

# Exploring Strategies to Prevent Harm from Web Search

Steven Edward Zimmerman

A thesis submitted for the degree of

*Doctor of Philosophy (PhD)*



Computer Science and Electronic Engineering

University of Essex

Awarded January 2021



*primum non nocere*

first, do no harm



TO GERULF AND HENRY

Words cannot express my gratitude for the support  
and inspiration you provide.



# Abstract

Web search, the process of seeking and finding information online, is an ubiquitous activity engrained in the lives of many individuals and much of broader society. This activity, which has brought many benefits to individuals and society, has also opened the door to many harms, such as echo chambers, loss of privacy and exposure to misinformation. Members of the information retrieval (IR) community now recognize the dangers of the search technologies commonplace in our daily lives. The upshot of this recognition are growing efforts to address these dangers by the IR community. These efforts focus heavily on system oriented solutions, but give limited focus on behavioural and cognitive biases and behaviours of the search and even less attention to interventions designed to address these biases and behaviours. As such, a theoretical framework is proposed, with behavioural and cognitive strategies as a core component of interactive Web search environments designed to minimize harm.

Using the framework as the foundation, this thesis presents a number of offline and online studies to evaluate *nudging*, a popular intervention strategy rooted in the field of behavioural economics, and *boosting*, a successful intervention strategy from the cognitive sciences, as strategies to reduce risk of harm in Web search. Overall the studies produce findings in line with the theories underlying the behavioural and cognitive strategies considered. The key takeaway from these studies being that both *boosting* and *nudging* should be considered as viable approaches for harm prevention in Web search environments, in addition to pure system and algorithmic solutions. Additional contributions of this thesis include methods of study design for the comparison of multiple paradigms that promote improved decision making, along with a set of evaluation metrics to measure the success of the IR system and user performance as they relate to the harms being prevented. Future research is needed to confirm the effectiveness of these strategies for other types of harms.





## Acknowledgements

The foundation for independent research was laid during my second journey into academia in 2010, which included studies in maths, atmospheric sciences and computing at Cornell in Ithaca. The first class teaching and the pressures of these courses taught me how to think and persevere through difficult problems. Thank you to all of my course instructors.

However, the journey of this thesis truly began in 2013 when I was introduced to the field of information retrieval by my now current advisor. To *Udo Kruschwitz*, you have been integral in this journey. Without the feedback, suggestions, connections to industry and freedom to develop my research you have given, I would not stand here today.

Thank you to my co-supervisors, *Jon Chamberlain* and *Chris Fox*, who have also provided helpful guidance along the way. As we are all part of the Human Rights Big Data and Technology (HRBDT) Project, our time together was only possible through the Economic and Social Research Council funding (grant number ES/M010236/1) that supported our efforts. And of course, I must thank the *HRBDT administration* for ensuring all costs (including payment of participants) was covered.

Autumn 2017, was a pivotal point in my research and PhD, where I attended the Autumn School for Information Retrieval and Foraging (ASIRF) at the Schloss-Dagstuhl Leibniz-Zentrum für Informatik. I am grateful for the funding bodies that support the event, as it was my introduction to experimental IIR with human subjects and my first step away from the pure system view. At this event, I met *David Elweiler*, for whom I am grateful for our many talks and feedback on design of my last 2 studies during my short time in Regensburg.

Much of my knowledge in behavioural and cognitive interventions is grounded in the 2018 Summer School for Bounded Rationality at the Max Planck Institute (MPI) for Human Development, for which I am

again grateful to the funders. During the summer school, MPI researcher *Stefan Herzog* took notice of my research, and we have carried discussions on ever since which have greatly nurtured my proficiency of decision making interventions.

A third hallmark event I attended was the presentation of my research at the 2018 SIGIR Doctoral Consortium (DC), for which awesome feedback from the mentors was only possible through the SIGIR student travel grant. It is at the DC where *Mark Smucker* suggested the search tasks used in the thesis, and he later connected me to his student *Amira Ghenai* for a copy of their evaluation test set (the starting point for test sets in this thesis).

To *Jaime Arguello*, thank you for introducing the basics of data collection for interactive studies, and for great laughs at conferences!

Sending much appreciation to *Annette Jäckle* (a professor in survey methodology in the building next door to my office) who introduced me to excellent resources for survey development and also provided constructive input for refinement of the surveys used in my studies.

A massive thank you to Cliqz and Ghostery as organizations and *Arjaldo Karaj*, *Sam Macbeth* and *Josep M. Pujol* as individuals for compiling the [WhoTracks.me](https://whotracks.me) data set used in several of my studies.

Without *anonymous participants*, none of the studies in this thesis would have been possible, thank you for your efforts!

*Alistair Thorpe*, without your connections, I would not have been able to recruit participants with such ease! Our many discussions about risk and decision making in the gym, along with statistical methods, have been so important for this project. Words cannot express my appreciation. And don't let me forget to return your copy of Nudge!

