explained in that chapter and is supported by research in the field which shows that this approach is more effective than more traditional closed-style questions directed at individuals (Bachtiar, Fariz and Arif, 2024).

This approach was adopted, with the only exception being the obtaining of direct Likert responses as recorded in the results document in *Appendix 9*.

All participants were given the domains that were being explored in sequential order to ensure that they understood the objectives of the Focus Group, as well as the specific reason for questions that were being asked. This also enabled an informed discussion around the specific domains and allowed for accurate recording of points that were raised.

The specific domains that were explored within the testing Focus Group are contained within *Chapter 2.4.3*.

To accurately cover each domain, the framework was presented and explained to the participants, with each domain then being mapped to each area of the framework to ensure that discussion was effectively directed to the relevant sections.

Following this, Likert questions were given to the group, and each individual was asked to provide a score of between 1 and 5 (1 being disagree and 5 being agree) that each element or domain was met by the framework. The Likert system was explained at the time of asking these questions, and an explanatory image of the scoring was shown, the same definitions were used in both the Focus Group and this thesis for ease of understanding, and are contained within (AlMutawa, Ikuesan and Said, 2024).

The only exception to the Likert scale was *Domain 4* as this was rather used for the participants to provide feedback on what had been said within the Likert questions and ensure that corrections were made which would assess each of the points raised.

6.3.1.4 Results of Focus Group and Analysis

The results of the Focus Group can be found in *Appendix 9*.

At the time of the Focus Group, the Framework that was presented was very different from the Framework that has been given as the completed version in *Chapter 5.4*, which is the version that was created following the feedback of the Focus Group that was undertaken.

The version of the Framework that was shown to the Focus Group is shown below in Figure 41:

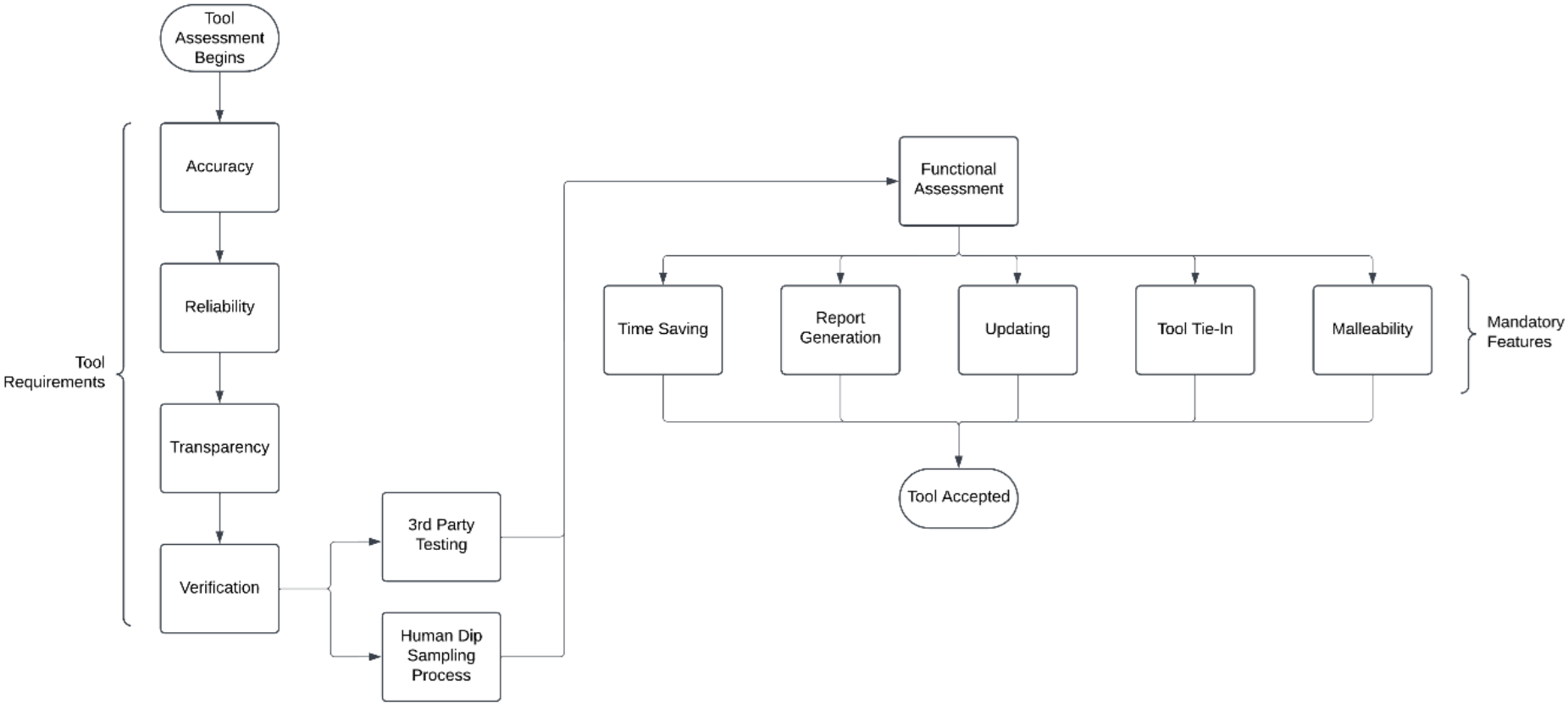


Figure 41 - Framework - Focus Group Version

The summarisation of this feedback is shown below in Figure 42:

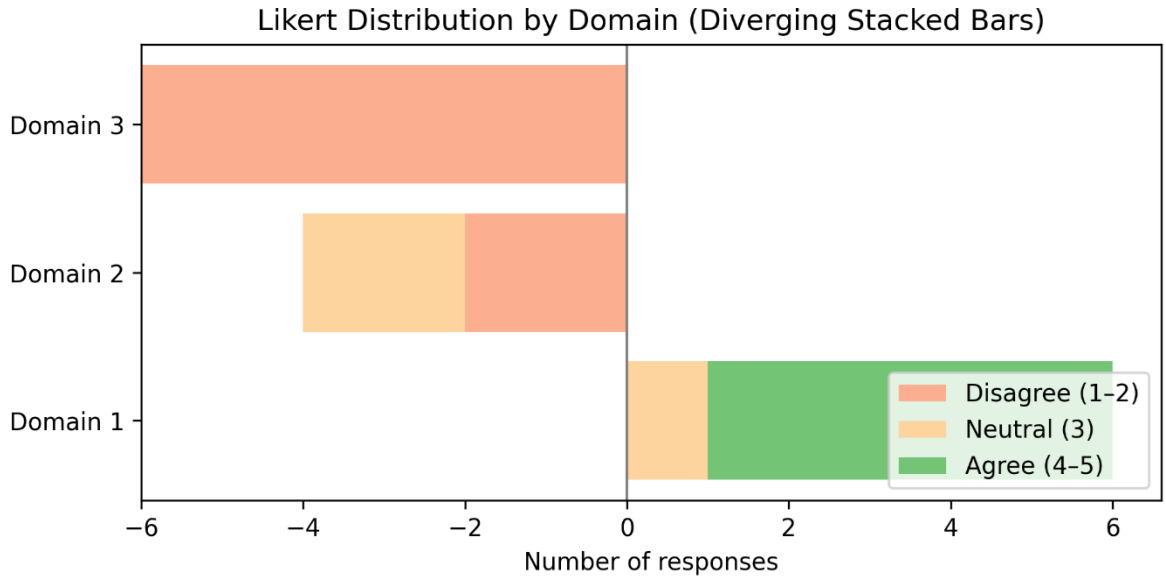


Figure 42 - Likert Results of Framework Testing

The main correction was for Domains 2 and 3 in relation to the usability of the framework. These domains are specifically in relation to the practicality and usability of the framework.

As can be seen from the results, the findings were significant in that the participants universally agreed that the Framework, whilst relevant and valid, was impractical and was not usable.

There was a significant amount of discussion held within the Focus Group around the Domains and the scores/comments that were made in respect of these domains.

These will be addressed in turn in the following section however, the full response notes can be found under *Appendix 9*.

The comments, analysis and amendments were as follows:

DOMAIN 1 – Relevance and Applicability

Within Domain 1, several key themes from the participants comments were identified.

The participants stated that a significant part of the work undertaken with DF, as well as the CJ sector in general, was around accreditation and ISO standards. They identified that the maintenance of this accreditation is a key aspect of any work in this area, and that any additional element that is added to existing workflows and assessments of tools or procedures would have to be in compliance with the relevant ISO standards and accreditations obtained from external partners such as the UK Accreditation Service (UKAS).

This element was carefully considered for inclusion into the final framework as a result of it being so prevalently mentioned. However, the purpose of the framework is to assess the tools themselves for use within the sector and the DF domain, rather than looking at the implementation of passing tools into existing workstreams and processes. The accreditation requirement would be organisational for the receiving organisation and would have to be considered at national level by the relevant bodies. It would not be an element to which the developer of a tool had control, other than to respond for requests for information or for adaptions and changes to be made based upon accreditation feedback and requirements. Therefore, direct accreditation was considered to be not required for the framework. However, there are key areas that accreditation would be required which have been included, such as the requirements for transparency and third-party testing. Therefore, it is thought that the accreditation concern raised by the participants has been addressed by the framework.

In addition to the above within *Domain 1*, the participants identified that a significant issue with the initial framework design was in relation to the updating of tools that are considered. Several of the participants were uncertain around the ongoing assessment of tools, for example once a tool updates with significant functional changes and the effect that this could have. This has since been effectively covered by the manual accompanying the framework, shown in *Chapter 5.5*. Within this guide, it is made clear that the process of assessing tools is an ongoing one, with a mandatory requirement that any tool receiving significant functional or backend updates in relation to the AI models that form the decision-making executive of said tool, would need to be re-evaluated via the Framework. The extent of the re-

evaluation would be a matter of organisational policy, and the onus for declaring said update level would be on the developer. These particular elements are not possible to codify within the Framework, as they are external decisions rather than functional assessment. Tool updates themselves are included within the Framework as a separate assessment component. As a result, this concern has now effectively been covered by the Framework shown in *Figure 40*.

DOMAIN 2 – Practicality and Feasibility

This was a significant domain for the Focus Group. Whilst the relevancy of the Framework was accepted, the usability and accessibility of the first generation of the Framework was called into question.

On delving further into these concerns, a key aspect was that whilst the overall concept of the Framework was around the technical definitions that were used within the created gates. The concern was raised that the terminology that had been used would not always be clear, nor accessible, to the potentially non-technical assessors from within the assessing organisation and would be potentially unclear in scope or specifics to the technical developers. In the version of the Framework that was displayed, the gates were displayed as titles only for the purposes of readability, as shown in Figure 40. Following the feedback of the participants, this was amended to include brief definitions that were being used within the Framework assessment procedure which accompanies the Framework itself, an example of this from the final version of the Framework is shown below in Figure 43:

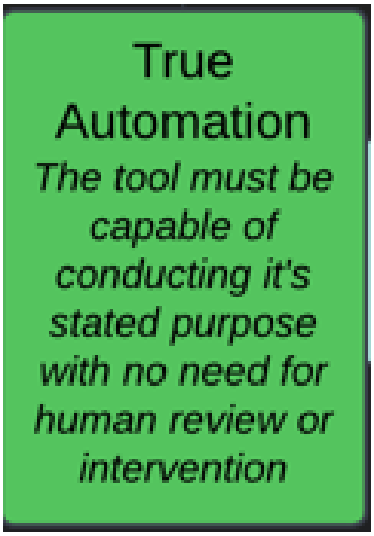


Figure 43 - Definition Inclusion Example

Additionally, the participants raised some concerns regarding the readability of the Framework, leading to the suggestion of including colour coding for the different elements that were contained, and a legend to go along with these colours.

This inclusion was accepted by the participants, with colours being added to show the different elements that were contained, as shown below in Figure 44:



Figure 44 - Colour Coding Legend

Each element of the Framework is based upon the requirements that were given within the primary research conducted and explained in previous chapters.

- Requirement – This element shows that the displayed gate is a primary assessment requirement that is placed on the tool that is being assessed, rather than a function of the tool itself.
- Requirement sub-set – This element shows that the displayed gate is a sub-element to a primary requirement, in the case of this Framework the sub-requirements relate to the breaking down of the primary requirements for Verification and PR. The purpose of these sub-requirements is to show that the assessed primary gate can be broken down, for example the primary

research conducted showed a strong requirement for verification to be included, however there are differing requirements for verification from participant responses and accreditation requirements. The inclusion of a breakdown of the gate enables the use of the Framework across both elements and ensures that the assessment system is relevant and valid to the field of DF.

- Feature – This element shows that these assessed components are in relation to how it functions, or a capability that it must possess in order to pass the relevant gate of the Framework.
- Framework Position relates to the flowchart style that has been adopted for readability and does not relate to the assessment system.

Finally, within this domain the participants suggested the inclusion of a 'plane' based approach, to illustrate the different domains that were being addressed within the complete Framework.

This approach was reviewed by the author, with similar approaches being found in various technological assessment frameworks (Abdi *et al.*, 2024), and so a review was undertaken of the different areas that were being assessed by the Framework, and how these could be grouped and demonstrated more clearly. Based on this review, the following 'planes' were identified as layers to the Framework:

- Framework and Tool Suitability

  This layer assesses whether a tool meets the requirements of the PR metric explained within Chapter 5.3 to this thesis and ensures that a tool is not unfairly assessed when it does not meet these requirements. The Framework that has been developed is specific in relation to the type of tool, that being that it must be designed or intended to be used without a 'human-in-the-loop' element in everyday function and must be designed to be used within the DF domain, specifically in relation to the assessment of image files. Any tool that does not meet this criterion would not be suitable for assessment via the Framework in the first instance and therefore would be assessed against metrics that would not be suitable, leading to skewed results. Therefore, this is the first layer that

must be passed. This layer addresses the requirements of the research project as a whole.

- Backend Function Assessment

  This layer assesses the backend function of the tool, such as the Large Language Model (LLM) or other model in use within the tool. The requirement for this layer addresses participant concerns over the 'black box' element of software development, ensuring that the method by which the tool reaches a decision can be understood, demonstrated and evidenced for the court environment, as well as disclosure requirements, and also ensures that the claims made by the developer can be adequately tested for accreditation, due diligence and legal concerns. The inclusion of this layer directly addresses concerns that were raised within the testing focus group.

- QA and Compliance

  This layer ensures that any tool passing this Framework has the appropriate and suitable verification capabilities that have been identified as being needed for accreditation and reliance purposes throughout the project. A tool passing this layer would meet the minimum requirements that would be expected from a tool being submitted for accreditation under the current ISO standards as outlined above in Chapter 3.3. This layer directly addresses the concerns that were given by the participants in the testing focus group, as well as in other aspects of the project, including the interview stage and the initial focus group in Chapter 1.

- Tool Function Assessment

  This layer addresses the requirements that the DF domain within UK law enforcement would have for a tool that is designed to run autonomously. With this being a novel concept, a tool running without human intervention, there is a focus on the tool's capabilities beyond that of standard tools as there is no human to perform immediate interpretation of the results. As such, the requirements that are contained within this layer would constitute the minimum capabilities that a tool must possess in order to be accepted by DF practitioners, in addition to the designed image assessment capability. This layer directly addresses participant requirements from both the interview/survey stage, and the focus group comments for this domain.

As a result of these steps, it is seen that the concerns raised about the initial Framework have been addressed in the creation of the *Figure 40* Framework version.

DOMAIN 3 – Usability and Confidence

The comments made by the participants to the focus group in relation to Domain 3 were very similar to those that were included within *Domain 2*, and as such they have been addressed by the additions made previously. There are no additional updates in relation to this domain.

DOMAIN 4 – Feedback and Improvements

The comments made in respect of Domain 4 are largely in relation to the cultural concerns of the implementation of '*True Automation*' and therefore are addressed earlier in this thesis, specifically in *Chapter 4.3*, however there are comments made in relation to the method by which several of the gates proposed in the Framework are assessed;

- False Positive concerns were raised by the participants, with the specific comment being that these were the biggest concern in relation to the assessment of the 'accuracy' of the tools. This has been addressed within the PR metric as the FP rate of any tool is given appropriate focus and weighting throughout as has been discussed within *Chapter 5.3*. Based upon the testing of the PR metric that has been conducted in *Chapter 7*, it is seen that these concerns have been addressed.
- Dynamic assessment over static assessment is commented as being significantly more desirable within the assessment of tools that are passed through the Framework.

  This has been researched to ascertain whether this is the accepted standard for the assessment of automated software with the DF domain, and the reasons for this.

  The result of this research is that the use of dynamic assessment methods is vastly preferred within the industry over static analysis. The primary reason for this is that static assessment is likely to miss important updates which would change the function of a tool over the lifecycle of the software, and as

such a dynamic approach is preferable (Claij-Swart *et al.*, 2025). Additionally, this method ensures that the accreditation requirements of the DF domain are met, and that new datasets and information is included within tools that are used.

Given this finding, and in response to the concerns that were raised, the guidance document accompanying the Framework was updated to include guidance that the Framework should be a living process that is built into the assessment strategy of the organisation making use of it, rather than it being a 'one-off' system that is used only at the inception of a tool into the organisational workstream. This was considered for inclusion into the Framework itself as a separate component, however, as with the updating element discussed above in this chapter, this was considered to be an organisational requirement which would vary between organisations and workflows, and as such would be an issue in the implementation of the Framework rather than a component in its own right.

This inclusion of the guide element in relation to Framework implementation appears to satisfy this concern.

- An additional concern that was raised was bias of the datasets that are used to train AI models, and how this would impact onto tools that are assessed as part of this Framework. Whilst this is a cultural concern that has been discussed at length within *Chapter 4.3*, participants did request that this was included within Framework considerations.

  As a result, this was considered within the assessment methods for the tools that are submitted to the Framework.

  This concern is addressed within the PR metric proposed and included within this Framework by the inclusion of both the F-Sore and the MSS Score that have been discussed. These metrics, both individually and in combination, seek to address potential bias within a model, which would appear to address this concern as much as possible, however, it is accepted that bias within datasets is still a concern across the sector.

  There are methods of dealing with bias, such as the use of multiple models to form an answer to a query/task and then the comparison of results to form

the most reliable (Shinoda *et al.*, 2024), however, this is a matter for the developer themselves to address within their development of the tool – rather than being an assessment criteria in its own right. The result of the PR metric will show whether the results of the developer's efforts in this regard are successful enough to pass the gate, which is the most effective inclusion of this concern into the Framework.

- Finally, it was considered by the participants that there is a legal process behind the work that is undertaken by AI powered tools. A specific example that was raised was the 'producing' of the findings of the tool for legal purposes i.e. evidential presentation of the results and other documentation surrounding the analysis that is conducted. Concerns were raised regarding whether the investigating officer could produce these when they have not been involved in the analysis process, as is currently the case with *human-in-the-loop* systems.

  Although this is a valid concern, this was considered as not something that could be dealt with in the Framework itself as it would be an organisational, or national, strategic decision which would have to be undertaken by the receiving entity rather than being an assessment point in its own right.

Following the additions and amendments discussed in this section, the Framework was adapted to the version that is seen within Figure 40, with this version being agreed by the Focus Group participants as successful in addressing the concerns that were raised.

6.3.1.5 Expert Validation Conclusion

The conclusion to the Expert Validation process discussed and analysed above was that the concept of the Framework was a success with several of the features being successful following the additions and corrections that have already been discussed.

The Framework that has been created, as a result, is seen as having been successfully tested against the given domains via Expert Validation and as such has credence for implementation in a wider scale.

The gates that were presented were accepted as being relevant to the field, and logical in their flow and implementation.

The results of the Focus Group discussions showed that there are cultural concerns and blockers that remain, which could not be addressed directly within the Framework, which would be dealt with by the simple implementation of an assessment system. The main comment that was repeated around this in both the Focus Group and the interviews that were conducted, was that if the organisation were to accept the Framework, and the results of that Framework, then they would be willing to make use of the tools that had been assessed.

The results of the Expert Validation, as a result of this, are seen as successful.

6.3.2 Gate Passing Metrics

The original version of the Framework that was presented to the Focus Group discussed in *Chapter 6.3.1* did not include specific values for the passing of the gates were not given,

The rationale for this initial decision was that various tools, functions and models would present differing challenges within the DF and legal environment and so the inclusion of specific values for passing would introduce an element of potential confusion, as well as give avenues for criticism of the Framework in the legal sense.

However, throughout the Focus Group sessions, as well as in the interviews that were conducted, there were repeated comments around the usability and readability of the Framework, especially in relation to the assessment of complex technical elements by non-technical personnel. A surprising anecdotal finding from within the work was that there were a large number of staff who would be asked to feed into an assessment of automated technologies, but who had limited technical understanding of those technologies.

As a result, to address usability concerns, guidance was implemented to the Framework, within the guidance documentation provided, which gave a value that should be met as a minimum standard for passing each gate.

The values that are given are designed to be considered as decision thresholds, with the intention being that they support consistency across assessors and organisations. They are intended to be initially calibrated to a generalised position on the specific gates that they reference, rather than being set values that cannot be changed. These should be tuned for operational risk appetite and organisational stance.

The values for this, as well as the rationale for each value, are given in *Chapter 5.5*.

The metrics used were specifically selected as a result of the primary research that has been undertaken in this thesis. As the framework, and the gates included, are specific to this thesis, and are considered novel in the field, it was decided that the use of any existing metrics for individual assessment of passing values would not be suitable as each element had been tailored to the requirements of the DF field and the unique use case presented in this thesis.
As such, the passing values supported by primary research were used over any existing ones.

The method of testing these values and ensuring that they were valid and accurate, as well as usable in a practical sense, was via Expert Validation in the same manner as the Framework itself, however, due to the late stage of the project that these were added, it was not possible to re-convene the full expert panel that had been used in the Framework testing, and as such Single-Expert Review was used for the testing of these specific values, rather than a full panel as had been used in the previous, and more complete, testing. This was additionally considered proportionate to this stage of the thesis as the purpose of this testing stage was to conduct a focussed assessment of whether the thresholds were practically credible, rather than seeking to establish a domain-wide acceptance.
The selected expert's relevant technical and operational knowledge enabled informed scrutiny within these constraints.

There were limitations to this approach to the testing of these values. Significantly, reliance on the single expert naturally limited the diversity of professional perspectives represented in the testing and meant that there was limited generalisability to the conclusions that were drawn. As such, the conclusions found

in this particular aspect of the thesis should be treated as preliminary only with additional future work required to ratify these with a more diverse stakeholder testing base. These limitations are additionally explored and discussed in Chapter 8. However, although there are limitations to this approach, there is evidence of this method being used in expert evaluation contexts (Kävrestad, 2022), as well as being accepted as a necessary step in evaluating automated forensic processes (Bollé, Casey and Jacquet, 2020). Therefore, although limited, the conclusions drawn from this can be said to be valid, albeit lacking representation of the wider forensic, legal or organisational communities.

6.3.2.1 Participant Selection & Testing Scope

The testing of the metrics was conducted via a single expert, as shown above.

The selection of the participant was, at its core, based upon the availability of staff at the late stage of the project, however, there were also core considerations around:

- The previous involvement of the staff member. It was considered that the reviewer had to have been involved in the previous research actions that had been conducted as this would ensure that they were aware of the project, the scope and the purpose of the Framework, as well as the context in which the metrics were being presented.
- The experience of the staff member. The reviewer had to have significant experience in both the forensic review of images, the legal process, the requirements of that legal process and the tools that would be used within the analysis of image files.
- The knowledge level of the staff member. As this review only involved one participant, the selected staff member had to be knowledgeable in both the above legal issues as well as the technical information that is included in the Figure 40 Framework.

Following the above considerations, a serving Police Officer from the OCAIT unit involved in the previous surveys, interviews and testing focus group was found to be available.

The officer has 10 years of experience in investigations, works in the OCAIT unit which requires the review and analysis of CSE image files, and also has previous

experience working in telecommunications giving them a unique insight into the technological background of this thesis. The personal information of the reviewer was not recorded, as this was not necessary for the review. There was no coding required as there was only one participant. All participant information and rights were given, as per the testing documentation outlined in previous chapters of this thesis.

Once the reviewer had been located, the scope of the review was set out. The sole purpose of the review was to provide comment and insight into the passing levels that had been set into the Framework. The reviewer was asked to review the Framework (which they had previously seen and agreed as above in *Chapter 6.3.1*) and the new passing levels for each metric that had been created.

They were then asked to either agree or disagree with the level and provide comment alongside this.

6.3.2.2 Test Results

The testing documentation, and results, for the passing values is provided below in *Appendix 10*.

In summary, of the 8 passing levels that were given to the reviewer, 7 were agreed with minimal changes and comments, and 1 was not agreed at the level that was set – although this was agreed once the changes described below were implemented.

This represents an 87.5% acceptance rate, moving to 100% with changes. This is shown below in Figure 45.

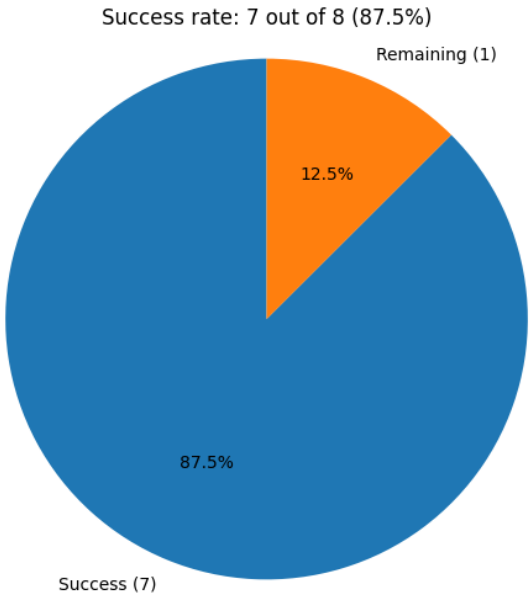


Figure 45 - Passing Level Acceptance Rate

The specific comments and results of this testing are discussed per value below:

- Practical Reliability – Provided level: 5 - AGREED

The comment that was made in respect of the PR metric was that a 1.0 (100%) score at this metric was not required. This was in line with the comments that had been made in the Focus Group testing and was reflective of the level that had been included within the 5 score. Although this is the highest level of score available, it has a built-in variance of 0.9 to 1.0 which was agreed by the reviewer.

This passing level was agreed by the reviewer.

- Time Saving – Provided level: 3 – AGREED

The comment that was made in respect of time saving was that the actual time saved would be dependent on the tool that was being used, and the reviewer accepted that the actual amount of time that would be saved would be malleable based on other processes and requirements around the analysis that was conducted.

They agreed that some time saving was necessary to justify the use of the tool over a human review and agreed that they level and definition given was acceptable as a minimum standard.

This passing level was agreed by the reviewer.

- Transparency – Provided level: 3 – AGREED

The comment that was made in respect of transparency was in line with the other comments made in the Focus Group, and the comments made within the rationale for the level being set in Chapter 5.5. That being, the technical information should be limited and is not always necessary. Transparency is an extremely important aspect of the assessment system and is absolutely necessary for both CPIA and Accreditation requirements, however, there is an accepted limitation to this which has been found with the level that has been set here.

This passing level was agreed by the reviewer.

- Verification – Provided level: 3 – AREED

The comment that was made in respect of verification was that, in the view of the reviewer, it was very similar to transparency, and that it was for the OIC to verify rather than the tool or developer itself.

It was clarified with the reviewer that the point in respect of verification was that the tool offered the capability, rather than conducting this itself. On that basis, the reviewer was happy with the level that was set and the requirement.

This was clarified in the documentation, and the definition that was given for the level. The reviewer then saw this and agreed with it.

The passing level was agreed by the reviewer.

- Report Generation – Provided level: 4 – AGREED

The comment that was made in respect of report generation was that it was an extremely important metric, and something that was most definitively required for a tool to be accepted, especially in light of the 'True Automation' definition that was given for the work.

There were no other comments to make in respect of this matter – the definition and passing level for the metric were agreed.

- Updating – Provided level: 3 – AGREED

The comment that was made in respect of the updating provision was that automated updating was not the preferred option. This was already explained within the rationale,

and this was explained to the reviewer, who agreed with this rationale as well as the level that had been set. There were no other comments made.

The passing level was agreed by the reviewer.

- Tool Tie-In – Provided level: 3 – AGREED

There was no comment made by the reviewer in respect of this metric. The rationale and the level that were both given were agreed with no additions being required.

The passing level was agreed by the reviewer.

- Malleability – Provided level: 4 – NOT AGREED

The malleability score that was given originally was not agreed by the reviewer.
The reviewer pointed out that malleability of parameter, as well as adapting the analysis based on key factors such as victim identification, were of paramount importance to Police work, with the primary focus of the service being the preservation of life and the protection of victims. If a tool were to be trusted to do this thesis without the human intervention that is currently relied on, then this would have to be assured by the developer.
On reflection, this does tie into comments that had been made throughout the process by other law enforcement personnel, who pointed out that whilst the technical aspects of development would focus on key areas such as PR, the law enforcement aspect would prefer that elements such as malleability and adapting to change were focussed on.

Following these reflections, the definition for the levels was updated to reflect this more prominently, and the minimum level was updated to 5 to reflect the importance of this element.

Once this had been completed, the reviewer was once again given this metric with the new elements included, and it was then agreed.

With corrections, this new passing level of 5 was accepted as well as the new definitions.

6.3.2.3 Passing Level Testing Conclusions

The results of the metric passing level testing was that in the main, the rationale that was provided with the levels given was largely correct and in the vein that was required from the law enforcement practitioners. Issues that were found were easily corrected, with a 100% acceptance rate being obtained once adaptions had been completed.

An interesting takeaway from this was the malleability score, which was found to be of significantly higher concern level that what could be traditionally thought of as more important aspects, such as accuracy and reliability as shown in the PR score. Although this was also seen as important to the reviewer, it was overshadowed by ensuring that victims could be identified.

This highlights a key finding of the thesis, that is more fully discussed within *Chapter 9*, that there is a significant need to keep Policing's priorities in mind when developing any form of software or assessment, as this is often not obvious to technical personnel who would not share those same priorities.

6.3.3 Technical Testing

Having completed the above research steps, the final stage of the framework testing was to assess several tools against the framework to ensure that the framework was compatible with real-world assessment and also to ascertain whether the chosen metrics and guidance as described in *Figure 40* and *Chapter 5.4*/*5.5* were both testable and valid.

6.3.3.1 Testing Method

The method of testing for the technical aspect of the Framework is described in *Chapter 2.4*, being technical testing.

The technical testing method involves testing a tool or system against defined test cases, such as the *Figure 40* Framework in this case.
There is precedent for the use of this approach in the testing of DF applications in the Computer Forensic Tool Testing (CFTT) project from NIST (NIST, 2025).

The experimental method was also considered for this testing, however, was discounted as in this case there is no changing of variables and so technical testing is more appropriate.

The FSR requirements on testing were also considered. The Draft Code of Practice (Version 2) 2025 was considered as the most recent version of this and requires that methods used within Forensic Science activities are valid through validation and verification, and that supporting documentation is provided to evidence that the method is fit for purpose. Given the rationale that is provided in this thesis and the supporting documentation around the testing of the Framework and the tools used, it is seen that this is complied with through the use of this accepted testing methodology (Forensic Science Regulator, 2025).

Each of the tools that are identified below were researched for publicly available data or information, and when relevant information was found then this would be used to assess the tool against the Framework in this case.

The purpose of this testing was to ensure that the Framework was valid for assessment of tools, and to ascertain whether there were any tools which would meet either the requirement for additional in-depth testing against the Framework, or which would be seen to potentially pass the Framework.

The criteria for the technical testing were the Framework gates themselves. Although for the purposes of testing, *'True Automation'* was ignored in respect of the first gate as current standards require human-in-the-loop systems and so tools would not be successful against this gate. However, where a tool requires significant human intervention or review, for example simply presenting images to be graded, then this would be assessed as a failure as this would not be an automated system. Additionally, the PR metric was ignored as this is tested in *Chapter 7*.

6.3.3.2 Testing Scope

This testing was achieved by selecting the most commonly used tools within the DFU in which the author served as a Digital Forensic Investigator, as well as academic research into existing technologies.

These tools were found to be (Sunde and Dror, 2021; Falcón-López *et al.*, 2025; Oh and Hwang, 2025), they are also mentioned within Chapter 3.4, although some tools from that chapter have had to be excluded for accessibility and budgetary reasons:

- Magnet Automate
- Magnet GriffEye
- XRY XAMN
- Cellebrite Physical Analyser / UFED
- Magnet GreyKey
- Magnet Axiom
- Magnet Outrider
- EnCase
- Autopsy
- FTK Imager
- ExifTool
- Amped Five / Authenticate
- JPEGSnoop
- Skin Sheriff
- CAID
- S21/LaseriX

6.3.3.3 Testing Results

In relation to each tool, the following was found:

- Magnet Automate – FAIL – Gate 2 – Lack of Analysis

Reference for information obtained in testing:

https://www.magnetforensics.com/products/magnet-automate/

When assessed against the Framework the tool, although one that is frequently used within the DFU of Police Forces in the UK, it is billed as a workflow automation tool, as well as an orchestration tool. It does not conduct analysis in its own right and therefore is not suited to the Framework, resulting in a FAIL on Gate 2 – DF Domain, however, there was confusion over this as discussed in Chapter 6.3.3.4.

- Magnet GriffEye – FAIL – Layer 2/3 - Lack of Information

Reference for information obtained in testing:
https://www.magnetforensics.com/products/magnet-griffeye/

When assessed against the Framework, the tool was found to meet both the gates that were required for Layer 1 (that being that the tool is used within DF image analysis and makes use of computerised automation in the use of the T3K.AI core), moving to the second layer for assessment. However, when a review was undertaken of the public information, it was found that there was insufficient information to make an informed assessment of any gate. For example, whilst there is a claim that the use of the tool makes time savings by streamlining the image detection process, there is no concrete information available which will show the level of time savings, and therefore the inherent risk of relying on an automated tool, as discussed above, is not mitigated effectively and therefore the tool must FAIL on Layer 2.

- XRY XAMN – FAIL – Gate 1 - Lack of Automation

Reference for information used in testing:
Researcher knowledge and https://www.msab.com/product/analyze

XAMN is a commonly used tool for a range of DF activities in relation to a range of devices. The software enables the extraction of devices through XRY itself and then presents these in a user interface for examiner review through the XAMN software There is categorisation of the file types that are present, however, the software at either element of this process does not include any form of automation or AI/ML powered process.

As a result of this, the framework does not apply to this software and any continuing through this would present false results as the original concept is not met.

Therefore, a result of FAIL under Gate 1 must be recorded.

- Cellebrite Physical Analyser / UFED / Premium ES – FAIL – Gate 1 - Lack of automation

Reference for information used in testing:
Researcher knowledge and https://www.ijfmr.com/papers/2025/2/41813.pdf

Cellebrite UFED, Physical Analyser and Premium ES are a commonly used and powerful extraction and display software suite which is used across the world for the extraction and analysis of a range of digital devices (Soni, Pawar and Shukla, 2025).

The software enables the extraction of devices through the Premium ES and UFED elements, as well as the display of identified artefacts for examiner review and categorisation.

However, in both the public information for the software, published articles and the researcher's own knowledge, there is no automation element to this software nor is there any specific use of AI or ML in relation to any of the functions.

A review of the further gates has now been conducted as the findings would be misleading due to the original premise of the Framework not being met.

Therefore, the result of FAIL under Gate 1 must be recorded.

- Magnet GreyKey – FAIL – Gate 1 – Lack of Automation
Reference used for assessment:
Researcher knowledge

GreyKey is a commonly used tool within the DF domain in the extraction of, and processing of, mobile phone extractions.

Whilst this would meet some of the aspect of the Framework, it does not perform any processing of image files and does not conduct any form of analysis on the data. It extracts information and presents this to an examiner.

Therefore, the tool must FAIL the Framework on Gate 1 – 'True Automation' as there is no automated analysis component to the tool.

- Magnet Axiom – FAIL – Layer 2/3 - Lack of Information

Reference used for information: Magnet Axiom User Guide: http://092f67184f02fcdb918c-b3d937de523d4a3d4cea730efa685a0d.r37.cf1.rackcdn.com/AXIOM%20docs/Magnet%20AXIOM%20User%20Guide.pdf

Magnet AXIOM is a data acquisition and analysis platform that is used throughout the DF domain across the world.

Layer 1 of the Framework is passed with the use of this software, it is designed for the DF domain and the analysis of relevant files, and the documentation that is provided shows that there is an AI analysis component with the use of the Magnet.AI engine, which performs a range of image analysis functions including categorisation.

However, the second layer of the Framework does not allow for reliance on AXIOM. Although there is mention within the documentation and academic research that the use of the tools will streamline the process of image analysis and processing of digital evidence sources, there is no information regarding the specific time saving that would be gained from the use of this software. As a result of this, the time saving cannot be – on the currently available information – quantified to the satisfaction of the gate and therefore a failure result must be recorded for this.

Additionally, although there is some information that is available for public review in relation to the AI function that is in use, there is little to no information regarding the model that has been used, nor regarding the training and backend information for the model. As a result, the tool must also fail the transparency and verification gates.

Finally, AXIOM is also assessed as failing to malleability requirement within the Framework, which requires that there is full malleability of the parameters used both at the outset of analysis but also during the analysis and in response to the findings of the analysis as they occur. From the information that is available on AXIOM, this is not a function that exists, there is nothing that would suggest that this is something that the software would be capable of conducting, and therefore the gate must be failed.

As a result of these, the Transparency, Time Saving, Verification and Malleability gates must be failed.

- Magnet Outrider – FAIL – Gate 1 - Automation

Reference used for information: Researcher knowledge and https://h11dfs.com/wp-content/uploads/2024/09/MF_Outrider_ProductReview_20240309.pdf

Magnet Outrider is a portable solution that allows for triage of digital devices at scenes and an assessment of sources at scene, without the need for extraction of the device.

The software is capable of the analysis of a range of file types, including images, at scene. This is an automated identification feature and does allow the software to identify CSAM or other images of offence. However, although this on the face of it this would meet Gate 1 for the Automation of analysis, this does not make use of any element over HASH function comparison. There is no AI or ML component, nor real automation over information comparison which must also then be confirmed by the examiner. There is no analysis that would meet the requirements of the Framework, nor this test.

As a result, the tool must be recorded as FAIL on this element.

- EnCase – FAIL – Layer 2/3 – Lack of Information

Reference for information used in testing:
https://www.opentext.com/uk/products/forensic

EnCase is one of the most commonly used, and widely trusted, DF software in use. It has a range of features and functions within multiple realms of DF.

There is a specific use case for the analysis and categorisation of image files, and from the information that is available, some of the most recent features in this area make use of AI for this purpose.

As a result, EnCase successfully meets the requirements of Layer 1 of the Framework – that being the conducting of automated analysis of image files for the DF domain.

However, although there are more information and documentation available in respect of the software with its prevalence and popularity, the claims made in publicly available information are that the use of the software is powered by AI, and that it reduces the time that is required for manual analysis and processing of image files.

There is no specific information available in either public company information nor in academic research which would answer the requirements of transparency and time saving to the extent that is required by the Framework.

Also, the malleability gate is not met. There is no information available which would show that the parameters that are used within the analysis conducted can be changed nor modified, and there is nothing to suggest nor state that the parameters would update in real-time in response to any of the findings on the device.

As a result of this, the tool must be shown with a fail in respect of all gates on Layer 2.

- Autopsy – FAIL – Gate 1 – Lack of Automation

Reference for information used in testing: Researcher knowledge and https://github.com/sleuthkit/autopsy_addon_modules?tab=readme-ov-file

Autopsy is a free, open-source GUI for a series of forensic modules, designed to allow the forensic examination and analysis of a range of digital evidence sources. It is widely used and accepted within the DF community, although there remains some question over its validity in terms of accreditation.

However, from a review of the public information available in respect of both the tool and the third-party modules that are available for use with the standard release, there is little information to suggest that there are ML or AI modules actually in use within the suite of available software.

As an example, the result of searching for ML is shown below in Figure 46*:*

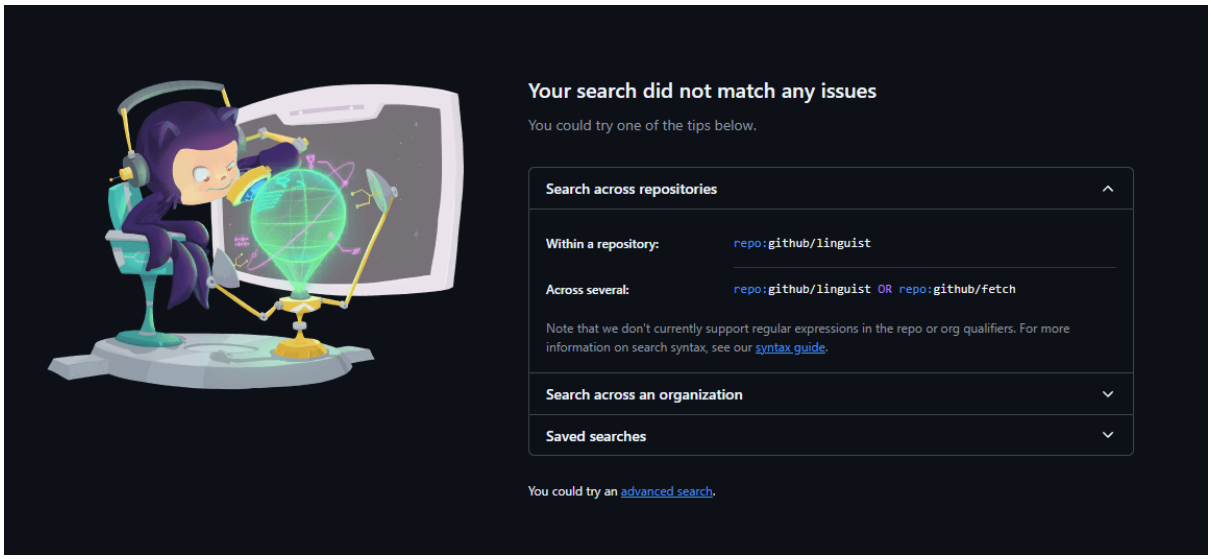


Figure 46 - Autopsy 3rd Party Module Search Result

There is some reference to the capability and consideration to these types of actions within the larger scope of the SleuthKit itself, but the available information, it appears that the primary analysis method of images in relation to Autopsy remains HASH comparison, as shown below in Figure 47*:*

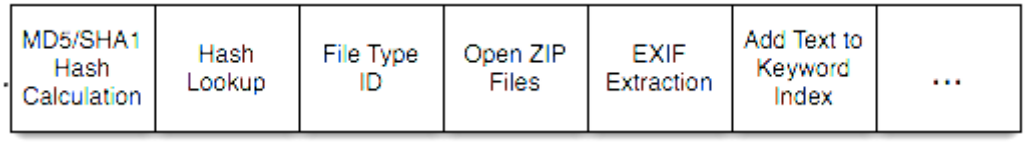


Figure 47 - Autopsy Function Pipeline (Autopsy, 2018)

As a result, it must be considered that even if there is a module that has not been located in this testing which would make use of automated technologies, the information on this is not readily available, which would preclude proper assessment via the Framework.

As a result of this, it must be taken at face value that there is limited to no use of AI nor ML in the current iterations of the SleuthKit nor Autopsy and therefore a result of FAIL must be recorded for this.

- FTK Imager – FAIL – Gate 1 - Lack of Automation

Reference for information used in testing: Researcher knowledge and *https://www.exterro.com/uploads/documents/FTK_7.4.2_UG.pdf*

FTK Imager is a popular and well used forensic extraction tool for a range of devices, including memory cards. It is designed around this extraction and therefore has the same concerns as other tools in respect of automation.

From the information available and the knowledge of the researcher, there is no use of AI, ML or other automated analysis function within FTK that would satisfy the requirements of the Framework, and therefore, a result of FAIL must be recorded in respect of Gate 1.

- ExifTool – FAIL – Gate 1 - Lack of Automation

Reference for information used in testing: Researcher knowledge and *https://exiftool.org/exiftool_pod.html*

ExifTool is a metadata review and extraction tool, which enables an examiner to review metadata of image files that are supplied.

It displays this data for the examiner; however, the tool conducts no analysis and no automatic function. As a result, this does not meet the requirements of the Framework for Gate 1 and so a failure result must be recorded.

- Amped Five / Authenticate – FAIL – Layer 2/3 – Lack of Information

Reference for information used in testing: https://ampedsoftware.com/authenticate

The AMPED Authenticate software suite is a range of tools which allow for the review and analysis of a range of file types, including image files.

The software meets Gates 1 and 2, in respect of its function – it is both intended for the use of image analysis within the DF domain, and certain elements of the software make use of AI elements to conduct some analysis of files, such as DeepFake detection.

However, from the information that is available, Layer 2 must be recorded as a failure for the purposes of the Framework.

From the information that is available, there is no specific information regarding the model that is in use, how this was trained, nor is there any information regarding actual time saving that is achieved from the use of this particular tool.

Therefore, as the information is not freely available, the tool cannot be said to have met the passing levels of these gates and as a result must be recorded as having failed to meet Layer 2 of the Framework.

- JPEGSnoop – FAIL – Gate 1 – Lack of Automation

Reference for information used: https://github.com/ImpulseAdventure/JPEGsnoop

JPEGSnoop is a tool that allows for the analysis of image files, including indicators for edits and a review of metadata.

There is no automatic function contained within this tool, nor is there any analysis that is conducted as a part of the tool's functions.

As a result, a result of fail must be recorded for Gate 1 of the Framework as there is no automated function available.

- SkinSheriff – FAIL – Gate 3 – PR Metric

Reference for information used: https://dl.acm.org/doi/10.1145/2598918.2598920

The available information for this tool provides the results under the F-Score for the tool as 0.496.

This value has been considered in relation to the PR Metric and the potential scores that would be available for this.

Given this score, it is considered that this value is too low to allow for a potential score under PR at the level that would be required for passing.

As such, the tool fails the PR Metric gate.

- CAID – FAIL – Gate 1 – Lack of Automation

Reference for information used: https://www.iicsa.org.uk/reports-recommendations/publications/investigation/internet/part-c-indecent-images-children/c2-detection-images.html#:~:text=%E2%80%9CYou%20can%20take%20an%20image,is%20the%20same%20or%20similar.%E2%80%9D

CAID makes use of hash functions for the storage, recall and comparison of image files.

Whilst this does use a technological solution to achieve this result and the comparison is conducted using only a technological component in the hash function, this is not considered to be sufficient to satisfy the requirements for automation under the above restrictions and therefore the tool must fail this gate.

- S21/LaseriX – FAIL – Layer 2/3 – Lack of Information *Reference for information used: https://www.semantics21.com/s21-laseri-x*

The S21 software has been reviewed and whilst this does meet the requirements in terms of automation and is in correct domain meaning that it does pass Gates 1 and 2,

there is no information that can be seen which would satisfy the requirements for Layer 2.

As a result of the lack of this information a result of fail must be recorded for these layers as a proper assessment is not possible.

6.3.3.4 Technical Testing Summary

A summary of the Technical Testing results, specifically where each tested tool was unsuccessful against the framework is shown below in Table 20, below:

(Blank elements show where a tool was successful)

Table 20 - Technical Testing Summary

METRIC	*Magnet Automate*	*Magnet GriffEye*	*XRY XAMN*	*Cellebrite PA, UFED*	*Magnet GreyKey*	*Magnet Axiom*	*Magnet Outrider*	*EnCase*	*Autopsy*	*FTK*	*ExifTool*	*AMPED*	*JPEGSnoop*	*SkinSheriff*	*CAID*	*S21*
True Automation	*X*		*X*	*X*	*X*		*X*		*X*	*X*	*X*		*X*		*X*	
DF Domain																
PR		*X*				*X*		*X*				*X*		*X*		*X*
Time Saving		*X*				*X*		*X*				*X*				*X*

Transparency		X				X		X				X				X
Tool Verification		X				X		X				X				X
Report Generation																
Updating																
Tool Tie-In																
Malleability																

6.3.3.5 Technical Testing Conclusions

One of the largest findings of the technical testing that was conducted was that the description for DF Domain was not sufficient. In the original version, this simply stated that the tool must be designed for the DF domain, as shown below in Figure 48.

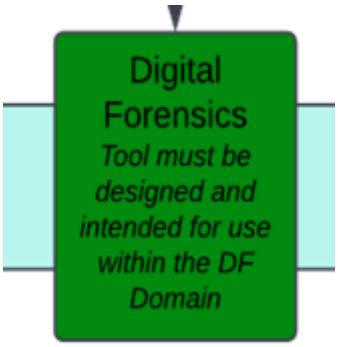


Figure 48 - Original DF Gate Definition

This caused some confusion for tools like *Magnet Automate* as this was designed for the DF domain but does not actually carry out any analysis itself. As a result, it is not suited for the Image Analysis specific version of the Framework that is presented in this thesis, however this presents an interesting use case for the Framework to be expanded beyond the current Image Analysis limitation to the DF domain as a whole. This is discussed further in *Chapter 8*.

To address this limitation, the definition of the DF Domain gate was expanded to be more specific, as shown in the *Figure 40* complete Framework.

The primary finding within the technical testing that has been conducted is that, although all tools submitted failed assessment by the Framework, this shows that the gates themselves are applicable as an assessment could be carried out against the tools that are currently in use within the DF domain.

From a review of the reasons for the tool's failures within the Framework, it can be seen that 62.5% of the tested tools failed assessment due to a lack of automation in the meaning of the testing that was conducted, whilst the remaining tools failed assessment due to a lack of meaningful documentation and testing which would allow for appropriate measurement of the benefits to reliance.

6.3.4 Comparative Testing Use and Suitability Summary

The below Table 21 shows a comparison between the testing methods that were used in this thesis to assist in understanding how each test method works towards validating and assessing the produced framework shown in Figure 40:

Table 21 - Framework Testing Comparison

Test	Use	Suitability
Expert Validation	Validates that the framework is suitable and applicable to the domain and the field in a practical sense.	Suitable for assessment as the framework is intended for operational use and therefore practitioner review by individuals with domain knowledge and expertise is required.
Gate Passing Metric (Single Expert)	Validates that the proposed passing metrics are fit for purpose and applicable to the practical domain.	Suitable as an initial validation as to whether each threshold is practically credible and understandable is required.
Technical Testing	Validates that the framework can be applied to real-world tools and performs initial assessment of existing tools.	Suitable as applying the framework to real-world tools is required to assure usability and demonstrates whether the criteria can be consistently applied to real world tools.

6.4 Testing Conclusions and Analysis

The testing of the Framework shows that the version of the tool that has been shown in *Figure 40* is both accepted by practitioners in the sector, and that it is suitable for the assessment of tools that are currently in use.

In relation to acceptance of the Framework by practitioners, as discussed in *Chapter 6.3.1 and 6.3.2*, once the concerns that were raised were dealt with by versioning the Framework, it was confirmed that it would be accepted for use. However, it must also be asked whether a tool passing the Framework would be accepted or not.

To answer this point, consideration was given to the surveys, interviews and questionnaires that had been disseminated earlier in the project, and which is discussed in *Chapters 1.3, 2.4.2,* and *5.2*.

Within these actions, the initial focus group participants stated that they would be

willing to accept a tool that had passed the Framework (*Appendix 1*), 92% of valid respondents within the practitioner survey and 73.9% of interview respondents stated that they would be willing to make use of automated tools without human review should the tool pass the produced Framework, as demonstrated below in Figure 49:

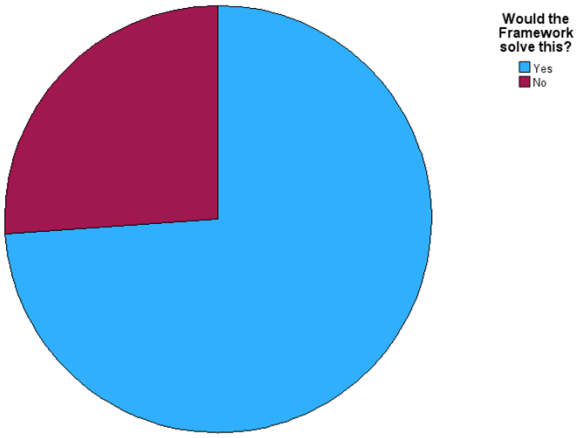


Figure 49 - Interview Respondent Representation

The combination of these findings and the results of technical testing showing that the Framework is relevant, accepted and applicable, shows that it is entirely possible for such an assessment to be conducted, and that if a tool were to pass this, that the tool would be accepted.

Additionally, the Framework has shown that despite there being tools available within the DF domain which claim to make use of AI-powered features or automated functions, there are no tools which have been tested which would meet the level that would be required for '*True Automation*'.

This finding is not surprising and is supportive of the approach that forms the current standard within the UK CJ sector, that being the *human-in-the-loop* approach that is discussed in *Chapter 3.3*. This shows that the current development of tools is not intended for autonomous deployment and that human review remains centric to the intended use of the tools that are being utilised.

6.4.1 – Future Work

Future testing of the Framework is suggested to be expanding this beyond the singular Force that was involved in the testing that has been conducted in this instance. This would allow for a wider scope of input and ensure that the Framework is more representative of the requirements of the domain as a whole rather than the single Force.

Additionally, longitudinal deployment of the Framework to real-world active units for the assessment of tools proposed would also enable a wider view of the real-world requirements than are currently present within the produced version and would also enable a more thorough assessment of the Framework's real-world applicability.

Finally, technical testing of the Framework could be expanded to include further tools that are being used or developed for the DF Domain in respect of image analysis, as well as a more in-depth analysis of the findings to ascertain the exact elements of the Framework that are causing concern.

This would additionally allow for consideration of including developers in the testing of the Framework to ascertain if the Framework is applicable from the technical development arena as well as the law enforcement one that has been considered in this thesis.

6.5 Chapter Summary

This chapter has evaluated the produced framework through a combination of expert review, gate validation and technical assessment against existing DF tools.

Following initial review and subsequent refinement, the framework achieved a 100% acceptance rate of gate thresholds and assessment criteria. This demonstrates that the framework was considered understandable, relevant and suitable for use within DF.

Additional technical testing further demonstrated that the framework could be applied to real-world tools and produce consistent assessment outcomes.

No tested tool met the requirements for full reliance under the framework, with 62.5% of tools failing due to a lack of sufficient automation, while the remaining tools failed due to a lack of transparency, verification or supporting documentation.

This is not surprising as the tools that were tested were not developed with *True Automation* as a concept, but the finding that the framework was applicable to them was the desired result.
Practitioner feedback on the framework showed strong support, with 92% of survey respondents and 73.9% of interview participants stating that they would be willing to solely rely on an automated tool that had successful passed an approved assessment framework.

These findings provide evidence that the framework is both operationally acceptable and technically applicable.

7.0 – Testing & Validation of *Practical Reliability* metric

The testing strategy for the novel assessment metric of PR, described in Chapter 5.3 was designed to ensure the technical accuracy of the proposed metric, as well as it's suitability for real-world analysis of appropriate tools.

This chapter explains the testing that was conducted in relation to the *PR Metric* and the rationale for the testing that was conducted. The same testing methodologies as described in Chapter 6.1 were considered for this purpose, and the method of *Expert Validation* and *Technical Testing* were adopted. Alternatives were considered, such as broader user involvement, and operational testing, however, it is considered that these are less suitable due to the purpose of this stage of testing being to establish foundational validity and applicability, rather than focussing on wider deployment suitability.

As such, there are two approaches, the expert validation conducted constitutes qualitative research and provides meaningful qualitative data in relation to the metric's understandability and usability, and the technical testing that was conducted in quantitative in nature and results in a quantitative analysis of the applicability of the metric to real-world tools and models.

Additionally, the results of this testing are provided, along with an analysis of these results which culminates in suggestions for future work in relation to the *PR Metric* in the context of the thesis.

7.1 Expert Validation

7.1.1 Methodology Justification and Participant Selection

Due to the late addition of this metric into the Framework project, there was limited time and availability of experts to form a complete review panel.

Additionally, with the findings in Chapter 6.3.1 and 6.3.3 that there was a lack of technical expertise in the available respondent pool, it was considered that as this was a mathematical assessment metric the available staff pool would not be able to provide a complete and reliable testing base.

Therefore, a single expert review was decided upon based on the available staff, as described in *Chapter 6.3.2* (and associated subchapters).
The participant that was available, as described in those chapters, has adequate experience in both law enforcement, as well as a previous role within telecommunications.

As a result, it was determined that the participant had sufficient knowledge and experience to conduct the role of the expert reviewer in this case.

7.1.2 Testing Scope

The sole scope of this testing was to obtain the view of the expert in the DF domain to ascertain whether the proposed metric in this case appeared to be suitable for assessment within the law enforcement arena.

As a result, the testing in this case covered two domains:

- How understandable and accessible is the metric?

- How understandable and accessible are the results that would be generated from the metric?

The concept behind the expert review was to assess the complexity of the metric and its results and ensure that it was interpretable in a real-world environment with real world users.

7.1.3 Testing Results

The full result document from the expert review can be found in *Appendix 11* to this thesis.

In summary, the findings of the expert reviewer were as follows:

- Understandability of the Metric

  The expert reviewer found that the metric itself was difficult to understand but accepted that this is due to the lack of mathematical or statistical training held. When the explanation accompanying the metric was explained, the understandability increased and the reviewer was accepting of the use case and rationale behind the combination of scores, as well as the calculation as a whole.

- Clarity of Calculation

  The reviewer accepted the calculation as making operational sense and understood the combination of technical scores to provide a more simplistic all-round response. There were no changes required.
- Ease of Calculation

  The reviewer accepted that, so long as the required information from the developers is available, then the calculation is easy to run and would be suitable for assessors to run within the workplace for the purposes of assessing tools in a real environment.
- Understandability of Output

  The reviewer accepted that, so long as the output of the calculation was as expected, it is easy to understand and interpret with the single figure being useful and the defined range as given in the passing metric assisting greatly in the analysis of the result.
- Overall Ease of Use and Understanding

  There were some concerns raised over the availability of information that would be required for the calculation to be run, and how reliable those figures would be when they were being sent from the developer themselves.

  However, on the whole the reviewer accepted that the calculation was a clear method of assessing complex scores.

  The reviewer pointed out that they would expect that the use of the Framework and the PR metric would require some training to be provided to the assessors, and that this would be needed to address the understandability and usability of the PR metric.

7.1.4 Testing Conclusions

The testing that was conducted within this section was limited, for a variety of reasons that are discussed in *Chapter 8*.

However, the finding of the expert reviewer was that the metric is fit for purpose and is understandable in both execution and result.

There are some areas that are unclear, as documented above, however, this would be an area that could be addressed by training users if and when the Framework was deployed as an assessment system within the organisation.

7.2 Technical Testing

In this section of the Testing conducted, specific tools meeting the automation criteria required of the Framework in *Chapter 6.3.3* were passed through the PR metric that was proposed in *Chapter 5.3*.

The criteria for the selection of these tools were that they had passed the *'True Automation'* gate test defined in *Chapter 6.3.3* and made use of some form of AI or ML in their function.

Public information on the tools was used to provide the information that is required for the calculation, and the results are presented and analysed below.

7.2.1 Testing Method

The method for testing the PR metric is the *Technical Testing* method, this is outlined in more depth within *Chapter 6.3.3*.

The rationale for the choosing of this method for testing this element of the PR metric is that there are defined metrics against which the tools are to be test, that being the PR metric itself, and the expected values being between 0 and 1 as defined in *Chapter 5.3*.

The tools that are to be tested against the metric are defined and technical in nature, with the scores required for testing (F-Score and MCC Score [*Chapter 5.3*]) also being technical in nature.

There is no variation of parameters nor variables proposed, and therefore the experimental method is not suitable in this case. This is a technical test under set and defined conditions.

As a result of this, the *Technical Testing* method was determined to be the most suitable in this case.

7.2.2 Testing Scope

The testing that was conducted in this case is, by virtue of the testing method and information available, limited.

The information scope was archival and secondary only, as it was not possible to conduct real-world testing of the tools to obtain the relevant F-Scores and MCC Scores firsthand.
The reason it is was not possible to access the tools was the time that the DFU was available for testing, and the access that would be required to known and test datasets, which was not available to the researcher.

The scope of the testing itself was singular, to ascertain whether the PR metric calculation, when applied to real-world tools that are currently in use, would result in the expected output in relation to the assessment of said tools.

7.2.3 Testing Results

The results of the testing in this case are documented below, along with the relevant scores and sources for each of the tools that were selected.

The rationale for each tool's selection into the testing is also given.

- *Magnet GriffEye*

No publicly available information exists in relation to either the F-Score or MCC Score in relation to this tool.

The documentation that accompanies GriffEye emphasises the time-saving that can be achieved with the use of the tool but does not provide the specific results for any scores.

- *Magnet Axiom*

No publicly available information exists in relation to either the F-Score or the MCC Score.

- *EnCase*

No publicly available information exists for these metrics.

- *AMPED Software Suite*

No publicly available information exists for these metrics.

As can be seen from the above, there is a distinct lack of information and data that is available for public review in relation to existing and common technical tools. Due to the fact that these metrics were not available for the selected tools, which is explored below in *Chapter 7.2.4* and *Chapter 7.3*, it was not possible to conduct a meaningful running of the *PR Metric* in this case. The data that was missing meant that the calculation could not be reliably performed as key assumptions would have to be made in respect of the models used and the application of those models within the tool, for example the implementation of the models, training and context would not be provided. As such, and to provide a meaningful test set, the search and selection parameters for this test were expanded to include task-specific DF algorithms which were found within academically valid research, rather than commercially available tools.

This is accepted as a strong limitation of this testing and is discussed in *Chapter 8.4.3*.

As the relevant scores are task-dependant, the algorithms that have been selected were specifically selected as they have been used in image manipulation and identification tasks.

All scores used within the calculation, as well as the table used, can be found in (Kong *et al.*, 2024).
This thesis utilises a dataset compiled by *Kong et al.,* 2024 for the testing of models, with the MCC to MCC^1 score calculated as shown in *Chapter 5.3*, this presented a good testing opportunity for the proposed PR Metric in this case whilst preserving equity with the source material, and as such the testing for the PR Metric on each model was also conducted using the scoring results from both datasets.

One of the specific tasks that was involved within the thesis was the identification of manipulated and deepfake images, and therefore the author is satisfied that the task that was given is relevant to the purpose of the Framework and therefore the scores are as relevant as is possible to use in the circumstances.

Once the relevant *PR Metric* score was calculated as per the above, the assessment matrix developed and showcased in *Chapter 5.5* and Table 10 was used to show whether the selected algorithm would have passed the assessment in the Figure 40 framework.

The results of this are shown below in Table 22:

Table 22 - PR Metric Scores (Kong et al., 2024)

Method	**Average F1**	**Average MCC**	**Calculated MCC[1] Score**	**PR Metric Score**	**PR Metric Passed?**
FCN	0.176	0.172	0.586	0.3810	No - 0
U-Net	0.129	0.112	0.556	0.3425	No - 0
DeepLabv3	0.202	0.193	0.597	0.3993	No – 0
MFCN	0.162	0.155	0.578	0.3698	No - 0
RRU-Net	0.139	0.126	0.563	0.3510	No - 0
MantraNet	0.208	0.099	0.550	0.3788	No - 0
HPFCN	0.105	0.086	0.543	0.3240	No - 0
H-LSTM	0.160	0.144	0.572	0.3660	No - 0
SPAN	0.148	0.137	0.569	0.3583	No - 0
ViT-B	0.168	0.158	0.579	0.3735	No - 0
Swin-ViT	0.211	0.204	0.602	0.4065	No - 0
PSCC	0.222	0.148	0.574	0.3980	No - 0
MVSS-Net++	0.284	0.258	0.629	0.4565	No - 0
CAT-NET	0.179	0.104	0.552	0.3655	No - 0
EVP	0.181	0.178	0.589	0.3850	No - 0
TruFor	0.310	0.300	0.650	0.4800	No - 0
PIM	0.333	0.318	0.659	0.4960	No - 0

7.2.4 Testing Summary and Analysis

As can be seen, no testing of the PR Metric could be conducted in relation to the specific tools that were named within the research, the reason for this is that the

named commercial tools do not provide specific information around their internal function. This is supported by other research in the area which points out that DF automation tools are a 'black box' with a lack of information hindering true assessment of function or reliability.

In relation to the testing of the identified models, the Table 22 running showed that the PR Metric is rapid to run, enabling quick assessment of the key F1 and MCC metrics that are otherwise difficult to reconcile and require technical understanding to interpret. This directly speaks to one of the concerns within the DF Domain around the time taken for tasks, although this is a tertiary task to the DF Domain, it demonstrates that the metric seeks to not only increase the understandability of technical information but also seeks to reduce the administrative burden that the assessment and approval of tools places upon the organisation.

In relation to the direct assessment of the metric's function, the above Table 22 data shows that the resulting figures are within the expected range of 0 – 1 that is described within *Chapter 5.3*, and also shows that the figures displayed are within an acceptable variance of the F1 and MCC scores that are used to generate it. This shows that the metric reliably represents the data that is used to generate the score and provides an easier to understand and interpret single figure for assessment, which enables the assessment of technical metrics by non-technical individuals within law enforcement units.

Although a key aspect is that no model tested would achieve the required standard, this is not unexpected in relation to the *PR Metric*, this is a novel metric, and the finding supports the current stance that automated tools require human review prior to use within current evidential work. There are models that are close to moving beyond the lowest level, but there is significant work that remains.

It should also be considered that the models that are being tested are running with a task of detection of image manipulation, a task that is known to be extremely difficult for current models (Verdoliva, L, 2020), and as such the finding that they are not sufficiently capable for *True Automation* reliance is not surprising nor unexpected.

In summary, from the technical testing that has been undertaken in this case, it appears that the metric has succeeded at its stated goals and functions as expected against real-world data.

7.3 – Testing Conclusions and Analysis

In conclusion to the testing that has been undertaken by both expert review and technical testing of the produced PR metric, it has been demonstrated that the novel metric moves significantly towards increasing the non-technical understanding of complex metrics in the F1 and MCC scores.

The expert testing has demonstrated that the metric enables this understanding and is itself more interpretable and accessible than the existing metrics that are available for the assessment of the complex technical tools that are required for modern law enforcement activities.

The technical testing that has been undertaken in respect of the metric also supports that the novel combinatory PR metric effectively represents the F1 and MCC scores, along with their respective benefits and drawbacks that are discussed in *Chapter 5.3* and as such achieves its stated goals of combining the metrics into an easy to create and evaluate singular assessment metric.
This success provides a proof of concept that such a metric is possible to create and would work within a real-world environment.

An additional finding in respect of testing is that the majority of commercial tools that are currently in use within the DF domain do not routinely provide the information that would be required for a proper and thorough assessment of their capability. This, in conjunction with concerns raised earlier in this thesis about the lack of technical understanding existing within DFU departments, that tools are currently being developed and used without a complete understanding of either their function or their reliability, accuracy or capability.

The proposed metric in this case would address this effectively in two ways – it provides for the easy assessment of capability as discussed and also allows for the assessment and review of the full capability of the tool that is being assessed.

This secondary benefit is achieved as the relevant score's requirement intimate knowledge of the models used within tools and how it functions.

As a result of this testing, it can be seen that the proposed metric achieves its stated aims and is effective at addressing key concerns from the DF community.

7.3.1 Future Work

The future work for the testing of the proposed metric in this case would be to vastly increase the scope of the testing beyond what has been done in this case.

Further models, tools and systems could be considered and included within testing for this metric, as well as widening the datasets used to provide a more in-depth review of the effectiveness of the metric, as well as it's applicability to the real-world.

Further experts could also be included within testing for the metric, which would increase visibility of the calculation and allow for a wider view as to the understandability and validity of the metric itself.

The testing that was conducted in this aspect was limited, as has been discussed above, and therefore it is envisaged that any increase in the visibility and volume of testing would be of benefit to this novel metric.

Finally, the testing that was conducted in this thesis was theoretical due to time and resource constraints. This meant that it was reliant upon the secondary research that was conducted and the findings of that research as outlined above. This should be expanded to include primary obtained results from real-world tools against known datasets to improve validity and reliability of the findings of testing.

7.4 Chapter Summary

In summary, the chapter has evaluated the proposed PR Metric through expert review and technical testing.

The expert validation testing showed that, although some explanation was still required, the metric was usable and understandable by non-technical practitioners. It successfully simplified complex performance metrics into a single usable metric whilst maintaining operational effectiveness for assessment.

Technical testing of the metrics demonstrated that the proposed calculation consistently produced results within the expected 0 – 1 range and accurately reflected the underlying F1 and MCC values used in its calculation. Across tested models, scores ranged from 0.3240 to 0.4960, with all results remaining proportionate to the source performance metrics.
The testing additionally highlighted a critical flaw in existing DF tool development, that being the lack of transparent technical performance data, which would be required for meaningful and robust assessment.

8.0 – Project Evaluation & Critical Review

The following chapter shows the critical evaluation of the thesis. It serves to highlight key limitations in all aspects of the thesis, from the methods that were used, the deliverables produced to the testing that was conducted.

The chapter showcases that following the completion of the thesis, the author has reviewed all decisions, processes and procedures that were adopted and used throughout with a critical view to locate areas for development, and improvement, in future.

The chapter also shows that the primary element to the project that should be addressed in future work, and which constitute a key flaw in this thesis, would be that the Framework, PR Metric and gates involved were not fully deployed to real-world cases.
The reason for this is that there are strict restrictions on what can and cannot be applied to live criminal case data, and there are also licence restrictions in relation to the tools that were used which precluded the live use of tools to generate primary data in respect of the capabilities of the tools involved.

Additionally, there were time, resource and availability constraints which precluded the longitudinal and more resource intensive nature of deployed testing.

The chapter then leads into the final chapter of the work, which suggest that now that the first generation of the Framework and PR Metric have been created, additional in-depth testing should be conducted to continue development and ensure that they are as valid, applicable and reliable as possible.

To aid readability and clarity, the additional identified points and discussions have been broken down into sections based upon topic and area within the thesis in the following sections.

8.1 Project Scope

8.1.1 Stakeholders and Police Force

A significant limitation in relation to the scope of the project, as well as the testing that was conducted, was that the focal stakeholders were comprised of primarily a

single UK Home Office Police Force. A key exception to this being the initial focus group discussed in *Chapter 1.3* as this was a separate Police Force, as well as the additional testing that was conducted in *Chapter 8.8*.

The reason that this is the case was that it was necessary due to accessibility and time constraints. The author is employed by this Force and so had access to experts and practitioners within the environment, but this did not extend to other Forces.

The effect of this is that the findings of the primary research that was conducted, is that the findings and input are limited to this one Force, having a potential effect on the universality of the findings and the produced framework, as well as the testing results and operational validity of the findings.

However, it is considered that UK Policing, as demonstrated throughout the thesis but especially within the literature review found in *Chapter 3*, is moving towards a significantly more integrated approach for example with the 2026 Home Office white paper emphasising nationalisation of approach towards data technology and AI within Policing (Home Office, 2026).
Additionally, it is considered that the processes involved in DF work, and the software that is used, is considerably universal already due to UKAS/ISO requirements for forensics requiring specific processes and assurances.

Accordingly, the findings within this research are not claimed to be generalisable across all forces. They are viewed as being applicable to like forces adhering to the same accreditation standards and requirements. As the requirements that are given are generally applicable to all UK Home Office forces, it is considered that the overall effect on generalisability is minimised when applied domestically, however, non-law enforcement organisations and non-UK entities may materially differ and so would require additional research.

8.1.2 Demographic Diversity

Within the research that was conducted, the experts that were consulted were for the most part specifically from DFU and associated areas such as OCAIT.

Whilst this was necessary for the purposes of ensuring the correct level of knowledge, experience and input to the relevant areas, it does mean that the wider areas of Policing such as safeguarding are missed.

These unique perspectives could have assisted in the development of the Framework and increased validity of the gates, metrics and understandability as it is likely that, whilst not technical, they would have had unique insight into the court process and the evidential requirements that they would have as the potential recipient of the end-product from these tools.

However, there were several different considerations and reasons that this was not the approach that was adopted during the course of the project.
Chief amongst these was availability of officers. The survey was initially sent to a wide range of units and areas, with the respondents being used for the later stages of the research, officers from other areas were not necessarily available to partake in this research. Additionally, even with the inclusion of only technical-related departments, there was confusion over some of the terms that were used and it was identified that there was a lack of knowledge around the key areas that were being discussed. It is very likely that, with the inclusion of wider areas of business, this issue would have been compounded.
This would have potentially led to unreliable and skewed data in the results, which would have had an undue impact on the project, both in terms of output and also time taken in analysis.

As a result of these concerns, it is considered that whilst this is a potential flaw which could be explored in future work now that the initial Framework design has been ratified (*Chapters 5 and 6*), it was the most suitable approach at this stage.

8.1.3 Problem Domain

Finally, it is acknowledged that the problem domain – image analysis – is very narrow and limits the work to a narrow field within the wider DF domain.

There are key areas that this affected, including the exclusion of key fields which would benefit from the use of AI/ML such as RAM Analysis and network forensics. Additionally, this narrowed the scope of the Framework and as a result the findings

are narrower than potentially necessary.
Finally, the transferability of the results is now limited as the testing, and development did not consider the other areas of business.

However, the reason that this was the approach that was taken is to ensure that the breadth of requirements placed upon tools and concepts in this area could all be accounted for, and to ensure that the experts available were selected appropriately. There were resource limitations to the available staff, and the project needed to focus on a particular aspect of the DF domain to ensure that the correct level of detail and viability could be assured.

As a result of this, it is considered that there is appropriate mitigation to this within the project, and the reason for this approach outweighs the drawbacks.

8.2 Research Methodology

8.2.1 Lack of Triangulation

The primary method of obtaining information for the development of the Framework in this case was the use of focus groups, interviews and surveys with serving Police Officers and investigators. However, there was no inclusion of developers, independent DF practitioners from outside of the law enforcement domain, legal professionals or ISO accreditation practitioners or assessors.

The result of this lack of inclusion is that there was no input from these separate DF related domains into the creation of the Framework, and therefore the applicability of the Framework to all areas of DF, even with the specific image analysis task, is limited.

It also means that the Framework has not been tested against these adjacent domains for validity and applicability, amongst the other metrics that were tested in this thesis.

However, a strong component to this was the availability of involved persons. The project was significant in length with the participants involved, and the inclusion of further ones would have significantly complicated the process further and introduced delays which could not be appropriately mitigated. Additionally, this would have introduced additional testing requirements which could not be

accommodated with the resources available in this thesis.
The issues that this presents notwithstanding, an additional consideration was that the Framework produced in this thesis is novel, with no existing skeletal structure or existing component to work from. As a result, the produced version is a prototype that would need additional testing as pointed out in *Chapter 6.4.1* and *Chapter 7.3.1*. It was considered that to include so many additional participants at such an early stage of development would introduce complexities and discourse that would actively hinder the development process. As a result, it is preferable to involve a single domain with limited participants at the initial stage of development and then increase this as the project develops and there is more substance to analyse and points to discuss around.

As a result of these considerations, it is considered that the benefits of the adopted approach outweigh the drawbacks to the project and that this was a suitable approach for this stage of the Framework's development.

8.2.2 Researcher Bias

It is a concern that the author of the research in this case is a serving investigator in the Force that has assisted in the development and testing of the Framework in this case.

This is an unavoidable element, as the access that this role granted to experts and technology, as well as knowledge, in the area was vital to the success of the project. However, there were issues that resulted as a consequence of this, such as:

- Unintentional Influence on Responses

The researcher knew the members of staff that were involved in both the surveys that were conducted and the interviews that followed. This means that there was a possibility of the results being skewed, as the respondents could have been reluctant to provide answers that appeared to be contradictory to the aims of the project or the activity. It also meant that during the testing of the Framework, and testing of the PR Metric, that the expert participants may be reluctant to provide negative feedback, to avoid conflict with the researcher.

However, it is considered that there is no evidence of undue influence in the data that was obtained and the results analysed, and the script that was provided shows that there were significant attempts to assure participants at every stage that there was not researcher influence in this case, honest answers were needed. This was colloquially emphasised in conversations around the topic as well, as the interviews and surveys were in action.
Additionally, it was noted and emphasised that all responses to the surveys were anonymous, to avoid this type of problem, and notes on the focus group testing were only obtained in group format with no comments made in respect of any individual in an attempt to avoid this particular problem.

To mitigate this, it would have been possible for the researcher to employ a proxy agent to administer all of these testing methods, with the identity of the researcher being anonymous.
There are identified problems with this approach, as it is also considered that the familiarity of the researcher to the participants mitigated some concerns around how the data would be used, and reassured participants regarding the process, leading to a higher-than-expected response and engagement rate with the testing. However, the mitigations could have been considered further at the time and employed at certain stages such as the interviews when the influence of the researcher was at its peak. This was not duly considered and as such as this opportunity to mitigate influence was missed.
The effect of this on the project could have been skewed results as above, however, there is no evidence that this had an impact and therefore, it is considered that although a missed opportunity, there is no undue effect on the project.

8.2.3 Sample Sizing

The sample size in the survey disseminated in *Chapter 4.2* is 42.

This sample size can be said to be limited, which could have conceivably resulted in non-representative data.

However, although this is a potential critique of the process, this actually represents a response rate of 42% which is higher than average for an organisational survey (Holtom and Holtom, 2008).

The initial pool of potential participants has been justified throughout *Chapters 2, 4 and 5*, this rationale remains the preferred approach in this case as explained within *Chapter 8.2.1*.

As a result, although on the face of it, a 42-respondent survey is limited, this is considered to be a positive position to be in when considering a work-based dissemination, which has traditionally lower response rates.

The limitations of the small sample size in the thesis include that relatively small misunderstandings may skew the data e.g. if a participant misunderstands a question, then this can lead to skewed results.
To mitigate this limited sample size, the sample that was included in the in-depth interviews and surveys was selected to include officers and staff who had specific knowledge and skills in the relevant areas, rather than utilising a less-specific and less-informed sample.

Additionally, several different methods were used at various points, with analysis conducted on each response to identify any significant variation in theme or trend, which would indicate a skew to the data.

8.2.4 Alternative Research Methods

Within the research that was conducted, there were several research approaches considered, which have been discussed within *Chapter 6.1 and 7.1.1*, as well as *Chapter 2*.

However, although the rationale for each decision was made is shown, there were further considerations that were not adequately considered at the time of the research being conducted.
An example of this would be the above-mentioned multi-stakeholder panels involving more departments and DF stakeholders, discussed in *Chapter 8.2.1*, or any consideration of multinational panels which would have again increased the validity of the Framework and testing that has been produced in this thesis.

However, although these are a valid consideration which must be considered at this point in the research, the rationale for the selection of the methods that were used

remains valid, and the concerns and comments made in respect of *Chapter 8.2.1* also remains valid.

Similarly to those arguments, at the time of the research being conducted, the strategies and approaches that were chosen were the most suitable to the research being conducted, with cognisance given to the aim, objectives, availability of resources and time available.

As such, it is considered that although this is a valid concern, this has been considered and discounted in respect of this thesis.

8.3 Framework Design

8.3.1 Dependence

A key aspect of the Framework that became apparent in testing as a potential weakness is that it requires, and relies upon, metrics and information being given freely and reliably as a source of assessment.

This strategy is flawed in that the vendors could, as is currently the case and shown in *Chapter 7.2.3*, to not release the data that is required, or they could change data and misrepresent the actual capability of the tools that they have produced. This is a flaw which the PR Metric seeks to address, as it is considered that it is entirely possible for vendors to use technical language and technical statistical misrepresentation to possibly change how their tool is presented as performing without the need to outright lie about what the results of testing are.

This also places a reliance upon those vendors using valid, tested and approved methods of testing, with approved and valid datasets, tasks and parameters. This is not something that can currently be controlled, and as shown in *Chapter 7.2.3* it is entirely impossible to perform adequate and proper testing of the tool via the Framework gates without this data, unless there was internal testing conducted, which also relies upon the vendor giving adequate access to the proper tool before approval for use, which is not assured. It is also considered that the real-world testing of tools to obtain the required F1 and MCC metrics (amongst others) would not be practical given the stated concerns within the sector around timesaving and efficiencies that are required.

However, it is first considered that the gates that are present, and the information required, have been placed as such based on the requirements of the law enforcement DF practitioners that were involved in the development and testing of the produced Framework. As such, this information is required and it must be accepted by developers that information will be required from them for an appropriate assessment of their tool to take place. The simple fact would be that vendors either refusing to provide this information, or misrepresenting this information, would not be approved and so would lose to competitors who are more open in relation to their tools.

It is accepted that there may be requirements placed upon developers by corporate entities to control the flow of data due to concerns around IP protection or commercial sensitivity, however, as discussed within *Chapter 5.2* such concerns, even when enforced by NDA, cannot be a limiting factor to use nor the providing of information when in the Criminal Justice system. As such, commercial entities will have to accept that, in order to gain the financial and reputational benefits that come with working within the law enforcement domain, comes information transparency requirements which are not trivial, and which must be addressed.

As such, although this is a valid concern, this speaks to the cultural shift that will be required in respect of this domain, as discussed in *Chapter 9*.

Therefore, it is noted that although this concern is valid, it has been addressed to the point that is possible within the bounds of this research.

8.3.2 ISO Requirements

As discussed in *Chapter 3.6*, ISO accreditation via UKAS is a key requirement of any DF process, which is discussed in-depth in *Chapter 3* to this thesis.

However, although this is a key requirement, and something that has only increased in importance since the 2026 Policing White Paper announced the move to potential nationalisation of automation and AI systems, it is not mentioned at any point in the Framework.

Although on the face of it, this is a key omission that must be rectified before the operational deployment of the Framework to any real-world DFU or associated unit, it

must be considered that the accreditation requirement does not apply to each individual process. It is something that, as discussed in the relevant chapters, applies to overall processes within the Units involved, or the tools that are used directly.

As such, direct involvement of UKAS for ISO accreditation of the Framework itself is not considered to be required at this stage, as adoption of the Framework and use within the Unit itself would be part of an overall assessment process which would be the subject of accreditation.

Additionally, it is noted that the ACPO Principles must be considered, as mentioned in *Chapter 3.6* to this thesis.

These principles require, in short, that there is no alteration of the exhibit, a full audit trail is created, there is appropriate examiner oversight and transparency of process for the judiciary.
This is a key consideration for not only the Framework, but the entire direction of both the technical and law enforcement sectors, as the move to '*True Automation*' as explored throughout this thesis would be seemingly contrary to the existing case law, guidance, doctrine and culture around automation analysis. It is no simple thing to remove the *human-in-the-loop* component to analysis which has dominated the early adoption of automated technology, and this is something that will require much deeper legal and technical exploration to address.

As such, although this is a valid concern, it is considered beyond the scope of this thesis to address the breadth of concerns in this case. The Framework that has been designed is intended to be the first step to addressing this potential future concern itself, and acts as a supplementary, rather than replacement, mechanism to ISO accreditation.

8.3.3 Missing Element

On review of the produced gates, metrics and other documentation for the Framework, there is a key potential omission within the considerations that were put forward which must be discussed. The Framework spends considerable time considering and assessing whether the function of a tool is *faster* than that of an

equivalent human investigator completing the same action, under the justification that a tool must be faster than a human to justify the inherent risks of reliance on an automated system.

There is also, in the form of the PR Metric, significant time and consideration given to the assessment of the function of the tool and how well this performs at the given metrics, discussed in *Chapter 5.3*. However, throughout all of these gates there is no proper assessment of whether the tool outperforms a human at the same task.

This is not something that was required by any of the primary research that was conducted in this thesis. It is not clear whether this was a deliberate omission on the part of the participants, or whether this was an oversight.
This is an element that requires further investigation and consideration as to whether this is something that needs to be considered and added to later versions of the Framework.

The method of achieving this would be to revisit some aspects of the primary research that has been conducted and undertake additional investigation with the participants to ascertain why this was not mentioned in the development of the gates.

It is considered that it may be that there is no requirement for the tool to outperform a human, in the view that the time saving is sufficient benefit for the assessment of the tool, and so this is not needed.

There would potentially be additional considerations in that if the tool outperformed a human in both time and ability, there would be concerns around the ability of the human to conduct dip-sampling and some of the verification elements that are required, which would potentially lead to a loss of employment or role in the event that the '*True Automation*' capability was adopted.

This is an element that requires further investigation and consideration as an element of further work.

8.3.4 Interview and Survey Misunderstanding

The development of the Framework relied heavily upon the interviews and survey that were conducted as part of this thesis.

However, on review of the responses and the notes that were made, a significant concern arose.

An example of this is the responses to questions fifteen and sixteen of the survey. Question 15 asked respondents if they would be willing to use an automated tool to conduct analysis, to which only 3 respondents stated no.

Question sixteen then asked, only those who answered no whether the use of a Framework would change their mind – however, 13 answered, with only one of these being a negative response.

This confusion indicates that not every participant to the survey was aware of what was meant by the terms that were being used, nor what was exactly meant by these questions.

Further opportunity could have been given for respondents to clarify points before submitting finalised answers which may have increased validity of the responses.

Additionally, it would have been beneficial to include a further introduction in the form of a video or other medium to explain key concepts and areas to potential respondents which could have again resulted in more reliable results.

However, the inclusion of an introduction video could have introduced skewed results if the explanation was not clear or in the manner that was required for the participants, and additionally if the explanation was too technical this could have dissuaded some from responding.

Additionally, there were resource and time constraints on the project and the researcher's time, which precluded these actions.

The approach that was adopted was the best that could be done at the time and so the researcher is satisfied that the results are as reliable as possible.

8.4 Testing & Validation

8.4.1 PR Metric Testing

The most apparent testing issue in this thesis was the inability to test the PR Metric against some of the most common tools that are in use currently, as shown in *Chapter 7.2.3* and *Chapter 7.2.4*.

The reasons for this are discussed in those chapters, in addition *Chapter 8.3.1* deals extensively with the reliance of the project, both the Framework and the PR Metric, on information being freely available and given in the tools that are being assessed.

However, at this stage of the critical evaluation, it is important to note this again as there would rightly be questions around the applicability of the PR Metric when it appears, on the face of it, that the metric cannot be tested in a real-world scenario and this would naturally affect the acceptance and validity of the Metric as well as the Framework.

As noted in the above-mentioned chapters, this is a serious concern which not only affects this thesis but also every workflow which seeks to assess these tools, and the lack of transparent information must be addressed as a potential cultural issue within the development sector. However, in this particular case, although the tools themselves could not be tested due to the lack of this information, there was ample alternative sources of relevant information as shown in *Chapter 7.2.3* which ensured that the testing could be adequately conducted. This evidences that although information may not always be readily apparent, this does not prevent the Metric from being tested, with successful results as has been shown here.

Therefore, it is considered that although this is a valid concern which does need to be addressed, there is information available which does allow for testing, and so the elements of the project which rely upon this testing are valid and the results of the testing can be utilised.

8.4.2 Focus Group Testing

There were two focus groups that were conducted within this thesis. The primary one was conducted to test and validate the Framework that was produced in this thesis and is displayed in *Figure 40*.

Following reflection in respect of this focus group after it had been conducted, there were several weaknesses which need to be discussed here.

The first of these issues is that the focus group testing was limited in scope because of the sample size (six) which was small. Although the participants were selected to be representative of the groups that had responded to surveys and interviews in

previous elements, the number of these participants at focus group stage was limited.

The reason for this was that there was resource limitations in the Units concerned, with officers not able to be released at the time that the focus group was conducted.

In hindsight, this could have been increased with additional sessions. However, there was no assurance that additional sessions would have allowed for further officers and there were also restrictions in relation to the author's availability. As such, at the time this was the only option available. Additionally, it is considered that although the sample size is limited, the experience of the participants and the diversity of opinion and role as documented in the above chapters, outweighs this to the point that the author is satisfied that the testing of the Framework is appropriate to this stage of development.

Linked to this is a secondary concern, similar to the that pointed out in *Chapter 8.1.2*, regarding the demographics of the conducted focus group. In this case, the focus group that was conducted comprised of law enforcement practitioners from the DFU and OCAIT as discussed. There was no inclusion of wider law enforcement, judiciary, or private sector individuals concerned with the conducting of DF activities. As such, the comments are restricted to only law enforcement opinion and input.

Whilst this is a valid criticism, the rationale for this is the same as shown in *Chapter 8.1.2*, this did not change between the different testing and development stages and so this was again necessary in the circumstances.

However, both of these points should be addressed in future work. Additional testing with a wider scope will be required to further this thesis and the respective fields.

Secondly, at the time of conducting the focus group, it was necessary to record notes only of the discussions and points that were raised. This was necessary due to privacy concerns raised by the participants. Most would not participate in any process involved in this research if it was audio and/or visually recorded. However, this had significant limitations both in terms of the points that were raised and the reliance of external parties on the results of this testing.
The points that were raised were now subject to interpretation from the researcher,

as the notes had to be made live and at the time. As such, there is scope for comments and nuances to be missed as comments and discussion could not be reviewed later.
External parties reviewing this thesis now also need to rely on the notes made and the assessor's work in taking these reliably and not missing anything that was said.

Both of these points mean that the written notes are less reliable than audio/visual recording, and so it would have been significantly preferable. However, given the concerns of the participants, this was the only approach that was possible.

8.4.3 Gate Passing Metric Testing

The Gate Passing Metrics for the gates on the Framework were a late inclusion, based on the researcher's desire to ensure that the Framework was as suitable and understandable as possible.

These metrics were only conceived and implemented at the end of the project, and so the time that was available for testing this was limited, as was the pool of resources and assets that were available.

The best approach that was deemed possible was taken, in the form of the Single Expert Review that is discussed in *Chapter 6.3.2.2*, however, there are significant flaws with this approach in that it relies on the opinion and input of only one person rather than a panel of experts as was used in other areas of testing.

This was mitigated as best as possible through the use of defined testing parameters, narrow scope and reliable expert with relevant experience in multiple fields, however, this was a flaw to the project that may still have had the effect of the results of the testing being less reliable than preferred.

In future work, this should be an area of significant focus to ensure that the metrics that are being used within the Framework are as reliable as possible.

The suggestion would be to conduct panel-based assessment of the gate passing metrics along with the longitudinal real-world deployment testing of the Framework and PR Metric that is suggested at the beginning of this chapter. This could ensure that the testing is as full as possible but could also include additional experts from

other associated fields and disciplines, which would additionally address the triangulation concerns raised in this chapter.

8.5 PR Metric

The PR Metric that has been developed has featured in several of the subchapters involved in this evaluation, chief among the concerns being that it was not tested in either a real-world scenario of using commercially available defined tools, rather the models that are used to complete tasks were tested as these were the elements that had data present.

The issues and considerations that this has have already been discussed in this chapter so will not be discussed further here.

However, there are further concerns that must be considered with the PR Metric, as shown in the following sections:

8.5.1 PR Metric Scoring

The PR Metric is a novel metric that has an intended purpose and has an anticipated and guided expectation and interpretation of the results that it produces, which must be between a 0 and 1 value.

However, the scoring has not been subject to real-world interpretation or analysis by experts in the field. Although there was single-expert review of the metric itself, and the result, in terms of clarity and understandability, it did not consider whether the values assigned to the levels of scoring that can be obtained are correct, nor did it review or consider the abstract interpretation of the scoring was correct, as shown in the Gate Passing Metric.

As a result of this, there is no way on the testing that has been done, to ascertain or interpret whether the scores that are provided as a result of running the metric present a correct assessment of a tool's reliability and accuracy, it only provides whether the numbers come out in the manner and range that was expected when the metric was conceived.

To address this in future work, the metric should be run in real-world tests against known datasets and tools, with a human review of the tool itself, and a comparative

analysis of the tool's function against the metric's output to ensure that there is correlation and that the scoring is being correctly interpreted in the real-world.

8.5.2 Edge Cases

The PR Metric attempts to combine 2 key technical scores, the F1 and MCC scores, into a singular calculation with a singular output which is easier to complete and analyse by non-technical practitioners.

The scores that are involved were carefully selected, as shown in *Chapter 5.3*, however there are some potential issues with the involvement of these metrics into a singular calculation.

The averaging together of these particular metrics can be seen to potentially miss high false negative rates, and mask class imbalance.

However, it is exactly these concerns that led to the inclusion of the MCC (1) score that is considered within the PR Metric itself.

Although the risk of this always remains, it is considered that the benefits of the singular metric for assessment in the first instance,

To address these potential concerns however, the overall impact of these potential missed elements will need to be investigated fully by comparative assessment of the tools against full metrics and then assessment via the new metric, with expert review of the results from both tests to compare and ensure that the substance of the scores are protected by the use of the new calculation.

8.6 Operational Challenges

There are several potential challenges to the operationalisation of the Framework and PR Metric, which were not possible to answer in this thesis due to the nature of the potential challenges.

8.6.1 Internal Resourcing

The Framework, and the PR Metric, are both specifically designed to remove the requirement for in-depth technical knowledge, and also to remove the time constraints around the assessment of novel tools and to encourage a cultural

change towards the acceptance of '*True Automation*' without a human-in-the-loop requirement.

However, although there is evidence from the testing that these elements are addressed within the produced components, there are wider concerns that would not necessarily be dealt with by the creation of these contributions, and even if they were, there are far wider considerations that simply the technical that this thesis has not been able to account for, such as procurement challenges, legal sign-off for tools, national strategy, competing opinion across the various Forces and agencies involved in this kind of analysis and pushback from the development community.

These issues will only be dealt with through the deployment of the Framework to a real-world Unit for testing, and then acceptance from that Unit that the components are for full adoption. At that point, the Force will begin to work through these challenges, which will potentially require the involvement of national entities to review and determine the strategic direction of British Policing in relation to this matter.

However, it is considered that this is actually one of the contributions of this thesis. By producing and providing this potential system of assessment, it seeks to encourage the discussions around the adoption of '*True Automation*' solutions over the existing human-in-the-loop requirements, which as can be seen from *Chapter 3* cannot remain the strategy forever.

8.6.2 Defence and Disclosure

A key issue that has been discussed multiple times throughout the project, and the evaluation, has been the lack of information forthcoming from the developers in respect of key metrics.

Chapter 8.3.1 discussed in depth the effect of refusing to provide this information and the need for change around this if the relevant developer or organisation wishes to enter the criminal justice system in a real way.

However, this also presents a barrier to operationalisation of this thesis and the concepts that are contained. There are multiple legislative requirements around disclosure and what must be declared to the defence, this is already the case but

would become significantly more contentious if a '*True Automation*' approach was adopted.

As such it is currently seen that this lack of information transparency would be a blocker to operationalisation as it is currently not possible to easily comply with disclosure and legal requirements around defence requirements.

However, it is considered that this is something that this thesis itself cannot address, this is something that would only be solved by the cultural shift that has already been discussed in this thesis and is a desired outcome to this thesis.

8.7 Evaluation Summary

In summary to the chapter, the key limitations from the critical evaluation that has been conducted are summarised thus:

- Restricted Scope and Representation
  The research in this thesis has focussed on the Police Force within the UK, and has not included the judiciary, legal teams on both prosecution or defence, nor the developers for the tools that are being made in this domain.
  As such, the findings from the primary research, development of the deliverables, and the conclusions drawn, are limited in scope and generalisability.
- Methodological Constraints
  Although mitigation was employed to include as many stakeholders as possible from within the identified Police Force, the participant numbers were limited and as such stakeholder triangulation was very limited. As a result of this, the generalisability of the findings is limited, and issues such as misunderstood questions and potential bias may have had an effect on the thesis.
- Limited Validation
  The produced deliverables in this thesis, the framework, the PR Metric and the gate thresholds, were not fully tested with independent practitioners or stakeholders, and the testing was not longitudinal in nature, the result of this is that the validation in the thesis can be seen as limited and requiring of further scrutiny.

- Dependence on Vendor Transparency
  Both the framework and the PR Metric in this thesis require the vendors and developers of tools to be clear and transparent with the data on the tools that they are manufacturing. This is a vulnerable point in that if the developers were not honest with the data, then the assessment could be skewed or unreliable, or if the data is not present, as has been found to be the case in the testing that was conducted in this thesis, then neither the framework nor the PR Metric can reliably be completed.
- Lack of focus on adoption over creation
  The thesis, from the initial focus group through to completion, heavily focusses upon the creation of the framework, the PR metric and the testing of these. There is a strong focus on the assessment of tools and how this can assist practitioners in the real-world domain. However, the thesis does not focus upon the organisational adoption of these, nor the logistical and policy requirements that such adoption would have on a forensic unit and the wider organisation.

In summary to this evaluation of the project, it is considered that although there are flaws with the approach and contributions, the most suitable and best available decisions and direction was taken at the time.

There are key identified elements for future work, with this intended to be the first move towards the required shift in the criminal justice system's approach to automation technology.

To summarise, the PR Metric is not yet fully adoptable due to limitations on the data that is released by developers, there are significant cultural barriers to adoption which this thesis alone will not address but seeks to encourage, and there is a need for further validation of the Framework and Metric against real world tools and wider expert practitioners.

8.8 Additional Validation

Following the completion of this thesis, and the undertaking of the evaluation shown in the above *Chapter 8* and associated sub-chapters, it was identified that a strong

and persistent issue with three aspects of this work was the reliance on single-expert review. These three important aspects were:

1. The framework (Figure 40)
2. The gate passing metrics (*Chapter 5.5*)
3. The *PR Metric* (*Chapter 5.3*)

The original rationale for this reliance was a lack of resource and availability of operational staff. However, in the time since that was originally completed, a further opportunity presented in the availability of additional staff members working within the AI Enablement Team and the DFU of the police force previously involved in the testing, as well as a serving academic from the University of Staffordshire, working within the Forensics department.

As all of these stakeholders had relevant experience and knowledge, which would be suitable for reviewing the above elements of this thesis, these experts were provided with a copy of the relevant *Chapter 5* and were asked to comment on all three elements, with a specific ask around the following points:

1. Whether the element is usable in practical terms
   The rationale for this question was that the intention of all deliverables in this thesis that it be usable in the real-world and that practitioners in the field are able to make use of the elements in a practical fashion.

2. Whether the element would be beneficial to either the organisation or the DF domain
   The rationale for the inclusion of this question is that the thesis aims to solve the initial questions around the development of tools which would solve the problems of efficiency and device numbers that are discussed throughout this thesis, and also that the tools assessed by these deliverables would be usable and accepted by the practitioners and organisations within the DF domain.

3. Whether the element is understandable in the form that it was presented
   The rationale for the inclusion of this question is that the produced deliverables, in order to achieve the usability and purpose of easy assessment, need to be

understandable in the form that they are presented.

4. Whether the expert agreed that the element was valid in the given use case
   This was the central reason for the additional expert contact. This question sought to ascertain whether the expert agreed with the element in question and whether they agreed that it met the stated objectives of the thesis, as originally agreed by the single expert.

The intention of all questions was that they would mirror, in a concise manner, the domains and questions that had been asked of the previous expert who had undertaken the single-expert review that the thesis originally relied upon.

The specific results of this ad-hoc additional validation that was conducted is contained in *Appendix 12* to this thesis.
Overall, the results showed that there was significant agreement with the original single-expert review, with the deliverables themselves being accepted and validated by the experts involved in this further work.

The results showed that all of the reviewers agreed with the framework and the *PR Metric.*

The only discrepancy came in respect of the gate passing metrics, where one of the respondents, from inside DFU, stated that they had a different idea as to two of the gate passing metrics.
This is, in and of itself, not surprising and is covered within the guidance of the framework in *Chapter 5.5*, in which it states that the metrics that are provided are given as examples and guidance only. Much wider stakeholder engagement will be required if the metrics are to be set at national level.

However, it is considered that the fact that this discuss and debate can happen at this early stage is a positive sign that the framework, and the gates contained, can be understood and practically applied, even at this conceptual stage. It is a positive indicator that the framework is applicable, with refinement needed to suit specific use cases.

9.0 – Research Conclusions and Recommendations

At the end of the project, the researcher has reviewed the findings and results found as a result of this thesis and compared this to the aim, hypothesis, research questions and objectives to consider whether the project has achieved what it set out to do.

These are discussed in the following sections, ending with an analysis of the project.

9.1 Context and Motivation

As shown throughout *Chapters 1* and *3*, there is a definitive and urgent requirement to increase the effectiveness and efficiency of DF processes and tools. This has been a developing need since as early as 2004 (Rogers and Seigfried, 2004), and has only increased since then with an estimated 90% of crime in the UK now containing a digital element requiring forensic examination (Secretary of State, 2025).

To address this, automation technologies have been included within DF analysis tools, in an attempt to leverage the gains in time and efficiency that automation can provide (Dunsin *et al.*, 2024), however, national policy requires manual verification of any automated process output (College of Policing, 2025) which can be seen to be potentially hindering the manifestation of these benefits in a real-world sense (HMICFRS, 2022).

As such, this thesis has sought to stimulate debate around the use of automation technologies without the need for the current human-in-the-loop process. To achieve this, the project has developed several proof-of-concept contributions which aim to provide a proof-of-concept that an assessment of such tools could be conducted, would be possible and would potentially be accepted by the DF practitioner community.

The artefacts that have been developed have been specifically created to combine practitioner verified requirements and technical assessment mechanisms, with a focus on timesaving, accountability, understandability and clarity.

The Framework that has been developed seeks to showcase the capability of potentially non-technical assessors to undertake complex assessment work in an accountable and verifiable manner whilst maintaining a time-efficient focus. The PR

Metric seeks to bridge the gap between non-technical assessor practitioners and the technical developers of tools to achieve the common goal of increasing the effectiveness and longevity of DF strategies and approaches in the face of evolving and rapidly developing criminality.

As has been shown within the EncroChat case, which was discussed at the beginning of this thesis, novel technologies are requiring novel solutions to solve. These novel solutions mandate a change in policy and culture around the investigation of digital evidence sources, otherwise law enforcement cannot be expected to keep pace and achieve meaningful evidence (Andreja, 2025). It is inevitable that traditional analysis methods cannot keep up with the changing face of technology and as such DF must keep pace. The Framework in this case provides a potential avenue to achieve this.

9.2 Research Aim

The stated aim for this thesis can be found in *Chapter 1.5*.

The aim in this thesis was to develop a novel framework to allow for the assessment of DF tools which utilised automated analysis of image files.

The concept behind this aim was to encourage and stimulate debate and discussion around the use of automated analysis via ML or AI, without the need for human intervention or oversight. This concept was coined as '*True Automation*' in *Chapter 1.2*.

This aim has been met within this thesis. The Framework has been created, in conjunction with multiple experts in the field, as is contained as *Figure 40*.

This aim has been the overarching direction to the entire of the project, with the Framework created specifically designed to address the need for assessment of complex tools in a verifiable and practitioner-informed manner.

From the work that has been conducted in this case, the aim of the research has been met with the creation of the *Figure 40* Framework.

9.3 Hypothesis

The hypothesis in this thesis stated that a framework, grounded in law enforcement requirements, increases investigator confidence and enables automation-based tool use in UK Criminal Justice.

From a review of the findings in this case, the hypothesis is **supported**.

This assertion is based on the following data from this thesis:

- Practitioner respondents to the survey, interviews and focus groups that were conducted throughout this thesis show that acceptance of tools when a framework assessment is completed increases, with significant numbers of respondents showing that they would make use of an automated tool without the need for human intervention should it pass the relevant framework, so long as that framework was accepted for use by the organisation as a whole. 92% of survey respondents and 73.9% of interview respondents stated that the would be willing to make use of an automated tool that passed a framework (*Chapter 6.4*).
This finding also shows that the use of this framework approach increases the number of respondents that would accept a tool after framework when they had originally said that they would not trust a tool, 66.7% (*Chapter 4*). This shows that not only would a framework encourage this change, but it also actively increases confidence.

- Practitioner testing has shown that the *Figure 40* Framework is accepted and valid for implementation into DFU processes and would be suitable for use within the field (*Chapter 6.3*).
Additionally, the Framework has been found to be technically applicable to tools. The *Chapter 6.4* testing showed that although tools are not yet ready for this level of function and reliance, the gates that are included within the were applicable and results could be obtained.

9.4 Research Questions

At the beginning of this thesis, several research questions were posed, which moved towards the aim of the overall project.

These questions are addressed in turn:

1. *What issues are facing the Digital Forensics industry, which may lead to the implementation and development of new automated tools, which can be used to generate court-admissible evidence?*

   There are several elements that have been identified and discussed in this thesis, most notably within *Chapter 3*.
   Primarily, the issues requiring addressing are the need for time-saving, efficiency gains and a reduction in the impactful welfare-affecting tasks such as the viewing CSAM material.
   These issues have been directly built into the *Figure 40* Framework and the PR Metric as key gates that must be passed, and as design elements.

2. *How has Machine Learning been integrated into Digital Forensics?*

   The use of ML and AI within DF tools have been established with *Chapter 3.3*, *3.4* and *6.3.3*.
   The work that has been undertaken shows that there is a significant use of automation technologies within key commercial offerings to the DF sector. However, the testing that was conducted against both the Framework and the PR Metric has shown that there is a key issue in the lack of transparency and internal information from these tools. This issue, as discussed in *Chapter 8.3.1* must be addressed before any meaningful adoption of the concepts in this thesis can be implemented.

3. *What requirements would be placed upon a fully automated Machine Learning based solution, for it to be relied upon by the Criminal Justice sector in the UK?*

   The requirements that would be placed onto an operational tool for it to be relied upon in the way that is required by the '*True Automation*' concept have been captured throughout the primary research that has been conducted in this thesis and shown in *Chapters 5.2*, *5.4* and *5.5*.
   These requirements have been built into the *Figure 40* Framework, the PR

Metric and the layers/planes and gate thresholds included in all artefacts.

4. *Can a framework be developed by which Machine Learning-based tools could be judged to ascertain their applicability to Criminal Justice?*

   Yes – The Framework that has been created and displayed in *Figure 40* has answered this question, with the testing that has been conducted throughout *Chapter 6* showing that this is applicable to the CJ sector and would be accepted by practitioners in the law enforcement field.

9.5 Research Objectives

Within the project there were 4 key objectives to ensure that the work that was conducted answered the questions posed and the aim of the project overall.

These are addressed in turn below:

1. *To investigate state-of-the-art approaches, solutions, and frameworks to identify current issues facing the DF industry, how the issues are being addressed, and how automation/Machine Learning is being used within the Digital Forensic Industry.*

   This objective has been met through the *Chapter 3* systematic literature review, and the subsequent analysis of findings around this topic. Additionally, there is information contained within the testing narrative shown in *Chapters 6* and *7* which showcases how the use of the technologies within the DF domain today.

2. *To determine what requirements would be placed upon a Machine Learning solution to allow for that tool to be relied upon in a criminal justice environment and meet these requirements.*

   This objective has been met through the primary research that has been conducted in this thesis, which has effectively mapped out the requirements that a range of law enforcement practitioners would have for the Framework

that has been developed.
These requirements have been implemented into the *Figure 40* Framework, and the associated guidance document that accompanies the Framework.

3. *To design, develop and implement a framework by which Machine Learning tools could be judged to ascertain their applicability to Criminal Justice in conjunction with real-world practitioners.*

   This objective has been met through the creation of the *Figure 40* Framework and the testing that was conducted throughout *Chapter 6* in relation to that Framework, ensuring that practitioner feedback and requirement was fed into it at all stages.

4. *To evaluate and validate the efficiency and performance of the proposed framework against state-of-the-art solutions.*

   This objective was met through the combined expert and technical testing that was conducted on all produced artefacts in this thesis. The findings of the testing have demonstrated the efficiency and performance of the proposed Framework and has highlighted cultural and commercial changes that are required for the effective analysis and testing of tools in the future.

9.6 Deliverables

Within the project, there were four deliverables that were identified as being required for the effective delivery of the aim, objectives, questions and hypothesis.

These are addressed in turn below:

1. *Results from primary research showing the requirements placed upon Machine Learning-based tools from the Criminal Justice sector.*

   This deliverable has been provided within the project as the results of the survey, interviews and focus groups, which constitute primary research as they were directly obtained during this thesis.

These requirements have then been codified within the *Figure 40* Framework, the PR Metric and the associated documentation that is provided alongside these artefacts.

2. *Analysis through primary & secondary research showing the current performance levels of common MLAs when applied to Digital forensic-based tasks.*

   This deliverable is challenging due to the lack of transparency that has been discovered within commercial tools that are commonly used within the DF domain.
   These tools are shown within *Chapter 3.4* to this thesis, with tools being tested for this within *Chapter 7.2.3*.
   Although this was challenging due to the lack of transparency, alternative methods of finding the results required for this deliverable were discovered in looking at the models themselves, rather than the tools that encapsulate them in the DF domain. Secondary research was conducted to ascertain the relevant F1 and MCC[1] scores for some of the most common MLAs in respect of image analysis and image manipulation detection. Primary research was then undertaken to apply the novel *PR Metric* to ascertain the applicability of the metric and to assess these MLAs against this new standard.
   This effectively obtains the required information in respect of MLA specific information; the deliverable did not state that tool-specific data was required.
   To ensure continuity with the rest of the project, although the deliverable does not specify a task, image analysis was used in the selection of tools and MLAs in this deliverable.
   The finding here is that currently MLAs in respect of DF are not currently performing to the required level for reliance as shown in Table 20.

3. *Conceptual operational framework to judge the applicability of MLA-based solutions to Criminal Justice and provide evidence and reassurance of that applicability if a tool meets all criteria set out.*

This deliverable has been met in the creation of the *Figure 40* Framework contained within *Chapter 5.4*.
The applicability of that tool has been covered in the development of the tool being practitioner led, as outlined in *Chapter 5.2*, and the testing that was conducted with both experts and technical information as shown in *Chapter 6*.
The requirements and input for this framework were obtained from the primary research that was conducted with serving officers and investigators. As such, the operationalised requirements that these participants fed into the thesis are present within the framework that was created and the testing that was conducted on that framework.

4. *Analysis and testing, via primary means where possible, of current automated Digital Forensic tools to see if they meet the developed framework, once complete.*

   This deliverable has been achieved by the testing and development of both the Framework and the PR Metric being practitioner-led to ensure that the developed artefacts were applicable and valid.
   The main element for this deliverable however is the expert and technical testing that was completed in respect of both artefacts, which tested the ability of tools to meet the Framework.
   This testing, and the analysis of such, is contained within both *Chapters 6* and *7*.

9.7 Contributions to Knowledge

Within this thesis, there were 6 contributions to knowledge that were expected.

These are discussed below:

1. *A detailed framework and assessment guidance, by which current and future Machine Learning based Digital Forensic solutions can be judged to ascertain whether they are applicable to the Criminal Justice sector and can be relied upon in Court situations.*

This Framework has been delivered as the *Chapter 5.4 Figure 40* Framework that is the main subject of this thesis. There is also guidance provided for the use of this Framework within *Chapter 5.5* and testing of it conducted in *Chapter 6*.

2. *A novel metric (Practical Reliability [PR]) to assist with the assessment of automated, ML and AI based tools within the DF domain, with the terms defined in True Automation.*

This contribution has been delivered as the *Chapter 5.3* PR Metric which is explained in that chapter and tested within *Chapter 7*.

3. *Coining and definition of a novel term – 'True Automation' – and updated definitions of technical terms in a novel lexicon for cross-domain working between AI, ML and DF.*

   This contribution has been delivered within *Chapter 1.2* to this thesis.
   The impact to the field for this conceptual contribution is additionally explored and explained in Table 2

4. *Findings on the lack of a defined and shared technical lexicon across technical and legal disciplines.*

   This is discussed and explored within the below *Chapters 9.8 and 9.9*.

5. *The results of targeted and comprehensive surveys / interviews to key stakeholders in the law enforcement sector (Chapter 2) pertaining to the requirements and attitudes towards Machine Learning based solutions and the move towards the need to be able to rely on automated analysis.*

These findings are shown and discussed within *Chapter 5.2* to this thesis.

6. *A comprehensive analysis of the current state of Machine Learning based accuracy when applied to a Digital Forensic setting*

   This is contained throughout this thesis, but is especially covered within *Chapter 3.3*, *3.4*, and *3.6*.
   The impact to the field for this conceptual contribution is additionally explored and explained in Table 2

9.8 Analysis

On analysis of the findings in this thesis, it has become apparent that there are several blockers to the adoption of '*True Automation*' that sits at the heart of this thesis.

These include the legal and ethical challenges that are discussed in the above chapters, but also logistical and cultural issues which are extremely difficult to overcome without longitudinal work in this case. It is apparent from the results of both the expert panels and especially the lack of transparency of information coming from developers of key tools, that a collaborative approach is needed moving forward, between both the DF practitioners and the software developers for the tools that are needed.
The issues that have been presented in relation to both sides can only be overcome by collaboration and understanding the competing demands that affect trust in tools.

A key challenge that was identified was a lack of technical experience and knowledge in the specific area of AI and ML on the part of the practitioner assessors who would be making use of the *Figure 40* Framework and the PR Metric.
This lack of experience led to confusion on the part of some participants in respect of the language that was used.
A colloquial finding was that terms such as accuracy were often misconstrued and needed to be clarified, although in most cases confusion remained.

This led to the conclusion that a key method for encouraging the above-mentioned collaboration would be an agreement of a shared lexicon between the two domains, as terms such as accuracy differing meanings between the technical and non-technical persons involved.

Additionally, this collaborative approach would enable the required testing of software to be conducted before contracts and agreements were signed, with issues only being discovered after the fact, leading to tension and a lack of trust.
If software were to be made available earlier with the agreement of both sides that it was a testing environment only, then trust could be built and issues addressed quickly, enabling benefits to be seen on both sides.
As it stands, the current asymmetrical development and assessment approach taken by both software developers and DF practitioners leads to a vital lack of information which precludes most tools from being properly assessed and hence relied upon to the extent that is required for a '*True Automation*' approach. An example of this and the impact that it can have has been seen in this thesis, with the PR Metric testing being limited due to a lack of information being released by tool developers.

That being said, it can also be seen from the testing that was undertaken that although there are automation features present in modern tools, and that AI may form a part of this, these are not sufficient for the level of autonomous decision making that would be required to fully address the above. Therefore, although this collaborative approach is required at this point to encourage discussion and longitudinal cultural change, it is also the case that '*True Automation*' requires technology to also advance further.

9.9 Recommendations

Future work has been addressed throughout this thesis and is mentioned in specific chapters.

One of the key elements which must be considered is the inclusion of key stakeholders from other areas of CJ such as the judiciary, prosecution and defence legal experts. This would allow for the widening of the scope of the framework and strengthen the claims of applicability of the produced artefacts in this thesis. This

inclusion would take the form of focus groups, expert validation and input as well as real-world testing as has been conducted in this thesis with law enforcement.

However, this thesis has provided a proof-of-concept for both an assessment metric and framework, designed with practitioners in the field, which shows that it is possible to appropriately test tools against the requirements of the DF domain and that as a result it would be possible to develop a tool which did meet those requirements, once automation technology itself advanced to a sufficient level.

There are key flaws in the project, which require more in-depth study and work to overcome and ensure that the metric and framework are as complete and suitable as possible.

The first step in this area is suggested to be opening full and open dialogue between developers and the DF domain to begin approaching the differing requirements.

This collaboration between development and DF practitioners is required to continue the discussion around this important area.

In practical terms, this collaborative effort could take the form of technical assurance system, for developers to provide DF Practitioners a minimalised version of the software for assessment purposes, rather than expecting developers to fully release their software without adequate commercial protection.

In response, the DF sector could provide a copy of the Framework and metrics in this case, as well as being clearer with developers around the disclosure requirements around their tools and function, so that informed decisions can be made at every stage and level.

The project has found that the above approach is possible, and that a meaningful assessment can be conducted which would encourage discussion around the use of tools without the current human-in-the-loop approach. The primary research also finds that if such a Framework were utilised and the challenges could be overcome, staff within the DF domain would be willing to rely upon the tools that are tested to the extent of '*True Automation*' so long as there were organisational acceptance of the Framework and its findings. As a result, the continuation of this discussion and

work only increases with importance as the DF issues around numbers of devices and time taken increases and the capabilities of AI models also increases.

In addition to the above, organisations undertaking to accept the framework and deploy this operationally would need to create a policy and procedure for the use, define acceptable parameters and clearly define roles and responsibilities for the human users and assessors.

Also, this dialogue is the main method by which the ethical and legal considerations described in *Chapter 3.7* can be addressed. These challenges can only be overcome by a collaborative and structured approach to changing the cultural and assumptive attitudes which act as a potential roadblock to the adoption of '*True Automation*' in a real sense.

The benefits for both sides can only be realised with this collaborative approach, as emphasised within the Information Commissioner's Office (ICO) AI Strategy, which emphasises the need for generative AI model developers to work with public authorities to ensure public trust and legislative compliance (Information Commissioner's Office, 2025)

It is no longer the case that the two domains can sit in isolation, and that real-world practitioners can no longer rely on a lack of knowledge and understanding. The issues that have led to this thesis have been ongoing and increasing for over 20 years, and the influence of digital devices and capabilities is ever-present in every aspect of daily life.

Additionally, to the above, a shared lexicon should be developed to agree key term definitions for elements such as accuracy, precision, recall, classes, positive and negatives as well as positive and negative. This will help to avoid confusion in the dialogue undertaken and ensure that both technical and non-technical individuals can communicate on equal footing, ensuring clarity of communication and assessment.

Finally, the establishing of a Core Assessor Role within the DF domain would be greatly beneficial, as this would enable the development of a minimum standard of training and understanding in the AI and ML fields and ensure that the staff

undertaking the assessment are properly trained and educated to the extent that is required by the expert feedback obtained in respect of the PR Metric and Framework. Processes would also need to be adapted in the same vein, to ensure an ongoing and continuous assessment process that is required to capture tool and technology updates.

In summary, the criminal justice system must accept that it needs to adapt its standards and requirements in the face of this ever-growing problem and accept that a human review cannot be the requirement for every piece of analysis as there is quite simply too much.

Technology may provide the solution to this problem, but the only means by which the benefits can be fully realised would be to begin the process of adapting criminal justice to work with technology and developers, rather than viewing it as a 'black box' as has been the case until this point.

9.10 Future Work

For future work to continue this thesis, there are several aspects which require further investigation, correction and addressing before the produced deliverables, especially the *True Automation* term, titular framework and the *PR Metric* can be reliably used, or considered for use, within CJ.

The most significant aspects are given below:

- Stakeholder Engagement

One of the most significant further works that is required for all three primary deliverables is the need for additional stakeholder engagement.
This thesis has presented an initial prototype to show that from a law enforcement practitioner level, the concept of assessing automated tools and moving towards removal of the *human-in-the-loop* safeguards is possible and would be accepted by practitioners in the right circumstances. As shown above, this has been achieved and the responses show that should this be adopted then practitioners are open to the idea of safeguard removal.

However, as has been discussed in preceding chapters, this has only considered one stakeholder group in the involvement of police at this stage. The DF domain and the

UK CJ system is vast and includes significant and diverse stakeholder groups. As such, further consultation should be conducted with additional stakeholder groups such as the judiciary, legal experts from the prosecution and defence, barristers and also technical developers and organisations to ensure that their requirements, considerations and concerns are also included within the produced deliverables, ensuring that the work is as general as possible to the field, and that the potential for acceptance of those deliverables within the wider CJ system and development domain is as maximised as possible.

- Testing

As has been outlined in the above chapters, the testing that has been conducted in this thesis requires additional input. Although this is as comprehensive as possible given the constraints of this project, the deliverables produced require significant further testing and validation before they can be considered as truly reliable or valid.

This increased and further testing should not only consider the existing testing strategies but should also pull upon the above point regarding additional stakeholders. Presently, the testing has been confined to police input, police expert review and technical testing on only the available data and tools. This should be expanded significantly to capture additional stakeholder requirements and testing scenarios, as well as include the further access, capability and skill that would be forthcoming from the inclusion of these wider groups.

A key example is the *PR Metric*. The testing for this metric found that there is a significant amount of data that is not present from the developers. This meant that the testing could not focus on the tools themselves but had to be done against models and algorithms instead. This is flawed and should be addressed with developers present, so that the testing of the metric can be conducted comprehensively.
Additionally, the metric should be discussed with wider diverse stakeholders to ensure that the understandability claims on which the metric is based is generally applicable.

- Gate Threshold Standardisation

Chapter 5.5 discusses the assessment of gate metrics and the assessment of those gates. It also provides a suggested level which would need to be achieved at a minimum for the gate to be seen as passed. As can be seen in *Chapter 8.8* these levels are not always agreed by experts in the field. As such, additional input should be sought from wider stakeholders in all aspects, including those discussed in this chapter, to agree a technical, organisational and legally acceptable level against which the tools could be assessed. This would allow for formal standards, which are acceptable domain-wide to be formalised, and would also allow for a wider assessment of whether the levels that are acceptable should vary depending on tool factors such as use case, case level, and risk.

This would significantly improve the credibility of the produced framework, as well as increase the acceptability of the tools that are assessed to have passed the framework across the domain.

- Policy and Organisational Acceptance

A significant amount of work will be required to move from the existing model of human verification to the new concept of *True Automation*. A large part of this will be organisational in that DF organisations, both law enforcement and forensic, would need to establish, ratify and agree policy and procedure in relation to the adoption of the framework in this thesis, as well as the tools that pass that framework. These policies and procedures would need to detail the different responsibilities of individuals including assessors and users and would need to codify the lines of accountability involved in the use of automated tools once they have passed the assessment framework.
Further, the policies would need to include the provisions for ongoing and continuous assessment to ensure that the assessment of tools is continuous, and that there is a mechanism for continuous monitoring.

The creation and inclusion of these policies and procedures would move further towards domain-wide acceptance of *True Automation* and the guided assessment of such tools and would also result in wider acceptance of the assessed tools by practitioners by codifying the organisational acceptance of assessed tools.

9.11 Final Word

In concluding this thesis, the original objectives, questions and deliverables have been answered, as shown in *Chapter 9*.

The overarching point of this thesis was to explore whether it was possible to move towards *'True Automation'* within DF.

When reviewing the results that have been obtained, it is easy to conclude that this is not the case, and that the lack of information and capability, as well as the cultural and legal issues that have been discussed would preclude this.

However, it is important to recall that the framework created within Figure 40 shows a pathway, not the conclusion. The response from practitioners in the field was that it would be possible, if there was appropriate organisational oversight.

This thesis has shown that this move is *possible,* with the appropriate cultural, technological and inter-domain changes that are required.

10.0 References

Abdi, A.H. *et al.* (2024) 'Security Control and Data Planes of SDN: A Comprehensive Review of Traditional, AI, and MTD Approaches to Security Solutions', *IEEE Access*, 12. Available at: https://doi.org/10.1109/ACCESS.2024.3393548.

ACPO (2006) 'Murder Investigation Manual'. Centrex. Available at: https://www.npcc.police.uk/documents/FoI%20publication/Disclosure%20Logs/Crime%20%20FOI/2011/073%2011%20Att%2001%20of%201%20Murder%20Investigation%20Manual.pdf.

ACPO (2012) 'Core Investigative Doctrine'. National Policing Improvement Agency.

Adeoye-Olatunde, O.A. and Olenik, N.L. (2021) 'Research and scholarly methods: Semi-structured interviews', *JACCP: JOURNAL OF THE AMERICAN COLLEGE OF CLINICAL PHARMACY*, 4(10), pp. 1358–1367. Available at: https://doi.org/10.1002/jac5.1441.

Al Fahdi, M. *et al.* (2016) 'A suspect-oriented intelligent and automated computer forensic analysis', *Digital Investigation*, 18, pp. 65–76. Available at: https://doi.org/10.1016/j.diin.2016.08.001.

AlMutawa, A., Ikuesan, R.A. and Said, H. (2024) 'Towards a Comprehensive Metaverse Forensic Framework Based on Technology Task Fit Model', *Future Internet*, 16(12), p. 437. Available at: https://doi.org/10.3390/fi16120437.

Amarullah, A., Ramll, K. and Regina, O. (2025) '(PDF) Illicit Cryotocurrency Investigation Digital Forensic Framework: Integrating Off-chain and On-Chain Analysis for Two Types of Crime', *International Journal of Electrical Computer and Biomedical Engineering*, 3(2). Available at: https://doi.org/10.62146/ijecbe.v3i2.135.

Amped Software (2026) *Amped Authenticate | Photo, Video, and Deepfake Forensics*, *Amped Authenticate*. Available at: https://ampedsoftware.com/authenticate (Accessed: 21 February 2026).

Andersen, D.B., Sunde, N. and Porter, K. (2025) 'Tool induced biases? Misleading data presentation as a biasing source in digital forensic analysis', *Forensic Science International: Digital Investigation*, 52, p. 301881. Available at: https://doi.org/10.1016/j.fsidi.2025.301881.

Andreja, M. (2025) 'Privacy in the Digital Criminal Landscape: redefining legal frameworks in the era of digital evidence', *Revue Lexsociété* [Preprint]. Available at: https://hal.science/hal-05296970 (Accessed: 14 February 2026).

Autopsy (2018) *Autopsy User Documentation: Ingest Modules*. Available at: https://sleuthkit.org/autopsy/docs/user-docs/4.5.0/ingest_page.html (Accessed: 28 January 2026).

Ayers, R., Brothers, S. and Jansen, W. (2014) *Guidelines on Mobile Device Forensics*. NIST Special Publication (SP) 800-101 Rev. 1. National Institute of Standards and Technology. Available at: https://doi.org/10.6028/NIST.SP.800-101r1.

Bachtiar, N.K., Fariz, M. and Arif, M. (2024) 'Conducting a Focus Group Discussion in Qualitative Research', *Innovation, Technology, and Entrepreneurship Journal*, 1, pp. 94–101. Available at: https://doi.org/10.31603/itej.11466.

Bentson, C. (2024) 'Cell Phones and Toothbrushes':, *The Progressive Orthodontist*, Business&Practice Development. Available at: https://bcp-advisors.com/wp-content/uploads/2024/05/Cell-Phones-and-Toothbrushes.pdf.

Bérubé, M. *et al.* (2025) 'From digital trace to evidence: Challenges and insights from a trial case study', *Science & Justice*, 65(5), p. 101306. Available at: https://doi.org/10.1016/j.scijus.2025.101306.

Bhatt, P. and Rughani, P.H. (2017) 'MACHINE LEARNING FORENSICS:A NEW BRANCH OF DIGITAL FORENSICS', *International Journal of Advanced Research in Computer Science*, 8, pp. 217–222. Available at: https://doi.org/10.26483/ijarcs.v8i8.4613.

Bianconi, F. (2024) 'Gantt Charts', *Data and Process Visualisation for Graphic Communication: A Hands-on Approach with Python*. Cham: Springer Nature Switzerland, pp. 175–181. Available at: https://doi.org/10.1007/978-3-031-57051-3_13.

Bollé, T., Casey, E. and Jacquet, M. (2020) 'The role of evaluations in reaching decisions using automated systems supporting forensic analysis', *Forensic Science International: Digital Investigation*, 34, p. 301016. Available at: https://doi.org/10.1016/j.fsidi.2020.301016.

Breitinger, F. *et al.* (2024) 'DFRWS EU 10-year review and future directions in Digital Forensic Research', *Forensic Science International: Digital Investigation*, 48, p. 301685. Available at: https://doi.org/10.1016/j.fsidi.2023.301685.

Butterfield, E.M. *et al.* (2018) 'Automated Digital Forensics'. Leeds Beckett University. Available at: http://eprints.leedsbeckett.ac.uk/5073/1/Automat.

Cambridge Dictionary (2022) *Framework - Cambridge Dictionary*. Available at: https://dictionary.cambridge.org/dictionary/english/framework (Accessed: 29 August 2022).

Campbell, A. (2011) *The Fingerprint Inquiry report*. Edinburgh: Published on behalf of the Fingerprint Inquiry by APS Group Scotland. Available at: https://www.thefingerprintinquiryscotland.org.uk/inquiry/files/TheFingerprintInquiryReport_Low_res.pdf?utm_source=chatgpt.com.

Cao, C., Chicco, D. and Hoffman, M.M. (2020) 'The MCC-F1 curve: a performance evaluation technique for binary classification'. arXiv. Available at: https://doi.org/10.48550/arXiv.2006.11278.

Cheshire Constabulary (2025) 'Cheshire Constabulary's IT and Digital Strategy 2025 - 2029'. Cheshire Police. Available at: https://www.cheshire.police.uk/SysSiteAssets/foi-media/cheshire/priorities-and-how-we-are-doing/it-and-digital-strategy-2025-29.pdf.

Chicco, D. and Jurman, G. (2023) 'A statistical comparison between Matthews correlation coefficient (MCC), prevalence threshold, and Fowlkes–Mallows index', *Journal of Biomedical Informatics*, 144, p. 104426. Available at: https://doi.org/10.1016/j.jbi.2023.104426.

Claij-Swart, A.A. *et al.* (2025) 'Automatically generating digital forensic reference data triggered by mobile application updates', *Forensic Science International: Digital Investigation*, 54, p. 301985. Available at: https://doi.org/10.1016/j.fsidi.2025.301985.

Cohen, A. and Nissim, N. (2018) 'Trusted detection of ransomware in a private cloud using machine learning methods leveraging meta-features from volatile memory', *Expert Systems*

with Applications, 102, pp. 158–178. Available at: https://doi.org/10.1016/j.eswa.2018.02.039.

College of Policing (2015) 'MCCT-2'. College of Policing.

College of Policing (2025) *Deciding when to use AI | College of Policing*, *Deciding when to use AI*. College of Policing. Available at: https://www.college.police.uk/guidance/building-ai-enabled-tools-and-systems/deciding-when-use-ai (Accessed: 8 February 2026).

CPS (2025) *How the CPS will use Artificial Intelligence | The Crown Prosecution Service*. Available at: https://www.cps.gov.uk/how-cps-will-use-artificial-intelligence (Accessed: 21 February 2026).

Crown Prosecution Service (2023) *Disclosure of Material to Third Parties | The Crown Prosecution Service*, *Disclosure of Material to Third Parties*. Available at: https://www.cps.gov.uk/legal-guidance/disclosure-material-third-parties (Accessed: 15 August 2025).

Dardick, G.S. and Lau, L.K. (2005) 'Interdisciplinary minor in digital forensics, security and law', *Proceedings of the 6th conference on Information technology education*. New York, NY, USA: Association for Computing Machinery (SIGITE '05), p. 371. Available at: https://doi.org/10.1145/1095714.1095797.

Du, X., Le-Khac, N.-A. and Scanlon, M. (2017) 'Evaluation of Digital Forensic Process Models with Respect to Digital Forensics as a Service', *arXiv:1708.01730 [cs]* [Preprint]. Available at: http://arxiv.org/abs/1708.01730 (Accessed: 22 June 2021).

Duarte (2025) *Alarming Average Screen Time Statistics (2025)*, *Exploding Topics*. Available at: https://explodingtopics.com/blog/screen-time-stats (Accessed: 20 July 2025).

Dunsin, D. *et al.* (2024) 'A comprehensive analysis of the role of artificial intelligence and machine learning in modern digital forensics and incident response', *Forensic Science International: Digital Investigation*, 48, p. 301675. Available at: https://doi.org/10.1016/j.fsidi.2023.301675.

Elsaesser, C. and Tanner, M.C. (2001) 'Automated Diagnosis for Computer Forensics', *Technical Report, The Mitre Corporation*, pp. 1–16.

EMSOU (2025) *Police pilot advanced data platform to tackle serious and organised crime*, *NPCC SOC Portfolio*. ROCU. Available at: https://www.rocu.police.uk/news/2025/july/police-pilot-advanced-data-platform-to-tackle-serious-and-organised-crime/ (Accessed: 21 February 2026).

Escalante-Viteri, A. *et al.* (2025) 'Artificial Intelligence in Software Testing: A Systematic Review of a Decade of Evolution and Taxonomy', *Algorithms*, 18(11). Available at: https://doi.org/10.3390/a18110717.

EU (2024) *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*. Available at: http://data.europa.eu/eli/reg/2024/1689/oj (Accessed: 28 February 2026).

Fadill, A. and Rochmadi, T. (2025) 'Developing a Delphi Validated Instrument for Assessing Digital Forensics Readiness Based on COBIT 2019', *International Journal of Advances in Data and Information Systems*, 6(3), pp. 749–762. Available at: https://doi.org/10.59395/ijadis.v6i3.1453.

Falcón-López, S.A. *et al.* (2025) 'Forensic Analysis of Manipulated Images and Videos', *Applied Sciences*, 15(23), p. 12664. Available at: https://doi.org/10.3390/app152312664.

Fish, L. *et al.* (2026) 'Children of the 2020s: home learning environment and screen time at age 2: research brief'. Available at: https://doi.org/978-1-83870-739-2.

Forensic Science Regulator (2025) 'Draft Code of Practice 2025 (Version 2)'. Home Office. Available at: https://assets.publishing.service.gov.uk/media/67daba1e594182179fe0883b/E03313596%2B-%2BCoP%2BForensic%2BScience%2BRegulator%2B2025_A4_v02_Web%2BAccessible.pdf.

Garfinkel, S. *et al.* (2009) 'Bringing science to digital forensics with standardized forensic corpora', *Digital Investigation*, 6, pp. S2–S11. Available at: https://doi.org/10.1016/j.diin.2009.06.016.

Geetha, R. and Thilagam, T. (2021) 'A Review on the Effectiveness of Machine Learning and Deep Learning Algorithms for Cyber Security', *Archives of Computational Methods in Engineering*, 28(4), pp. 2861–2879. Available at: https://doi.org/10.1007/s11831-020-09478-2.

Gentry, E. and Soltys, M. (2019) 'SEAKER: A mobile digital forensics triage device', *Procedia Computer Science*, 159, pp. 1652–1661. Available at: https://doi.org/10.1016/j.procs.2019.09.335.

Göbel, T., Breitinger, F. and Baier, H. (2025) 'Optimising data set creation in the cybersecurity landscape with a special focus on digital forensics: Principles, characteristics, and use cases', *Forensic Science International: Digital Investigation*, 52, p. 301882. Available at: https://doi.org/10.1016/j.fsidi.2025.301882.

Goundar, S. (2012) 'Research Methodology and Research Method - Methods Commonly Used by Researchers'. Victoria University of Wellington. Available at: https://prallagon.com/wp-content/uploads/2021/05/Research-Methodology-A2.pdf.

GriffEye (2020) 'Multi-Object Detection - Now in GriffEye Brain'. Available at: https://www.griffeye.com/multi-object-detection-now-in-griffeye-brain/.

Grossman, M. and Grimm, P. (2025) 'Judicial Approaches to Acknowledged and Unacknowledged AI-Generated Evidence', *Science and Technology Law Review*, 26(2). Available at: https://doi.org/10.52214/stlr.v26i2.13890.

Guest, G. *et al.* (2020) 'Comparing Interview and Focus Group Data Collected in Person and Online', *Patient-Centered Outcomes Research Institute (PCORI)*. Washington (DC). Available at: https://europepmc.org/article/NBK/nbk588708#free-full-text.

Hafdi, Z.S. and El Kafhali, S. (2025) 'A Comparative Evaluation of Machine Learning Methods for Predicting Student Outcomes in Coding Courses', *AppliedMath*, 5(2), p. 75. Available at: https://doi.org/10.3390/appliedmath5020075.

Hall, S.W., Sakzad, A. and Choo, K.-K.R. (2022) 'Explainable artificial intelligence for digital forensics', *WIREs Forensic Science*, 4(2), p. e1434. Available at: https://doi.org/10.1002/wfs2.1434.

Hargreaves, C., van Beek, H. and Casey, E. (2025) 'SOLVE-IT: A proposed digital forensic knowledge base inspired by MITRE ATT&CK', *Forensic Science International: Digital Investigation*, 52, p. 301864. Available at: https://doi.org/10.1016/j.fsidi.2025.301864.

Hasan, R., Mahmood, S. and Hasan, M.A. (2011) 'Artificial Intelligence Based Model for Incident Response', *2011 International Conference on Information Management, Innovation Management and Industrial Engineering*, pp. 91–93.

Hawkins, D. (2025) 'Deepfake false evidence & the justice system'. *ACE / OPCSA Deepfake Research Briefing Event*, London, November.

HMICFRS (2022) 'Digital forensics: An inspection into how well the police and other agencies use digital forensics in their investigations'. Available at: https://www.gloucestershire-pcc.gov.uk/SysSiteAssets/key-information/reports/external-reports/hmicfrs-reports/hmicfrs-digital-forensics-an-inspection-into-how-well-the-police-and-other-agencies-use-digital-forensics-in-their-investigations.

Holt, T.J., Bossler, A.M. and Seigfried-Spellar, K.C. (2022) *Cybercrime and Digital Forensics: An Introduction*. 3rd edn. London: Routledge. Available at: https://doi.org/10.4324/9780429343223.

Holtom, Y. and Holtom, B. (2008) 'Survey response rate levels and trends in organizational research - Yehuda Baruch, Brooks C. Holtom, 2008', *Human Relations* [Preprint]. Available at: https://journals.sagepub.com/doi/10.1177/0018726708094863 (Accessed: 21 February 2026).

Home Office (2022) 'Extraction of Information from electronic devices: code of practice'. OGL.

Home Office (2026) 'From Local to National: A New Model for Policing'. Available at: https://assets.publishing.service.gov.uk/media/69779267276692606c013862/260125_White_Paper.pdf.

Horsman, G. (2022) 'Forming an investigative opinion in digital forensics', *WIREs Forensic Science*, n/a(n/a), p. e1460. Available at: https://doi.org/10.1002/wfs2.1460.

Hughes, N. and Karabiyik, U. (2020) 'Towards reliable digital forensics investigations through measurement science', *WIREs Forensic Science*, 2(4), p. e1367. Available at: https://doi.org/10.1002/wfs2.1367.

Hunton, P. (2011) 'A rigorous approach to formalising the technical investigation stages of cybercrime and criminality within a UK law enforcement environment', *Digital Investigation: The International Journal of Digital Forensics & Incident Response*, 7(3–4), pp. 105–113. Available at: https://doi.org/10.1016/j.diin.2011.01.002.

ICO (2025) *Right not to be subject to automated decision-making*. ICO. Available at: https://ico.org.uk/for-organisations/law-enforcement/guide-to-le-processing/individual-rights/right-not-to-be-subject-to-automated-decision-making/ (Accessed: 21 February 2026).

Information Commissioner's Office (2025) *Preventing harm, promoting trust: our AI and biometrics strategy*. ICO. Available at: https://ico.org.uk/about-the-ico/our-information/our-strategies-and-plans/artificial-intelligence-and-biometrics-strategy/ (Accessed: 28 February 2026).

ISO/IEC 42001:2023 (2023) *ISO*. Available at: https://www.iso.org/standard/42001 (Accessed: 28 February 2026).

Isselé, Q., Heinze, A. and Fenton, A. (2024) 'The combined power of qualitative and quantitative insights', *Digital and Social Media Marketing*. 3rd edn. Routledge.

Jarrett, A. and Choo, K.-K.R. (2021) 'The impact of automation and artificial intelligence on digital forensics', *WIREs Forensic Science*, 3(6), p. e1418. Available at: https://doi.org/10.1002/wfs2.1418.

Kapetanidou, I.A. *et al.* (2025) 'Cognitive Computing Continuum: State-of-the-Art Review and ENACT Vision & Approach', *Journal of Grid Computing*, 23(3), p. 23. Available at: https://doi.org/10.1007/s10723-025-09810-9.

Kävrestad, J. (2022) 'Context-Based Micro-Training: Enhancing cybersecrity training for end-users'. University of Skövde 2022, Sweden. Available at: https://www.diva-portal.org/smash/get/diva2%3A1806953/FULLTEXT01.pdf.

Khalid, Z., Iqbal, F. and Fung, B.C.M. (2024) 'Towards a unified XAI-based framework for digital forensic investigations', *Forensic Science International: Digital Investigation*, 50, p. 301806. Available at: https://doi.org/10.1016/j.fsidi.2024.301806.

Khalil, M. and Er, E. (2023) 'Will ChatGPT get you caught? Rethinking of Plagiarism Detection'. arXiv. Available at: https://doi.org/10.48550/arXiv.2302.04335.

Kim, D. and Lee, S. (2020) 'Study of identifying and managing the potential evidence for effective Android forensics', *Forensic Science International: Digital Investigation*, 33, p. 200897. Available at: https://doi.org/10.1016/j.fsidi.2019.200897.

Kim, J. *et al.* (2025) 'Your forensic AI-assistant, SERENA: Systematic extraction and reconstruction for enhanced A2P message forensics', *Forensic Science International: Digital Investigation*, 53, p. 301931. Available at: https://doi.org/10.1016/j.fsidi.2025.301931.

Kong, C. *et al.* (2024) *Pixel-Inconsistency Modeling for Image Manipulation Localization*, *Pixel-Inconsistency Modeling for Image Manipulation Localization*. ArXiv. Available at: https://arxiv.org/html/2310.00234v2 (Accessed: 4 February 2026).

Lewis-Lim, S. *et al.* (2025) 'Analysing Chain of Thought Dynamics: Active Guidance or Unfaithful Post-hoc Rationalisation?', in C. Christodoulopoulos et al. (eds) *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. *EMNLP 2025*, Suzhou, China: Association for Computational Linguistics, pp. 29838–29853. Available at: https://doi.org/10.18653/v1/2025.emnlp-main.1516.

Luciano, L. *et al.* (2018) 'Digital Forensics in the Next Five Years', *ARES 2018: Proceedings of the 13th International Conference on Availability, Reliability and Security*. New York, NY, USA: Association for Computing Machinery (ARES 2018), pp. 1–14. Available at: https://doi.org/10.1145/3230833.3232813.

Mak, M.-K. and Luo, T. (2025) 'A framework for evaluating cultural bias and historical misconceptions in LLMs outputs', *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 5(3), p. 100235. Available at: https://doi.org/10.1016/j.tbench.2025.100235.

Malandruccolo, M. (2024) 'Harnessing the Power of Machine Learning and AI to Decrease Health Expenditures Worldwide - A Review - Premier Science', *Premier Journal of Science 2024* [Preprint]. Available at: https://doi.org/10.70389/PJS.100004.

Martinez, M.N. and Bartholomew, M.J. (2017) 'What Does It "Mean"? A Review of Interpreting and Calculating Different Types of Means and Standard Deviations', *Pharmaceutics*, 9(2), p. 14. Available at: https://doi.org/10.3390/pharmaceutics9020014.

Mohammad, R.M.A. and Alqahtani, M. (2019) 'A comparison of machine learning techniques for file system forensics analysis', *Journal of Information Security and Applications*, 46, pp. 53–61. Available at: https://doi.org/10.1016/j.jisa.2019.02.009.

Mohayeji, H. *et al.* (2025) 'Securing dependencies: A comprehensive study of Dependabot's impact on vulnerability mitigation', *Empirical Software Engineering*, 30(3), p. 89. Available at: https://doi.org/10.1007/s10664-025-10638-w.

Morgan, D.L. (2018) *Basic and Advanced Focus Groups*. SAGE Publications. Available at: https://www.google.co.uk/books/edition/Basic_and_Advanced_Focus_Groups/_hVcDwAAQBAJ?hl=en&gbpv=0.

NIST (2025) 'Mobile Device Forensic Tool Specification, Test Assertions and Test Cases'. NIST. Available at: https://www.nist.gov/system/files/documents/2025/01/17/Mobile%20Device%20Forensic%20Tool%20Test%20Specification%20V%203.3.pdf.

NPCC (2024a) 'NCSP Artificial Intelligence Large Language Models', *Cyber Standards* [Preprint].

NPCC (2024b) 'NPCC Artificial Intelligence (AI) Strategy'. Available at: https://www.npcc.police.uk/SysSiteAssets/media/downloads/publications/publications-log/science-and-innovation/2025/npcc-ai-strategy.pdf.

NPCC (2025a) 'Artificial Intelligence (AI) Playbook for Policing'. Home Office. Available at: https://library.college.police.uk/docs/NPCC/Artificial-intelligence-playbook-policing-2025.pdf.

NPCC (2025b) 'NPCC Responsible AI Checklist for Policing'. NPCC. Available at: https://library.college.police.uk/docs/NPCC/Responsible-AI-checklist-for-policing-2025.pdf.

NPCC (2026) 'NPCC AI Adoption: Key Recommendations'. (NPCC Artificial Intelligence Portfolio).

NPCC Homicide Working Group (2021) 'Major Crime Investigation Manual (MCIM)'. NPCC. Available at: https://library.college.police.uk/docs/NPCC/Major-Crime-Investigation-Manual-Nov-2021.pdf.

Oh, J. and Hwang, H. (2025) 'Advanced forensic recovery of deleted file data in F2FS', *Forensic Science International: Digital Investigation*, 54, p. 301976. Available at: https://doi.org/10.1016/j.fsidi.2025.301976.

OpenText (2026) *Digital Forensics Software | OpenText Forensic*, *OpenText Forensic (EnCase)*. OpenText. Available at: https://www.opentext.com/uk/products/forensic (Accessed: 21 February 2026).

Paleyes, A., Urma, R.-G. and Lawrence, N.D. (2022) 'Challenges in Deploying Machine Learning: A Survey of Case Studies', *ACM Computing Surveys*, 55(6), p. 114:1-114:29. Available at: https://doi.org/10.1145/3533378.

Parliament (2018) *Disclosure of evidence in criminal cases - Justice Committee - House of Commons*, *The failures in Disclosure*. Available at: https://publications.parliament.uk/pa/cm201719/cmselect/cmjust/859/85904.htm (Accessed: 15 August 2025).

Patil, A. *et al.* (2022) 'Roadmap of Digital Forensics Investigation Process with Discovery of Tools', *Cyber Security and Digital Forensics*. John Wiley & Sons, Ltd, pp. 241–269. Available at: https://doi.org/10.1002/9781119795667.ch11.

Platzer, C., Stuetz, M. and Lindorfer, M. (2014) 'Skin sheriff: a machine learning solution for detecting explicit images', *Proceedings of the 2nd international workshop on Security and forensics in communication systems*. New York, NY, USA: Association for Computing Machinery (SFCS '14), pp. 45–56. Available at: https://doi.org/10.1145/2598918.2598920.

PSNI (2025) *Police Service of Northern Ireland to pilot a new online service to the public titled 'Virtual Officer Attendance' | PSNI*. Available at: https://www.psni.police.uk/latest-news/police-service-northern-ireland-pilot-new-online-service-public-titled-virtual-officer (Accessed: 17 February 2026).

Qadir, A.M. and Varol, A. (2020) 'The Role of Machine Learning in Digital Forensics', *2020 8th International Symposium on Digital Forensics and Security (ISDFS)*. Beirut, Lebanon: IEEE. Available at: https://doi.org/10.1109/ISDFS49300.2020.9116298.

Qadir, S. and Noor, B. (2021) 'Applications of Machine Learning in Digital Forensics'. *International Conference on Digital Futures and Transformative Technologies (ICoDT2)*.

Rahman, M.A. *et al.* (2025) 'Enabling Intelligent Industrial Automation: A Review of Machine Learning Applications with Digital Twin and Edge AI Integration', *Automation*, 6(3), p. 37. Available at: https://doi.org/10.3390/automation6030037.

Rappert, B., Wheat, H. and Wilson-Kovacs, D. (2021) 'Rationing bytes: managing demand for digital forensic examinations', *Policing and Society*, 31(1), pp. 52–65. Available at: https://doi.org/10.1080/10439463.2020.1788026.

Räz, T. (2024) 'ML interpretability: Simple isn't easy', *Studies in History and Philosophy of Science*, 103, pp. 159–167. Available at: https://doi.org/10.1016/j.shpsa.2023.12.007.

Ribaux, O. and Margot, P. (2003) 'Case-Based Reasoning in Criminal Intelligence using Forensic Case Data', *Science & Justice*, 43(3), pp. 135–143.

Robertson, D. *et al.* (2016) 'Face Recognition by Metropolitan Police Super-Recognisers | PLOS ONE'. Available at: https://doi.org/10.1371/journal.pone.0150036.

Rogers, M.K. and Seigfried, K. (2004) 'The future of computer forensics: a needs analysis survey', *Computers and Security*, 23(1), pp. 12–16. Available at: https://doi.org/10.1016/j.cose.2004.01.003.

Rosenfeld, M.J., Thomas, R.J. and Hausen, S. (2019) 'Disintermediating your friends: How online dating in the United States displaces other ways of meeting', *Proceedings of the National Academy of Sciences*, 116(36), pp. 17753–17758. Available at: https://doi.org/10.1073/pnas.1908630116.

Ryan, B. (2024) *Police commissioner says AI will modernise force and help the crackdown on crime*, *Northwich Guardian*. Available at: https://www.northwichguardian.co.uk/news/24635511.cheshire-police-make-use-ai-crackdown-crime/ (Accessed: 28 November 2024).

Sarker, I.H. (2021) 'Machine Learning: Algorithms, Real-World Applications and Research Directions', *SN Computer Science*, 2(3), p. 160. Available at: https://doi.org/10.1007/s42979-021-00592-x.

Saunders, M., Lewis, P. and Thornhill, A. (2019) *Research Methods for Business Students*. Available at: https://www.researchgate.net/profile/Mark-Saunders-10/publication/330760964_Research_Methods_for_Business_Students_Chapter_4_Understanding_research_philosophy_and_approaches_to_theory_development/links/5c53056f299bf12be3f0e2cf/Research-Methods-for-Business-Students-Chapter-4-Understanding-research-philosophy-and-approaches-to-theory-development.pdf (Accessed: 30 March 2021).

Saunders, M.N.K. and Tosey, P. (2013) 'The Layers of Research Design', *Rapport*, pp. 58–59.

Schatz, B.L. (2019) 'AFF4-L: A Scalable Open Logical Evidence Container', *Digital Investigation*, 29, pp. S143–S149. Available at: https://doi.org/10.1016/j.diin.2019.04.016.

Science & Technology in Policing (2023) *Covenant for Using Artificial Intelligence (AI) in Policing*. Available at: https://science.police.uk/delivery/resources/covenant-for-using-artificial-intelligence-ai-in-policing/ (Accessed: 21 February 2026).

Secretary of State (2025) *Disclosure in the Digital Age: Independent Review of Disclosure and Fraud Offences (accessible)*, *GOV.UK*. Parliament: UK Gov. Available at: https://www.gov.uk/government/publications/independent-review-of-disclosure-and-fraud-offences/disclosure-in-the-digital-age-independent-review-of-disclosure-and-fraud-offences-accessible (Accessed: 8 February 2026).

Semantics 21 (2025) 'S21 Global Alliance Database | Semantics 21'. Available at: https://www.semantics21.com/s21-global-alliance-database/ (Accessed: 21 February 2026).

Shahzad, F. *et al.* (2022) 'Cyber Forensics with Machine Learning', in D. Phung, G.I. Webb, and C. Sammut (eds) *Encyclopedia of Machine Learning and Data Science*. New York, NY: Springer US, pp. 1–6. Available at: https://doi.org/10.1007/978-1-4899-7502-7_987-1.

Shayau, Y.H. *et al.* (2019) 'Digital Forensic Investigation Reduction Model (DIFREM) For Windows 10 OS', *IEEE 9th International Conference on System Engineering and Technology (ICSET)*. Shah Alam, Malaysia: IEEE, pp. 459–464.

Shinoda, K. *et al.* (2024) 'Learning from Multiple Annotator Biased Labels in Multimodal Conversation'. *Proc. Interspeech 2024*, pp. 4089–4093. Available at: https://doi.org/10.21437/Interspeech.2024-1197.

Sibe, R.T. and Kaunert, C. (2024) 'Digital Evidence, Digital Forensics, and Digital Forensic Readiness', in R.T. Sibe and C. Kaunert (eds) *Cybercrime, Digital Forensic Readiness, and Financial Crime Investigation in Nigeria*. Cham: Springer Nature Switzerland, pp. 57–83. Available at: https://doi.org/10.1007/978-3-031-54089-9_3.

Sikos, L.F. (2021) 'AI in digital forensics: Ontology engineering for cybercrime investigations', *WIREs Forensic Science*, 3(3), p. e1394. Available at: https://doi.org/10.1002/wfs2.1394.

Sokolova, M. and Lapalme, G. (2009) 'A systematic analysis of performance measures for classification tasks', *Information Processing & Management*, 45(4), pp. 427–437. Available at: https://doi.org/10.1016/j.ipm.2009.03.002.

Solomon, M. *et al.* (2011) *Computer Forensics Jumpstart*. 2nd edn. Wiley Publishing Inc.

Soni, P., Pawar, M. and Shukla, P.K. (2025) 'Exploring Mobile Forensic Investigations of Smartphones Using the Cellebrite UFED Tool', *International Journal for Multidisciplinary Research (IJFMR)*, 7(2). Available at: https://www.ijfmr.com/papers/2025/2/41813.pdf.

Souppaya, M. and Scarfone, K. (2022) *Guide to Enterprise Patch Management Planning: Preventive Maintenance for Technology*. National Institute of Standards and Technology. Available at: https://doi.org/10.6028/NIST.SP.800-40r4-draft.

Stacey, J. *et al.* (2025) 'A responsible artificial intelligence framework for forensic science', *Forensic Science International*, 375, p. 112548. Available at: https://doi.org/10.1016/j.forsciint.2025.112548.

Stella Polaris (2021) *S21 LASERi Suite: Laseri-v (Video)*, *S21 LASERi Suite: Laseri-v (Video)*. Childhood. Available at: https://stellapolaris.childhood.se/natural-language-processing-datorlingvistik/verktyg-i-databasen-som-anvnder-ljud-ai/s21-laseri-suite-laseri-v-video (Accessed: 21 February 2026).

Strickland, C.R. (2022) *Understanding the impact of digital forensic analysts' experiences of working with online child sexual exploitation and abuse*. d_fcpd. University of Birmingham. Available at: https://etheses.bham.ac.uk/id/eprint/13174/ (Accessed: 10 June 2024).

Sunde, N. and Dror, I.E. (2019) 'Cognitive and human factors in digital forensics: Problems, challenges, and the way forward', *Digital Investigation*, 29, pp. 101–108. Available at: https://doi.org/10.1016/j.diin.2019.03.011.

Sunde, N. and Dror, I.E. (2021) 'A hierarchy of expert performance (HEP) applied to digital forensics: Reliability and biasability in digital forensics decision making', *Forensic Science International: Digital Investigation*, 37, p. 301175. Available at: https://doi.org/10.1016/j.fsidi.2021.301175.

Terven, J. *et al.* (2025) 'A comprehensive survey of loss functions and metrics in deep learning', *Artificial Intelligence Review*, 58(7), p. 195. Available at: https://doi.org/10.1007/s10462-025-11198-7.

Thacker, K. (NO DATE) *The TOEX Programme - Bringing Criminals to Justice*, *Science & Technology in Policing*. Science & Technology in policing. Available at: https://science.police.uk/delivery/case-studies/the-toex-programme-bringing-criminals-to-justice/ (Accessed: 21 February 2026).

Thesing, T., Feldmann, C. and Burchardt, M. (2021) 'Agile versus Waterfall Project Management: Decision Model for Selecting the Appropriate Approach to a Project', *Procedia Computer Science*, 181, pp. 746–756. Available at: https://doi.org/10.1016/j.procs.2021.01.227.

Transforming Forensics *et al.* (2020) 'Digital Forensic Science Strategy'. NPCC. Available at: https://www.npcc.police.uk/Digital%20Forensic%20Science%20Strategy%202020.pdf.

Tümen Akyildiz, S. and Ahmed, K.H. (2021) 'An Overview of Qualitative Research and Focus Group Discussion', *International Journal of Academic Research in Education*, 7(1), pp. 1–15. Available at: https://doi.org/10.17985/ijare.866762.

UK Gov (2005) *Freedom of Information Act 2000*. Statute Law Database. Available at: https://www.legislation.gov.uk/ukpga/2000/36/section/17 (Accessed: 28 November 2024).

UK Gov (2021) *Forensic Science Regulator Act 2021*, *Legsilation.gov.uk*. King's Printer of Acts of Parliament. Available at: https://www.legislation.gov.uk/ukpga/2021/14 (Accessed: 13 August 2026).

UK Gov (2024) *Child abuse image database (CAID)*, *GOV.UK*. UK Gov. Available at: https://www.gov.uk/government/publications/child-abuse-image-database/child-abuse-image-database-caid-privacy-notice (Accessed: 21 February 2026).

USA Gov (2014) 'Carter Exhibit 30'. United States of America. Available at: https://htv-prod-media.s3.amazonaws.com/files/carter-exhibit-30-1497356322.pdf.

Veksler, M., Akkaya, K. and Uluagac, S. (2025) 'Practitioner-driven framework for AI adoption in digital forensics', *Forensic Science International: Digital Investigation*, 55, p. 302030. Available at: https://doi.org/10.1016/j.fsidi.2025.302030.

Verdoliva, L. (2020) 'Media Forensics and DeepFakes: An Overview', *IEEE Journal of Selected Topics in Signal Processing*, 14(5), pp. 910–932. Available at: https://doi.org/10.1109/JSTSP.2020.3002101.

Vinayagam, P.S. (2025) 'Mobile Forensics: Investigation and Tools', *International Journal of Computer Trends and Technology*, 73(6), pp. 7–15. Available at: https://doi.org/10.14445/22312803/IJCTT-V73I6P102.

Wen, C. *et al.* (2025) 'Leveraging large language models for thematic analysis: a case study in the charity sector', *AI & SOCIETY* [Preprint]. Available at: https://doi.org/10.1007/s00146-025-02487-4.

Yeow, W.L., Mahmud, R. and Raj, R.G. (2014) 'An application of case-based reasoning with machine learning for forensic autopsy', *Expert Systems with Applications*, 41(7), pp. 3497–3505.

Young, M. *et al.* (2020) 'The Spectrum of Inductive and Deductive Research Approaches Using Quantitative and Qualitative Data', *Academic Medicine*, 95(7), p. 1122. Available at: https://doi.org/10.1097/ACM.0000000000003101.

11.0 Appendices

Appendix 1 – Focus Group Notes

SOMU Focus Group

30/08/2023 - SOMU and Staffordshire Forensic Partnership

SOMU - Staffordshire Police

Forensics - [redacted]

SFP - [redacted]

1 - Tell me about any use of automation that you are aware of within Digital Forensics.

CADE for analysis of images (HASHes)

ADF - Image classifiers using machine learning - what the object / focus of a photograph.

Semantics 21 - Laser Eye X - [redacted] - AI built in to categorise images. DO NOT USE THE AI AS WE DO NOT KNOW IF WE COULD TRUST IT - picks out elements of the photograph such as school logos, and can determine whether the subject of the photograph is naked. This also links in to CADE. Not sure how actively the AI is used outside of us. Could argue that there is too much in it.

Most of this is in relation to photographs.

Fingerprint analysis is automated as far as we are aware.

2 - Are you aware of the use of volatile memory in Digital Forensics?

Not aware of any of this.

Aware that this CAN be done, but it is not done at the moment. This might be used through routers.

3 - Describe any concerns that you would have regarding using an automated tool for conducting analysis of DF sources.

I would have issues with developer bias, they would not want to say that the tool does not do things. Problems are found after the fact. A lot of the developers would not want to admit this - transparency issues.
ISO also requires verification of data. Regulatory issues.
Would be nice to have through a tool that you CAN use - if you can have error from a human you can have machine error - ACCURACY - if I do not know it then how can I expect a machine to know it?

Others would not have an issue - sounds like it is going to save a lot of time.

What kind of time saving and accuracy would you require?

95% of accuracy at the task that is given. As an absolute minimum.

Further concerns over the criteria that it would use to do things like grading of images or data.

It does depend on the level of the tool and what the purpose of the tool would be.

Would need some kind of transparency as to how this was done and how the results were obtained.

4 - Describe what requirements you would place on such a tool, should you need to rely on it?

Accuracy - at least 100% for this

Proven to save time - Not worth the risk if there was no reward

Readable format for the output in an evidential report that displays the criteria that was used.

Legalities of the matter, how has it confirmed that the law has been complied with

Location data if that was relevant to the investigation

Confirmatory testing - it would have to run the tests against the another tool or database as suitable.

Declaration from the creator - how it has been built and how it complies with regulatory and legal requirements - approval from somewhere.

5 - Should such a tool be able to meet these requirements, would you be willing to rely on that tool for conducting analysis?

Working practice wise this would be fine, yes we would take this.

6 - Would you be willing to answer a survey or talk more in depth about this at a future date?

YES

Appendix 2 – Ethics Forms

RESEARCH ETHICS
Proportionate Review Form

The Proportionate Review process may be used where the proposed research raises only minimal ethical risk. This research must: focus on minimally sensitive topics; entail minimal intrusion or disruption to others; and involve participants who would not be considered vulnerable in the context of the research.

PART A: TO BE COMPLETED BY RESEARCHER

Name of Researcher:	Christopher HAWKINS
School	School of Digital, Technologies and Arts

Student/Course Details (If Applicable)			
Student ID Number:			09002003
Name of Supervisor(s)/Module Tutor:			Dr Asma PATEL
PhD/MPhil project:	☒		
Taught Postgraduate Project/Assignment:	☐	Award Title:	
Undergraduate Project/Assignment:	☐	Module Title:	

Project Title:	The Applicability of Machine Learning-based Digital Forensic Solutions to Criminal Justice in the UK		
Project Outline:	The project seeks to develop a framework against which a future ML-based DF tool could be judged, to ascertain and evidence whether this tool can be relied upon for the purposes of Digital Forensics and Criminal Justice.		
Give a brief description of participants and procedure (methods, tests etc.)	The majority of this work will be theoretical in nature - however, there is a requirement for surveys to be disseminated to various persons from a variety of fields including the Police, Criminal Law and HMCS. All surveys will be private and anonmyised when received. A full information sheet, as well as contact and opt out information will be given at the same time as the survey itself.		
Expected Start Date:	01/09/22 (relevant)	Expected End Date:	01/09/23

Relevant professional body ethical guidelines should be consulted when completing this form.

Please seek guidance from the School Ethics Coordinator if you are uncertain about any ethical issues arising from this application.

There is an obligation on the researcher and supervisor (where applicable) to bring to the attention of the School Ethics Coordinator any issues with ethical implications not identified by this form.

Researcher Declaration

I consider that this project has no significant ethical implications requiring full ethical review	☒

I confirm that:		
1.	The research will **NOT** involve members of vulnerable groups. Vulnerable groups include but are not limited to: children and young people (under 18 years of age), those with a learning disability or cognitive impairment, patients, people in custody, people engaged in illegal activities (e.g. drug taking), or individuals in a dependent or unequal relationship.	☒
2.	The research will **NOT** involve sensitive topics. Sensitive topics include, but are not limited to: participants' sexual behaviour, their illegal or political behaviour, their experience of violence, their abuse or exploitation, their mental health, their gender or ethnic status. The research must not involve groups where permission of a gatekeeper is normally required for initial access to members, for example, ethnic or cultural groups, native peoples or indigenous communities.	☒
3.	The research will **NOT** deliberately mislead participants in any way.	☒
4.	The research will **NOT** involve access to records of personal or confidential information, including genetic or other biological information, concerning identifiable individuals.	☒
5.	The research will **NOT** induce psychological stress, anxiety or humiliation, cause more than minimal pain, or involve intrusive interventions. This includes, but is not limited to: the administration of drugs or other substances, vigorous physical exercise, or techniques such as hypnotherapy which may cause participants to reveal information which could cause concern, in the course of their everyday life.	☒
6.	The research **WILL** be conducted with participants' full and informed consent at the time the study is carried out: • The main procedure will be explained to participants in advance, so that they are informed about what to expect. ☒ • Participants will be told their involvement in the research is voluntary. ☒ • Written consent will be obtained from participants. *(This is not required for self-completion questionnaires as submission of the completed questionnaire implies consent to participate)*. ☒ • Participants will be informed about how they may withdraw from the research at any time and for any reason. ☒ • For questionnaires and interviews: Participants will be given the option of omitting questions they do not want to answer. ☒ • Participants will be told that their data will be treated with full confidentiality and that, if published, every effort will be made to ensure it will not be identifiable as theirs. ☒ • Participants will be given the opportunity to be debriefed i.e. to find out more about the study and its results. ☒	YES ☒ N/A ☐
7.	A risk assessment has been completed for this research project	YES ☐

		N/A ☒

If you are unable to confirm any of the above statements, please complete a **Full Ethical Review Form**. If the research will include participants that are **patients,** please complete the Independent Peer Review process.

8. Information and Data Please provide answers to the following questions regarding the handling and storage of information and data:
a) How will research data be stored (manually or electronically)? Electronically within researcher's encrypted storage
b) How is protection given to the participants (e.g. by being made anonymous through coding and with a participant identifier code being kept separately and securely)? All resulting completed surveys will be anonymised through the use of a coding system - only the researcher will have access to the originals. Coding system information will be held securely and privately - only the researcher will have access to this information.
c) What assurance will be given to the participant about the confidentiality of this data and the security of its storage? The method of storage, including within a password protected and encrypted machine, is provided to the participant at the time of responding, and they are also informed of the coding of their data so that their information is not revealed.
d) Is assurance given to the participant that they cannot be identified from any publication or dissemination of the results of the project? Yes - within the information sheet.
e) Who will have access to this data, and for what purposes? The author, for review and action upon the results including obtaining conclusions and evaluations from the respondents as described. No other person will have access to the completed questionnaires. Anonymised versions will be provided with the final report as appendices, but these will not contain personal information.
f) How will the data be stored, for how long, and how will it be discarded? The response questionnaires will be retained as part of the report by the author for the length of the project. They will be be discarded. Anonymised copies will be retained indefinitely, with the consent of the participants - for future review in the case that the work is questioned.

Supporting Documentation

All key documents e.g. consent form, information sheet, questionnaire/interview schedule are appended to this application.	☒

Signature of Researcher:	C HAWKINS	Date:	15/06/2022

NB: If the research departs from the protocol which provides the basis for this proportionate review, then further review will be required and the applicant and supervisor(s) should consider whether or not the proportionate review remains appropriate. If it is no longer appropriate a full ethical review form **MUST** be submitted for consideration by the School Ethics Coordinator .

Next Step:

STUDENTS: Please submit this form (and supporting documentation) for consideration by your Supervisor/ Module Tutor.

STAFF: Please submit this form to your Head of Department or a Senior Researcher in your School. Once they have reviewed the form, this should be forwarded to the Research Administrators in RIIS (ethics@staffs.ac.uk) who will arrange for it to be considered by an independent member of the School's College of Reviewers .

PART B: TO BE COMPLETED BY SUPERVISOR/MODULE TUTOR (If student) OR Head of Department/ Senior Researcher (if staff)

I consider that this project has no significant ethical implications requiring full ethical review by the Faculty Research Ethics Committee.	☐
I have checked and approved the key documents required for this proposal (e.g. consent form, information sheet, questionnaire, interview schedule).	☐

Signature of Supervisor/ Head of Department/ Senior Researcher:		Date:	

Next Step: Please forward this form to the Research Administrators in RIIS (ethics@staffs.ac.uk) who will arrange for it to be considered by an independent member of the School's College of Ethical Reviewers , having no direct connection with the researcher or his/her programme of study.

PART C: TO BE COMPLETED BY A MEMBER OF THE SCHOOL'S COLLEGE OF ETHICAL REVIEWERS

This research proposal has been considered using agreed University Procedures and is now approved.	☐
Or	
This research proposal has not been approved due to the reasons given below.	☐

Recommendation (delete as appropriate): Approve/ Amendments required/ Reject	

Name of Reviewer:		Date:	
Signature:			

Signed (School Ethical Coordinator)		Date:	

RESEARCH ETHICS

Proportionate Review Form

The Proportionate Review process may be used where the proposed research raises only minimal ethical risk. This research must: focus on minimally sensitive topics; entail minimal intrusion or disruption to others; and involve participants who would not be considered vulnerable in the context of the research.

PART A: TO BE COMPLETED BY RESEARCHER

Name of Researcher:	Christopher HAWKINS
School	School of Digital, Technology, Innovation and Business

Student/Course Details (If Applicable)			
Student ID Number:			09002003
Name of Supervisor(s)/Module Tutor:			Dr Russell CAMPION & Mr Christopher HOWARD
PhD/MPhil project:	☒		
Taught Postgraduate Project/Assignment:	☐	Award Title:	
Undergraduate Project/Assignment:	☐	Module Title:	

Project Title:	Automated Digital Forensics - Towards acceptance of Machine Learning and Artifical Intelligence within forensic analysis of image files.
Project Outline:	The project seeks to develop a framework against which a future ML-based DF tool could be judged, to ascertain and evidence whether this tool can be relied upon for the purposes of analysis of image files, and the infentification of images of offence, within the realms of Digital Forensics and Criminal Justice.
Give a brief description of participants and procedure (methods, tests etc.)	*** THIS ETHICS FORM IS A SUPPLEMENT TO THE EXISTING ETHICS FORM IN RELATION TO THIS PROJECT - IT HAS BEEN PREPARED AS THE PROJECT EVOLVED AND THE METHODS USED HAVE DEVELOPED *** In addition to the surveys that have already been conducted as part of this research, and which were the subject of the original ethics form that has been approved, the methods that are being used in this project are as follows: - Focus Group with members of Staffordshire Police SOMU (Sexual Offender Management Unit) regarding initial concepts for the framework, and a broad overview of concerns and requirements on automated tools. - Structured Interviews with members of Cheshire Police's DFU (Digital Forensics Unit) and DMIU (Digital Media Investigation Unit) as well as OCAIT (Online Child Abuse Investigation Team) regarding specific concerns and detailed requirements that they would place onto any framework, or automated tools that would be developed.

	There is no deception involved in this project, and there is no anticipated harm or risk associated with this project, nor posed to the participants. All participants will be proviced with a full information sheet created by the author and checked by the supervision team, which will fully infomr them of their rights and how their information shall be used. They will also be informed of the right to withdraw - the end date for this is given as 30/08/2024 as this is the expected submission date and after this it would not be possible to remove their response. All responses are anonymous, the only personal data of any kind collected will be name of the participant to be mapped against participant number - this will only be used in the vent that the participant wishes to withdraw their participation. Recording of the interviews will be done via typed notes, no audio nor visual recording is being conducted. The interviews that will be conducted ask non-personal questions, and are scripted. Script will be checked by the supervision team prior to being deployed to participants. All participant involvement will be authorised by their line manager and head of unit prior to the interviews taking place.		
Expected Start Date:	01/09/2021	Expected End Date:	01/09/2024

Relevant professional body ethical guidelines should be consulted when completing this form.

Please seek guidance from the School Ethics Coordinator if you are uncertain about any ethical issues arising from this application.

There is an obligation on the researcher and supervisor (where applicable) to bring to the attention of the School Ethics Coordinator any issues with ethical implications not identified by this form.

Researcher Declaration

I consider that this project has no significant ethical implications requiring full ethical review	☒

I confirm that:		
1.	The research will **NOT** involve members of vulnerable groups. Vulnerable groups include but are not limited to: children and young people (under 18 years of age), those with a learning disability or cognitive impairment, patients, people in custody, people engaged in illegal activities (e.g. drug taking), or individuals in a dependent or unequal relationship.	☒
2.	The research will **NOT** involve sensitive topics. Sensitive topics include, but are not limited to: participants' sexual behaviour, their illegal or political behaviour, their experience of violence, their abuse or exploitation, their mental health, their gender or ethnic status. The research must not involve groups where permission of a gatekeeper is normally required for initial access to members, for example, ethnic or cultural groups, native peoples or indigenous communities.	☒
3.	The research will **NOT** deliberately mislead participants in any way.	☒

University Research Ethics Committee (February 2018)

4.	The research will **NOT** involve access to records of personal or confidential information, including genetic or other biological information, concerning identifiable individuals.	☒
5.	The research will **NOT** induce psychological stress, anxiety or humiliation, cause more than minimal pain, or involve intrusive interventions. This includes, but is not limited to: the administration of drugs or other substances, vigorous physical exercise, or techniques such as hypnotherapy which may cause participants to reveal information which could cause concern, in the course of their everyday life.	☒
6.	The research **WILL** be conducted with participants' full and informed consent at the time the study is carried out: • The main procedure will be explained to participants in advance, so that they are informed about what to expect. ☒ • Participants will be told their involvement in the research is voluntary. ☒ • Written consent will be obtained from participants. *(This is not required for self-completion questionnaires as submission of the completed questionnaire implies consent to participate)*. ☒ • Participants will be informed about how they may withdraw from the research at any time and for any reason. ☒ • For questionnaires and interviews: Participants will be given the option of omitting questions they do not want to answer. ☒ • Participants will be told that their data will be treated with full confidentiality and that, if published, every effort will be made to ensure it will not be identifiable as theirs. ☒ • Participants will be given the opportunity to be debriefed i.e. to find out more about the study and its results. ☒	YES ☒ N/A ☐
7.	A risk assessment has been completed for this research project	YES ☐ N/A ☒

If you are unable to confirm any of the above statements, please complete a **Full Ethical Review Form**. If the research will include participants that are **patients,** please complete the Independent Peer Review process.

8. Information and Data Please provide answers to the following questions regarding the handling and storage of information and data:
a) How will research data be stored (manually or electronically)? Electronically within researcher's encrypted storage
b) How is protection given to the participants (e.g. by being made anonymous through coding and with a participant identifier code being kept separately and securely)? All data will be retained only within the researcher's own encrypted OneDrive in line with University procedure and policy.

University Research Ethics Committee (February 2018)

No personal data is being sought, nor will be recorded with the exception of name to participant number in the interviews to ensure that data can be removed should the participant apply for this. There will be no complete document with both answers and names mapped together, answers will be recorded via participant number and participant number to name will be kept separately.
c) What assurance will be given to the participant about the confidentiality of this data and the security of its storage? The method of storage, including within a password protected and encrypted machine, is provided to the participant at the time of responding, and they are also informed of the coding of their data so that their information is not revealed.
d) Is assurance given to the participant that they cannot be identified from any publication or dissemination of the results of the project? Yes - within the information sheet.
e) Who will have access to this data, and for what purposes? The author, for review and action upon the results including obtaining conclusions and evaluations from the respondents as described. No other person will have access to the completed questionnaires or answer documents. Anonymised versions will be provided with the final report as appendices, but these will not contain personal information.
f) How will the data be stored, for how long, and how will it be discarded? The response questionnaires will be retained as part of the report by the author for the length of the project. They will be be discarded after the project has completed + 10 years, in line with University regulations.

Supporting Documentation

All key documents e.g. consent form, information sheet, questionnaire/interview schedule are appended to this application.	☒

Signature of Researcher:	C HAWKINS	Date:	17/02/2024

NB: If the research departs from the protocol which provides the basis for this proportionate review, then further review will be required and the applicant and supervisor(s) should consider whether or not the proportionate review remains appropriate. If it is no longer appropriate a full ethical review form **MUST** be submitted for consideration by the School Ethics Coordinator .

Next Step:

STUDENTS: Please submit this form (and supporting documentation) for consideration by your Supervisor/ Module Tutor.

STAFF: Please submit this form to your Head of Department or a Senior Researcher in your School. Once they have reviewed the form, this should be forwarded to the Research Administrators in RIIS (ethics@staffs.ac.uk) who will arrange for it to be considered by an independent member of the School's College of Reviewers .

PART B: TO BE COMPLETED BY SUPERVISOR/MODULE TUTOR (If student) OR Head of Department/ Senior Researcher (if staff)

I consider that this project has no significant ethical implications requiring full ethical review by the Faculty Research Ethics Committee.	☒
I have checked and approved the key documents required for this proposal (e.g. consent form, information sheet, questionnaire, interview schedule).	☒

Signature of Supervisor/ Head of Department/ Senior Researcher:	R.J.Campion	Date:	17/02/24

Next Step: Please forward this form to the Research Administrators in RIIS (ethics@staffs.ac.uk) who will arrange for it to be considered by an independent member of the School's College of Ethical Reviewers , having no direct connection with the researcher or his/her programme of study.

PART C: TO BE COMPLETED BY A MEMBER OF THE SCHOOL'S COLLEGE OF ETHICAL REVIEWERS

This research proposal has been considered using agreed University Procedures and is now approved. **Or** This research proposal has not been approved due to the reasons given below. **Recommendation (delete as appropriate):** Approve/ Amendments required/ Reject	☐ ☐

Name of Reviewer:		Date:	
Signature:			

Signed (School Ethical Coordinator)		Date:	

School of Digital, Technology, Innovation and Business

ETHICAL APPROVAL FEEDBACK

Researcher name:	Christopher Hawkins
Title of Study:	Automated Digital Forensics - Towards acceptance of Machine Learning and Artificial Intelligence within forensic analysis of image files.
Status of approval:	Approved

Your project ***proposal has been approved*** by the Ethics Panel and you may commence the implementation phase of your study. You should note that any divergence from the approved procedures and research method will invalidate any insurance and liability cover from the University. You should, therefore, notify the Panel of any significant divergence from this approved proposal. This approval is valid only for as long as you are employed at the University.

When your study is complete, please send the ethics committee an end of study report. A template can be found on the ethics BlackBoard site.

The Ethics Committee wish you well with your research.

Signed: **Date: 18.03.2024**



Esther MacCallum Steward

Ethics Coordinator

Appendix 3 – Survey Information Sheet and Question Sheet



SURVEY INFORMATION AND PARTICIPANT RIGHTS

Title of study

Automated Digital Forensics – Towards acceptance of Machine Learning and Artificial Intelligence within forensic analysis of live memory extractions

Invitation

I would like to invite you to participate in this research project which forms part of my Doctorate in Computer Science. Before you decide whether you want to take part, it is important for you to understand why the research is being done and what your participation will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask me if there is anything that is not clear or if you would like more information. My contact information is at the bottom of this section.
Thank you for looking at my survey, this is designed to take no more than 5 minutes to complete.

What is the purpose of the study?

The purpose of this project is to develop a novel framework for the assessment of digital forensic analysis tools, which automate the analysis of evidence sources via the use of Machine Learning and Artificial Intelligence.

This survey is designed to ascertain what existing uses there are for this type of tool, and what frameworks, policies and procedures exist which may impact upon their use.
The survey is also designed to ascertain whether such a tool would be used by practioners and whether a framework could be developed which would allow for reliance on such a tool.

The responses obtained from this survey will be used to validate and verify that such a framework does not already exist, that it would be useful in a real-world scenario and what kinds of criterion would be the most desirable in such a framework.

Do I have to take part?

Response to this questionnaire is deemed to be participation and is entirely voluntary. There is no obligation to complete this in any form.

If there are any questions contained within this questionnaire that you do not wish to answer, please do not answer them. There is no obligation to complete the every question, even if you want to participate – partial completion is welcomed.

Can I withdraw later?

At any point either before, during, or after completion of this questionnaire, you have the right to withdraw from participation. If this is the case, please contact the author and request withdrawal and all data / responses will be removed from the work. Responses cannot be withdrawn once the research is complete.

What are the possible risks of taking part?

They should be no risk involved in this study as all personal details will be anonymised, If you decide to go ahead, you will be asked to sign a consent form to make sure that you fully understand what you are agreeing to. The research has been approved by the Ethics Committee at Staffordshire University.

What happens after the study?

Your responses, in anonymous form, will be analysed by the author, and the statistics found used to inform the survey going forward.

You have the right to be debriefed after completion of this questionnaire. If you wish a debrief, either verbally or written, please contact the author directly.

If you wish to be made aware of any outcome of the work, an update on the status of the project, or any additional information – please contact the author.

** Author contact: christopher.hawkins@staffs.ac.uk - Phone: 07881228016 **

Data Protection and Confidentiality

Data handling and Confidentiality

All data will be treated with confidentiality. No personal data, nor information, is requested - all responses are completely anonymous. Any data referenced in the completed report for this project will be anonymised by referring to respondents by a given code or number.

The questionnaires will be anonymised and the document containing this kept separately from the questionnaires and statistical results. Research data and related material it will be retained for a minimum of 10 years after the study has been completed but with the option to keep data indefinitely as long as it is still useful and there is an intention to use them further for research. After these points electronic data will be deleted. The data will only be shared with my research supervisor and possibly an external examiner.

Anonymised responses will be kept for the length of the project, and for a minimum period of 10 years after the completion of the study, in line with Staffordshire University Policy.

Your data will be processed in accordance with the data protection law and will comply with the General Data Protection Regulation 2016 (GDPR). The data controller for this project will be Staffordshire University. The University will process your personal data for the purpose of the research outlined in this information sheet. The legal basis for processing your personal data for research purposes under GDPR is a 'task in the public interest'. You can provide your consent for the use of your personal data in this study by completing the consent form that has been provided to you. You have the right to access information held about you. Your right of access can be exercised in accordance with the GDPR. You also have other rights including rights of correction, erasure, objection, and data portability. Questions, comments and requests about your personal data can also be sent to the Staffordshire University Data Protection Officer. If you wish to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk.

Thank you for reading this information sheet and for considering taking part in this research.

CONSENT TO PARTICIPATE

This section confirms your consent to participate in this survey.

Please contact the author: christopher.hawkins@staffs.ac.uk if you have any concerns before proceeding.

1. I have read and understood the instruction and information contained on the previous sections.

   ◯ YES

   ◯ NO

2. I understand that I have the opportunity to ask questions and confirm that I am happy to proceed.

   ◯ YES

   ◯ NO

3. I understand that the data will be aggregated for analysis purposes immediately upon submission of my responses and that it will not be possible to disaggregate the data after submission.

   ◯ YES

   ◯ NO

4. I consent for the data collected to be used for publication in academic journals and/or potential presentation in academic forums (conferences, seminars, workshops etc) or can be used for teaching purposes and understand that all data will be so presented **anonymously.**

   ◯ I agree

   ◯ I disagree

5. I understand that I may withdraw my data from the project up to the point that the research is completed. The date for this is not yet known, but is likely to be around August 2024.

   ◯ YES

   ◯ NO

6. I hereby give consent to take part in this study.

   ◯ YES

   ◯ NO

QUESTIONS

Question responses are entirely anonymous. No personal data is requested nor required.

For the purposes of the below questions, references to 'images' refers to any picture, photographs or other graphical files.

Analysis and other related terms have their traditional meaning.

7. What is your age?

- ◯ 18-24
- ◯ 25-34
- ◯ 35-44
- ◯ 45-54
- ◯ 55+

8. What is your length of service within law enforcement in the UK?

- ◯ 0 - 5 years
- ◯ 5 - 10 years
- ◯ 10 + years

9. Do you conduct image analysis as part of your role?

- ◯ Yes
- ◯ No

10. Are you aware of the concepts of Machine Learning and AI in relation to the automation of tasks?

- ◯ Yes
- ◯ No

11. Are you aware of any automated tools that exist which could automate the process of image analysis within digital forensics, or criminal investigations (regardless of whether these are used or not)?

- ◯ Yes
- ◯ No

12. Are you aware of any **use** of automated tools, or any form of automation, within digital forensics or image analysis within the investigations that you undertake?

- ◯ Yes
- ◯ No

13. Are you aware of any frameworks, or minimum standards, that exist within your investigations or organisations, which control the use of automated tools, or automation, within the analysis of suspect images?

- ◯ Yes
- ◯ No

14. Could you name these tools, and mention briefly what they do?

15. Would you be willing to make use of an automated tool to conduct image analysis as part of your investigations?

- ◯ Yes
- ◯ No

16. IF YOU ANSWERED NO TO QUESTION 15:

Would you be willing to use such an automated tool if a framework, or set of rules, were developed against which the tool could be judged, and it's quality assured?

- ◯ Yes
- ◯ No

17. Assuming that you would be willing to use such a tool - whether now, in the future, and with or without a control framework - To what extent would you be willing to rely on an automated tool for such analysis?

- ◯ Not at all - I would not be willing to use such a tool at all.
- ◯ Not at all - I would be willing to use the tool for speculative checks, but would have to conduct a full review myself as part of the investigation.
- ◯ Partially - I would still require a human review to confirm the findings.
- ◯ Wholly - If the tool was authorised for use, I would accept the findings of any such tool.

18. What concerns would you have about using automated tools within image analysis?

19. In relation to a control framework, if one were to be developed, what requirements would you look for in terms of the assessment criteria?

    For example, 'the tool would have to be X% accurate at location images of offence'

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.

Microsoft Forms

Appendix 4 – Interview Information Sheet

RESEARCH DATA - NOT PROTECTIVELY MARKED



Doctoral Research Interview

Automated Digital Forensics – Towards acceptance of Machine Learning and Artificial Intelligence within forensic analysis of image files

Author: Chris HAWKINS (Staffordshire University)

Contact:

Christopher.hawkins@staffs.ac.uk

07881228016

PARTICIPANT RIGHTS AND INFORMATION

Why have I been invited to take part?

You have been invited to participate as you are a serving officer within an English Police Force. Your responses are greatly valued and will help provide valuable evidence supporting my hypothesis for my Doctoral studies.

What will happen if I take part?

The purpose of this research is to develop a novel framework for the assessment of digital forensic analysis tools, which automate the analysis of image-based evidence sources via the use of Machine Learning and Artificial Intelligence.

This interview is designed to ascertain what requirements would be placed on such a framework by active investigators, and what concerns they would have about the use of – or reliance on – automated tools.
Further, the interview is designed to ascertain qualitative information about what those concerns and requirements mean to the investigator.
Finally, the interview is designed to ascertain whether such a framework – if it were to address the concerns and requirements – would provide the investigator with the confidence to rely on automated analysis as part of criminal investigations.

The responses from this, and other interviews conducted as part of this research, will be used in conjunction with other quantitative data obtained in the project to produce the framework outlined above.

Do I have to take part?

Participating in this interview is entirely voluntary. There is no obligation to complete the interview, or answer any questions.

If there are questions that are asked which you do not want to answer, please do not answer them – indicate this to the interviewer.

Partial completion of the interview is welcomed as well as full completion – if you do not wish to answer a question, this does not preclude your answers from being considered.

Participation is voluntary, consent will be obtained via signature of the participant number on a recorded sheet.

What are the possible benefits of taking part?

There are no direct benefits to yourself other than assisting with my study. It is recognised that participation in research projects may cause emotional distress and anxiety in some individuals. If you feel that your psychological wellbeing has been affected, the university has a free counselling service which you can access at by telephone on 01782 294977 or at:
http://www.staffs.ac.uk/study/disabled/mental_wellbeing/index.jsp

What if I change my mind about taking part?

You are free to withdraw at any point of the study, prior to completion, without having to give a reason. Withdrawing from the study will not affect you in any way. You are able to withdraw your data from the study up until 30th August 2024, after which withdrawal of your data will no longer be possible due to the data would have been anonymised/processed and analysed as a part of the final report on the project.

If you choose to withdraw from the study, we will not retain any information that you have provided us as a part of this study.

What will happen to the results of the study?

The results of this study will be fed into the project considerations and will form part of my thesis, there is no immediate plan to publish the results, but there may be in the future.

All published results are anonymous and no names or personal information of participants will be presented – unless you wish to be credited in the final submission, if this is the case please inform the interviewer.

Who should I contact for further information?

If you have any questions or require more information about this study, please contact me using the following contact details: christopher.hawkins@staffs.ac.uk or on phone: 07881228016

What if I have further questions, or if something goes wrong?

This research is being undertaken for the purpose of completing a Doctorate in Computer Science at Staffordshire University. If you have any queries or questions related to this research, please contact me by email at christopher.hawkins@staffs.ac.uk. If you have any concerns about this research, please feel free to contact my supervisors, Dr Russell Campion (R.J.Campion@staffs.ac.uk) and Chris Howard (c.howard@staffs.ac.uk).

DATA PROTECTION AND CONFIDENTIALITY

Data handling and Confidentiality

All data will be treated with confidentiality. No personal data, nor information, is requested in relation to answers - all responses are completely anonymous – the only personal data obtained is name, mapped to participant number – this is solely for the purposes of ensuring that responses can be removed on request. The mapped names will not be stored with the responses, mapped names will be stored separately and will be encrypted. Any data referenced in the completed report for this project will be anonymised by referring to respondents by a given code or number if this is deemed necessary.

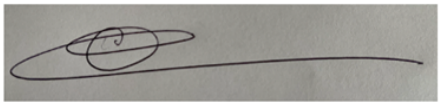
The responses will be anonymised and the document containing this kept separately from the responses and statistical results.

Research data and related material will be retained for a minimum of 10 years after the study has been completed but with the option to keep data indefinitely as long as it is still useful and there is an intention to use them further for research. After these points electronic data will be deleted. The data will only be shared with my research supervision team and possibly an external examiner.

Anonymised responses will be kept for the length of the project, and for a minimum period of 10 years after the completion of the study, in line with Staffordshire University Policy.

Your data will be processed in accordance with the data protection law and will comply with the General Data Protection Regulation 2016 (GDPR). The data controller for this project will be Staffordshire University. The University will process your personal data for the purpose of the re- search outlined in this information sheet. The legal basis for processing your personal data for re- search purposes under GDPR is a 'task in the public interest'. You can provide your consent for the use of your personal data in this study by completing the consent form that has been provided to you. You have the right to access information held about you. Your right of access can be exercised in accordance with the GDPR. You also have other rights including rights of correction, era- sure, objection, and data portability. Questions, comments and requests about your personal data can also be sent to the Staffordshire University Data Protection Officer. If you wish to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk

Thank you for reading this information sheet and for considering taking part in this research.



C HAWKINS – Staffordshire University

Christopher.hawkins@staffs.ac.uk || 07881228016

Appendix 5 – Interview Question Script & Answer Recording Document

NOT PROTECTIVELY MARKED

Doctoral Research Interview

Automated Digital Forensics – Towards acceptance of Machine Learning and Artificial Intelligence within forensic analysis of image files

Doctoral Research Interview – QUESTION SCRIPT & ANSWER RECORDING

Welcome to my PhD Research Interview. Thank you for agreeing to take part. This process is designed to take no more than approximately 10 minutes.

Here is an information sheet which explains the interview process, as well as your rights in relation to this process. Please take a moment to read through this information, and feel free to ask me any questions that you wish.

Have you had time to read the document? Do you understand your rights and the information contained?

I remind you, before we begin, that you have the right to withdraw at any time, and that you do not need to answer any question that you are not comfortable in answering.
If – at any time – you have any concerns, please let me know and I will be happy to help address them if I am able.

To confirm, all responses to this will be anonymous – no personal data is requested nor required in this process.

The recording of your answers will be via my typing them into this document. ***There is no audio or visual recording taking place of any kind.***

Are you happy to proceed to the interview itself?

QUESTION 1:

Within the context of your criminal investigations, and your investigative strategies, of suspect image files – would you consider the use of automated tools to conduct the analysis of such files, and the identification of images constituting an offence *without human intervention or review*?

ANSWER:

QUESTION 2:

Tell me about any concerns that you would have about relying on such automated analysis.

ANSWER:

QUESTION 3:

The aim of this project is to develop a framework against which such tools could be assessed, if a tool meets the framework's requirements, then the tool could be authorised for use in criminal investigations **without the need for human intervention or review.** For such a framework to be created, investigators such as yourself need to be confident in the criteria and tests that have been applied.

In this case, *tell me* what requirements on these tools would you like to see / need, to be confident in their use to the point that you would be able to rely on those tools – it does not matter if you feel that this cannot be met at the moment, it is simply what you would need.

ANSWER:

QUESTION 3a:

Please tell me more about these requirements, explain what you mean.

ANSWER:

NOT PROTECTIVELY MARKED

QUESTION 4:

If a framework were to be applied, meeting the criteria that you have described to me, and a tool passed the tests involved, do you believe that you would feel confident in the reliance on that tool, as a result of the framework being applied?

ANSWER:

Appendix 6 – Testing Focus Group Participant Information Sheet

NOT PROTECTIVELY MARKED



Doctoral Research - Testing

Automated Digital Forensics – Towards acceptance of Machine Learning and Artificial Intelligence within forensic analysis of image files

Doctoral Research Testing – Focus Group Feedback Document

(Semi-Structured)

Participant Information Sheet

Thank you for agreeing to continue participating in my doctoral research.

This phase of the research involves the deployment of the framework that we spoke about during the previous interviews that we had.

Following those interviews, the framework has been created – as we spoke about during the interviews, the purpose of this framework is to begin the process of moving towards the acceptance of automation within forensic analysis of image files. The idea behind this framework is to provide a set of minimum standards, against which automated tools would be assessed to ascertain whether they are suitable for being reliable upon during criminal investigations.

The purpose of this session is to obtain your feedback and thoughts on the produced framework – specifically to find out if this framework is relevant to your work, addresses your concerns and provides you with the confidence to consider using automated tools in your criminal investigative strategies.
The format of the session will be that I will introduce the framework and explain this to you, you will also be shown the framework.
I will then ask a series of questions and take a note of your responses and comments.

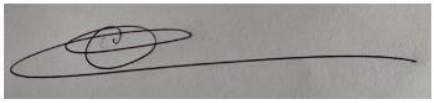
The only recording taking place is via these notes. There is no audio-visual recording taking place.

As before, you have several rights and entitlements regarding your participation in this project.
These are explained on following pages under *Participant Rights and Entitlements.*

Thank you for agreeing to continue taking part in this project, if you have any questions or concerns, please contact me as below.

Thank you

Chris Hawkins



Christopher.hawkins@staffs.ac.uk

07881228016

Participant Rights and Entitlements

Why have I been invited to take part?

You have been invited to participate as you are a serving officer within an English Police Force. Your responses are greatly valued and will help provide valuable evidence supporting my hypothesis for my Doctoral studies.

What will happen if I take part?

The purpose of this research is to develop a novel framework for the assessment of digital forensic analysis tools, which automate the analysis of image-based evidence sources via the use of Machine Learning and Artificial Intelligence.

This focus group is designed to ascertain whether the created framework meets the above-mentioned factors.

The responses from this session, as well as other data obtained during the course of the project, will be used to assess the suitability of the produced framework, and the project overall.

Do I have to take part?

Participating in this session is entirely voluntary. There is no obligation to participate, or answer any questions.

If there are questions that are asked which you do not want to answer, please do not answer them – indicate this to the interviewer.

Partial completion of the session is welcomed as well as full completion – if you do not wish to answer a question, this does not preclude your answers from being considered.

Participation is voluntary, consent to participate is implied by virtue of your continued attendance at the focus group. Personal information is NOT requested, required nor recorded. Individual responses are not recorded.

What are the possible benefits of taking part?

There are no direct benefits to yourself other than assisting with my study. It is recognised that participation in research projects may cause emotional distress and anxiety in some individuals. If you feel that your psychological wellbeing has been affected, the university has a free counselling service which you can access at by telephone on 01782 294977 or at: http://www.staffs.ac.uk/study/disabled/mental_wellbeing/index.jsp

What if I change my mind about taking part?

You are free to withdraw at any point of the study, prior to completion, without having to give a reason. Withdrawing from the study will not affect you in any way. You are able to withdraw your data from the study up until 30th August 2024, after which withdrawal of your data will no longer be possible due to the data would have been anonymised/processed and analysed as a part of the final report on the project.

If you choose to withdraw from the study, we will not retain any information that you have provided us as a part of this study.

What will happen to the results of the study?

The results of this study will be fed into the project considerations and will form part of my thesis, there is no immediate plan to publish the results, but there may be in the future.

All published results are anonymous and no names or personal information of participants will be presented – unless you wish to be credited in the final submission, if this is the case please inform the interviewer.

Who should I contact for further information?

If you have any questions or require more information about this study, please contact me using the following contact details: christopher.hawkins@staffs.ac.uk or on phone: 07881228016

What if I have further questions, or if something goes wrong?

This research is being undertaken for the purpose of completing a Doctorate in Computer Science at Staffordshire University. If you have any queries or questions related to this research, please contact me by email at christopher.hawkins@staffs.ac.uk. If you have any concerns about this research, please feel free to contact my supervisors, Dr Russell Campion (R.J.Campion@staffs.ac.uk) and Chris Howard (c.howard@staffs.ac.uk).

DATA PROTECTION AND CONFIDENTIALITY

Data handling and Confidentiality

All data will be treated with confidentiality. No personal data, nor information, is requested in relation to answers - all responses are completely anonymous – no personal data is recorded or obtained. All focus group information is only included as part of the group response, and therefore there is nothing to identify any participant in any way.

Research data and related material will be retained for a minimum of 10 years after the study has been completed but with the option to keep data indefinitely as long as it is still useful and there is an intention to use them further for research. After these points electronic data will be deleted. The data will only be shared with my research supervision team and possibly an external examiner.

Anonymised responses will be kept for the length of the project, and for a minimum period of 10 years after the completion of the study, in line with Staffordshire University Policy.

Your data will be processed in accordance with the data protection law and will comply with the General Data Protection Regulation 2016 (GDPR). The data controller for this project will be Staffordshire University. The University will process your personal data for the purpose of the re- search outlined in this information sheet. The legal basis for processing your personal data for re- search purposes under GDPR is a 'task in the public interest'. You can provide your consent for the use of your personal data in this study by completing the consent form that has been provided to you. You have the right to access information held about you. Your right of access can be exercised in accordance with the GDPR. You also have other rights including rights of correction, era- sure, objection, and data portability. Questions, comments and requests about your personal data can also be sent to the Staffordshire University Data Protection Officer. If you wish to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk

Thank you for reading this information sheet and for considering taking part in this research.

Appendix 7 – Testing Focus Group Domains Script

NOT PROTECTIVELY MARKED



Doctoral Research - Testing

Automated Digital Forensics – Towards acceptance of Machine Learning and Artificial Intelligence within forensic analysis of image files

Doctoral Research Testing – Focus Group Feedback Document

(Semi-Structured)

NOT PROTECTIVELY MARKED

Introduction Wording

Welcome to the focus group for my research.

As you will see on your sheet, the purpose of this session is to showcase the framework that has been produced as part of my research.

In this session, I will show you the framework that has been created and explain the different elements of this to you.

Following this, I will ask you a series of questions to encourage a discussion on the framework and obtain your views and feelings on the elements that are needed for the testing of the framework.

DOMAINS

Relevance and Applicability

How relevant do you find the framework to criminal investigations?

Practicality and Feasibility

Do you believe the framework is practical for real-world deployment?

Are there any implementation challenges or resource constraints that could hinder its adoption?

Usability and Confidence

Would using this framework increase your confidence in the tools or solutions that meet its criteria?

How user-friendly is the framework in terms of understanding and applying it?

Feedback and Improvements

What improvements or enhancements would you suggest for the framework?

Are there any additional factors or considerations that should be included?

NOT PROTECTIVELY MARKED

LIKERT SCALE ANSWERS

Guidance:

1 = Disagree -> 5 = Agree

To what extent do you believe that *each overall domain* has been met by the Framework that has been presented?

We will then follow up the responses with discussion amongst the group

Participant	Domain 1	Domain 2	Domain 3	Domain 4
1				NOT ASSESSED VIA LIKERT SCALE
2				
3				
4				
5				
6				

Appendix 8 – Primary Research Conducted in 2021

Although a literature review has found that there is a large amount of academic research showing that academia is concerned about the efficiency of Digital Forensic processes, the researcher also considered the views of officers working on the frontline of these investigations. In order to ascertain whether this academic concern is shared with those officers conducting real-world investigations, the same twelve officers as responded to the researcher's questionnaire about the use of the Volatility Framework, were also asked if they considered that there was a problem with the number of devices requiring analysis in digital investigations, and whether the amount of time that was taken for the conducting of that analysis was a problem.

[GRAPHS SHOWN IN Section 3.1 OF THIS THESIS]

As can clearly be seen from the responses, most respondents stated that there are concerns around both the number of devices requiring analysis, and the amount of time that it takes for said analysis.

Appendix 9 – Framework Testing Focus Group Results

NOT PROTECTIVELY MARKED



Doctoral Research - Testing

Automated Digital Forensics – Towards acceptance of Machine Learning and Artificial Intelligence within forensic analysis of image files

Doctoral Research Testing – Focus Group Feedback Document

(Semi-Structured)

Introduction Wording

Welcome to the focus group for my research.

As you will see on your sheet, the purpose of this session is to showcase the framework that has been produced as part of my research.

In this session, I will show you the framework that has been created and explain the different elements of this to you.

Following this, I will ask you a series of questions to encourage a discussion on the framework and obtain your views and feelings on the elements that are needed for the testing of the framework.

DOMAINS

Domain 1 - Relevance and Applicability

How relevant do you find the framework to criminal investigations?

Do you believe that the Framework that has been produced would be of value to the Criminal Justice sector?

Would a tool that achieved compliance with the framework be usable?

Domain 2 - Practicality and Feasibility

Do you believe the framework is practical for real-world deployment?

Are there any implementation challenges or resource constraints that could hinder its adoption?

Domain 3 - Usability and Confidence

Would using this framework increase your confidence in the tools or solutions that meet its criteria?

How user-friendly is the framework in terms of understanding and applying it?

Domain 4 - Feedback and Improvements

What improvements or enhancements would you suggest for the framework?

Are there any additional factors or considerations that should be included?

The Focus Group for the Framework testing was conducted on 13/12/2024.

The participants within this group were respondents from the online surveys and the interviews, only participants who had participated in both elements were accepted as focus group members.

The participant identities were not recorded and no personal information was requested nor required as part of this focus group.

All ethical considerations were met and the information sheet was given to every participant.

NUMBER OF FOCUS GROUP MEMBERS: 6 (multiple units)

COMMENTS AND FOCUS GROUP RESPONSES:

DOMAIN 1

General Comments

Any use of an OR gate would need to be to do with verified to ISO standards. This is often missed when software is developed. Even if this is already done, the cops would have to do it again to ensure that accreditation is maintained.

Updating in any software is a problem – in relation to the whole sector there are concerns about safeguards when something is updated. Safeguards should be in place which would ensure that any updates to either the software or the data that it uses for decision making is clearly stated and is assessed to the same level as the tool when it is first considered.

I believe that the framework would be valid and relevant to investigations and it appears to meet the requirements that I would have in respect of assessing a tool. I have concerns over the explainability of the framework in a practical sense but can talk about this in a later question.

How relevant do you find the framework? Do you believe that the framework would be of value to the sector?

The framework is relevant to the work that is undertaken, and a tool that meets these criteria would certainly be usable in this sense, although I am not sure that there is a tool that would meet these criteria at the moment from the work that is being done.

There is a strong desire to adopt this kind of technology within the organisation and as a result there is a definitive need for this kind of assessment as we look to explore how we can use this. At the moment, we need a human to review everything, and this is something that we need to consider moving forward. We need to have some way of assessing tools for any real automated activity if we are going to move away from needing a human.

The technology is not going to stop developing and so we need to make sure that we are assessing properly. This framework is a good way of doing that.

Do you believe that a tool meeting the framework would be usable?

Yes, I think that if a tool met all the criteria of the framework, this would be something that could be used, although there are other constraints such as national rules around how AI is used. However, this is something that will have to change in the future and if a tool were to meet these criteria then this is something that would be supported.

There would need to be some consideration to third party accreditation requirements, but this would be organisational and not something that would have to be in the framework.

DOMAIN 2

General Comments

The framework may be clear to someone who is technically minded or has the time to work on understanding it, but it could do with being clearer in respect of the actual definitions that are being used.

Practical for real-world deployment?

The framework may be practical insofar as it flows from one point to another and there is a logical flow to it which helps with understanding.

There are some issues around the understandability of the metrics that are being used to assess whether a gate is passed or not. This may be contained in the explanatory notes, and that would be fine, but to go with this there should be more of an explanation of what a gate means.

A definition of the gate would be massively helpful in ensuring that the framework is usable to operational officers, and could be used to quickly assess tools – which is something that is required.

There needs to be clarity in what is being assessed and how, which would also be assessed by making the definitions clearer. It would also be beneficial to ensure that there is a 'plane' approach to the framework so it could be looked at as to where tools are failing for the industry and future development.

Implementation Challenges?

The only thing that can be seen is that there will be a need for repeated updating of the software as thing change. At the moment, software like Cellebrite updates very often so it needs to constantly be running through the framework as this happens. It would be more of a living process rather than a one-off and the industry needs to be aware of that.

As has been said before, there are also accreditation issues with software but the framework seeks to address this rather than get in the way of it so this is not a concern at the moment.

DOMAIN 3

General Comments

So long as the framework was accepted at organisational level, and a tool passed the relevant aspects of the framework, then this would be acceptable and we could make use of the tool in the manner that passed.

Would using this framework increase your confidence in the tools or solutions that meet its criteria?

Absolutely, so long as there was organisational acceptance of the framework as well and there was documentation to show that the tool had been assessed and had passed then this would be fine.

It is not certain that there are tools that would already pass this framework, but that is part of the research.

How user-friendly is the framework in terms of understanding and applying it?

As has been said in other questions, the framework itself has a logical flow and seems to meet all the requirements that are present. The logic of it is present and appears to be justifiable.

However, to make it more usable and accessible to officers and the industry who have to meet this framework, the main thing would be to make it a little more readable, maybe put in some colour to it and also add in the definitions of what is being assessed so that it could be assessed at a glance.

This would make it much more user friendly as even with the manual, it needs to be accessible quickly for non-technical people as well as those with existing knowledge.

DOMAIN 4

General Comments

Organisational acceptance, the short version is that If my boss tells me that it is accurate and it works then that's fine, I would use it.

Random dip sampling is important. No so concerned about 100% accuracy.

False positives are the biggest concern rather than anything else in terms of accuracy.

The human element – this the current norm, and is present in all of what we do with AI. It needs to be present in the research and needs to be part of the dip sample process that is already present.

The responsibility of this is large (imprisoning people) – the weight of this needs to be put on companies that are developing these tools.

Culture, understanding and education will be a big element to this – there may be a tool that passes the framework, but it is foreseen that officers and practitioners would be reluctant to make use of the tool as they do not understand it. There is a strong cultural element in law enforcement around understanding what you are using, how it works and being able to explain the results. At the end of the day, it is you and not the tool that would be on the stand in the courtroom.

What improvements or enhancements would you suggest for the framework?

We need to make sure that there is something that allows for robust testing to deal with any bias in the dataset that have been used to train the models that would be relied on.

NOT PROTECTIVELY MARKED

The testing for the software will need to be living rather than static.

Ensure that testing is a key factor in this.

Are there any additional factors or considerations that should be included?

Welfare over time saving should be mentioned.

There is a legal process behind anything like this – for example, who would produce a report or finding from the AI model? Although this may be more of a sector / organisational decision rather than something that could be addressed in the framework.

LIKERT SCALE ANSWERS

Guidance:

1 = Disagree -> 5 = Agree

To what extent do you believe that *each overall domain* has been met by the Framework that has been presented?

We will then follow up the responses with discussion amongst the group

Participant	Domain 1	Domain 2	Domain 3	Domain 4
1	3	2	2	NOT ASSESSED VIA LIKERT SCALE
2	5	3	1	
3	5	2	2	
4	5	2	2	
5	4	3	2	
6	5	2	1	

Appendix 10 – Metric Passing Value Testing Results

NOT PROTECTIVELY MARKED



Doctoral Research - Testing

Automated Digital Forensics – Towards acceptance of Machine Learning and Artificial Intelligence within forensic analysis of image files

Doctoral Research Testing – Single Expert Review

NOT PROTECTIVELY MARKED

Thank you for agreeing to participate in the testing element of my doctoral research.

This phase of the research involves the review and feedback on the 'passing values' that were given for the produced Framework that we spoke about during the Focus Group session.

The following questions are designed to obtain your views on the produced metrics. You will also be given a copy of the guidance document which explains the rationale for these metrics being selected.

As we spoke about previously, the purpose of this framework is to begin the process of moving towards the acceptance of automation within forensic analysis of image files. The idea behind this framework is to provide a set of minimum standards, against which automated tools would be assessed to ascertain whether they are suitable for being reliable upon during criminal investigations.

As before, you have several rights and entitlements regarding your participation in this project.
These are explained on the attached Rights Sheet.

Thank you for agreeing to continue taking part in this project, if you have any questions or concerns, please contact me as below.

Thank you

Chris Hawkins



Christopher.hawkins@staffs.ac.uk

07881228016

QUESTIONS

Please complete the following table, having taken time to review both of the following documents:

- **Document 1 – Framework Metrics**
- **Document 2 – Explanatory Documentation**

If there are any questions, please make contact with the researcher as found on Page 2.

QUESTION 1 – Passing Levels

Having reviewed the documentation that has been provided, do you agree with the level given for each of these metrics?

METRIC	VALUE	AGREED?	COMMENTS
PRACTICAL RELIABILITY	5	Y	100% not necessary, 90% is acceptable here.
TIME SAVING	3	Y	Tool dependant but some time saving necessary.
TRANSPARENCY	3	Y	Technical info not required.
VERIFICATION	3	Y	Similar to Transparency - for the OIC to do not the tool/developer.
REPORT GENERATION	4	Y	Extremely important.
UPDATING	3	Y	Be careful of auto updating.
TOOL TIE-IN	3	Y	Rationale agreed.
MALLEABILITY	4	N	malleability is key – victim ID is paramount.

Appendix 11 – PR Metric Expert Testing Results

NOT PROTECTIVELY MARKED



Doctoral Research - Testing

Automated Digital Forensics – Towards acceptance of Machine Learning and Artificial Intelligence within forensic analysis of image files

Doctoral Research Testing – Single Expert Review

PRACTICAL RELIABILITY METRIC

Thank you for agreeing to participate in the testing element of my doctoral research.

This phase of the research involves the review and feedback on the 'Practical Reliability Metric' that has been created for the assessment of tools through the produced Framework that you have already seen as part of this project.

The following questions are designed to obtain your views on the produced metric.

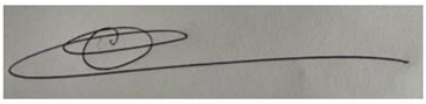
As we spoke about previously, the purpose of this framework is to begin the process of moving towards the acceptance of automation within forensic analysis of image files. The idea behind this framework is to provide a method of assessment for tools that are submitted in a clear and understandable form, with results that are easy to interpret.

As before, you have several rights and entitlements regarding your participation in this project.
These are explained on the attached Rights Sheet.

Thank you for agreeing to continue taking part in this project, if you have any questions or concerns, please contact me as below.

Thank you

Chris Hawkins



Christopher.hawkins@staffs.ac.uk

07881228016

QUESTIONS

NOT PROTECTIVELY MARKED

Please complete the following table, having taken time to review both of the following documents;

- **Document 1 – PR Metric Calculation**
- **Document 2 – Explanatory Documentation**

If there are any questions, please make contact with the researcher as found on Page 2.

QUESTION 1 – PR Metric

Having reviewed the documentation that has been provided, please comment on the following:

METRIC	COMMENTS
UNDERSTANDABILITY OF METRIC	Difficult to understand as I'm not a mathematician. However, with more info became clearer.
CLARITY OF CALCULATION	Makes operational sense. A single metric is preferable than multiple technical calculations
EASE OF CALCULATION	Concerned that some info may not be available. Actual maths appears easy to do.
UNDERSTANDABILITY OF OUTPUT	Giving a defined result range works. Using this will help.
OVERALL EASE OF USE AND UNDERSTANDING	Lack of info will impact usability but is clear in itself. Training will be needed.

Appendix 12 – Additional Validation Results

Academic:

Your experience	*University of Staffordshire - Senior Lecturer in Forensic Investigation & Course Leader for MSc Digital Forensic Investigation* Since 2021, I have taught and assessed digital forensics across Levels 4–7, supervised student projects and contributed to the research/academic Staffordshire Forensic Partnership with Staffordshire Police. Previously, I conducted digital forensic examinations for CCL Forensics, Staffordshire Police and the Metropolitan Police Service, holding Specialist and Coordination roles, on an estimated 400+ cases. My experience also includes auditing ISO and FSR Code compliance, unit management, contributing to cross-department decisions and national tool-procurement projects and presenting research, most recently at Forensic Europe Expo 2026. I hold a first-class honours degree in Digital Forensics, a Postgraduate Certificate in Higher Education and 20+ relevant professional certifications.			
	Usable in practical terms?	**Beneficial to the organisation / Digital Forensics?**	**Understandable?**	**Any comments and whether you agree with the developed item**
PR Metric	Yes, because the calculation is clearly explained, straightforward to apply and can be performed by practitioners without advanced mathematical knowledge.	Yes, because this accessible metric would support quality management, accreditation and tool-procurement decisions.	Yes, because the underlying theory, practical application and benefits of the metric are clearly explained.	I agree with this item because it addresses key challenges within the field concerning the interpretation of accuracy and reliability of AI/ML automation,
Framework	Yes, because the framework is clear, transparent and practicable; it supports practitioners' operational needs while providing the explainability that may be required within the wider criminal justice system, including in court.	Yes, because the framework comprehensively addresses the key stages an organisation should consider when determining whether a tool is fit for purpose from a quality-assurance perspective.	Yes, because the framework is clearly structured, and its requirements, subsets and features are justified with reference to relevant research and practitioner experience.	I agree with this item because it is carefully constructed and demonstrates real-world applicability.
Gate Passing Metrics	Yes, because the proposed metrics and passing thresholds hold an appropriate balance between requirements/industry standards and practical tool use.	Yes, because the metrics would increase transparency in tool use and provide a measurable basis for dialogue with developers about any required improvements	Yes, because the metrics are clearly explained in relation to industry practice, and their inclusion is appropriately justified	I agree with this item because the proposed metrics and passing thresholds provide clear, measurable criteria for determining whether AI/ML automation is suitable for operational DF use.

AI Enablement 1:

Your experience	*I have been a police officer in Cheshire police for 7 years. I worked as a response officer, whilst obtaining my degree in policing. I then worked in the DFU team for 2 years, focusing on phone extractions and obtaining qualifications required for the role. I am now currently in an AI enablement team focusing on the development and implementation of AI solutions for the constabulary.*			
	Usable in practical terms?	Beneficial to the organisation / Digital Forensics?	Understandable?	Any comments and whether you agree with the developed item
PR Metric	The PR Metric is a clear formula that is straight forward to test with a suitable data set.	This is beneficial as it provides a more holistic scoring metric, building on the combined strengths of those that have been previously noted and allows for accessible comparison.	The workings and rational of the PR Metric are made clear.	The PR Metric takes into consideration both F1 and MCC encapsulating, within one result, more relevant elements required to make a judgment on the suitability of a solution.
Framework	The framework is clearly laid out and flows logically through the requirements that support the implementation of True Automation.	The Framework is beneficial, showing the sequential requirements and considerations for the True Automation Implementation.	This is easy to understand, each aspects description explaining the requirement. The Framework shows which elements sit together at each stage.	I agree with the developed Framework and the logic of the assessment that it presents.
Gate Passing Metrics	This is usable in practical terms with each gate having clearly defined scoring making assessment straight forward.	This would be beneficial, both for the organisation and DF as it allows easy assessment across various technical capabilities.	It is understandable as each gate is clearly explained as it the rational for the required score.	I agree with the Gate Passing Metrics. This makes a suitability assessment much more straight forward.

AI Enablement 2:

Your experience	I am currently employed within Cheshire Constabulary's AI Enablement team based within Police Headquarters, Winsford, Cheshire where I develop Copilot and Power Platform solutions. I have also separately developed a number of software solutions incorporating AI frameworks and ML elements.			
	Usable in practical terms?	Beneficial to the organisation / Digital Forensics?	Understandable?	Any comments and whether you agree with the developed item
PR Metric	Yes	Yes	Yes	I agree with the developed item. The PR element is well thought out drawing upon other known metrics as well as outlining its usability within the DF field.
Framework	Yes	Yes	Yes	The framework is clear as to its applicability within the DF field.
Gate Passing Metrics	Yes	Yes	Yes	I agree with the gate passing metrics, their outlined scores including the rationale for those scores.

DFU Manager:

Your experience	*I am currently employed by Cheshire Police as a Senior Digital Forensics Investigator / Supervisor based at Police Headquarters, Winsford and am responsible for the implementation and licensing of tools within the Digital Forensics department. I have worked in the field of Digital Forensics for 15 years and have experience in both the public and private industry. I have a number of professional qualifications including a BSC (Hons) in Computer Forensics.*			
	Usable in practical terms?	Beneficial to the organisation / Digital Forensics?	Understandable?	Any comments and whether you agree with the developed item
PR Metric	Yes	Yes	Yes	PR metric, whilst a little complex initially, was explained well and is usable within the DF field.
Framework	Yes	Yes	Yes	Appears to cover all aspects of pre-requisites in relation to DF based AI automation.
Gate Passing Metrics	Yes (with some comments as per final column)	Yes	Yes	From a DF management perspective, we would not be accepting tools that have a practical reliability of 90% (0.90). It is too low for our threshold therefore my view would be to amend the gate 5 scoring to be a practical reliability of 0.95-1. A similar assessment would also come with the updating gate, whereby a score of 3 would not be sufficient for use in the DF environment. 3 does not prompt any user interaction before updating however, as an accredited laboratory, we would require a score of at least 4 whereby updates are selectable and user interaction is required. An assessment needs to be made of each update prior to commencing it as to whether the tool would require revalidation or whether the update is indeed worth rolling out. The changes above would then also need to reflect in the table in Section 5.6.