To my friends and family who have been there through the years, love to you all. I look forward to a catch-up once this Covid madness has abated.

Finally, and certainly not least, to my parents *John*, *Janet* and *Ed*. You were so patient with me in my younger years and so supportive of my growth through challenging times in my later years, thank you, thank you, thank you!

# Contents

<b>List of Tables</b>	<b>xvii</b>
<b>List of Figures</b>	<b>xxi</b>
<b>I Theory</b>	<b>1</b>
<b>1 Motivation</b>	<b>3</b>
1.1 Harms from Web Search . . . . .	3
1.2 Scope of Harms . . . . .	8
1.3 Research Questions . . . . .	10
1.4 Thesis Contributions . . . . .	12
1.5 Presentation Style . . . . .	14
1.6 Thesis Structure . . . . .	16
<b>2 Background</b>	<b>19</b>
2.1 Overview . . . . .	19
2.2 The Search Process . . . . .	21
2.2.1 Searching the Library . . . . .	21
2.2.2 Information Behaviour and Information Seeking . . . . .	23
2.2.3 How Information is Used . . . . .	26
2.2.4 Evolutionary and Ecological Views of Search . . . . .	30
2.2.5 Economic View . . . . .	33
2.2.6 Concepts . . . . .	35
2.3 Information Retrieval System . . . . .	38
2.3.1 Indexing Process . . . . .	39
2.3.2 Query Process . . . . .	42
2.4 Sources of Harm . . . . .	44
2.4.1 Content Found on the Web . . . . .	45

# CONTENTS

---

2.4.2	Information Collected While Searching . . . . .	47
2.4.3	Personalization . . . . .	50
2.4.4	Ethics and Personalization . . . . .	51
2.4.5	Biases, Behaviours and Beliefs . . . . .	52
2.5	Interactive Information Retrieval (IIR) . . . . .	54
2.5.1	IIR: The Holistic View . . . . .	55
2.5.2	Ethics of IIR . . . . .	57
2.5.3	Search Interface and Design . . . . .	58
2.5.4	IIR and Conversational Assistants . . . . .	60
2.5.5	Decision Making Processes . . . . .	61
2.6	Behavioural and Cognitive Interventions . . . . .	62
2.6.1	Nudging . . . . .	63
2.6.2	Boosting . . . . .	64
2.6.3	Nutrition Labels and Fact Boxes . . . . .	66
2.6.4	Comparing Nudging and Boosting . . . . .	66
2.7	Evaluation . . . . .	68
2.7.1	Evaluation Test Sets . . . . .	69
2.7.2	Evaluating the Search System . . . . .	71
2.7.3	Evaluating Interactions . . . . .	73
2.7.4	The Science of Evaluation . . . . .	76
2.8	Summary . . . . .	77
<b>3</b>	<b>A Harm Intervention Framework for Web Search</b>	<b>79</b>
3.1	Overview . . . . .	79
3.2	Components . . . . .	81
3.2.1	Policy . . . . .	82
3.2.2	Behavioural and Cognitive Interventions . . . . .	84
3.2.3	Search System Design . . . . .	85
3.2.4	Evaluation . . . . .	91
3.2.5	Framework Combined . . . . .	92
3.3	Examples in Practice . . . . .	93
3.4	Summary . . . . .	94
<b>II</b>	<b>Experiments</b>	<b>97</b>
<b>4</b>	<b>Methodology</b>	<b>99</b>

4.1	Overview . . . . .	99
4.1.1	Hypotheses for General Research Questions . . . . .	100
4.1.2	Adapting Methods from Previous Research . . . . .	102
4.1.3	Cochrane Systematic Medical Reviews . . . . .	103
4.2	Procedure . . . . .	107
4.2.1	Search Tasks . . . . .	107
4.2.2	Search Systems (and SERPs) . . . . .	107
4.2.3	Design for User Studies . . . . .	116
4.3	Evaluation Test Sets . . . . .	119
4.3.1	Web Page Privacy . . . . .	120
4.3.2	Web Page Misinformation . . . . .	122
4.3.3	Test Set Metadata . . . . .	123
4.3.4	Test Set Progression and Summary . . . . .	124
4.4	Evaluation Measures . . . . .	126
4.4.1	Harm Prevention (Search Task Outcome) . . . . .	127
4.4.2	Compliance to Transparent Strategies . . . . .	128
4.4.3	Search System . . . . .	129
4.4.4	Demographics & Self Report Measures . . . . .	131
4.5	Statistical Tests . . . . .	134
4.6	Participants and Recruitment . . . . .	135
4.7	Ethics . . . . .	136
4.8	Summary . . . . .	137
<b>5</b>	<b>Investigating Nudges in an Offline Setting</b>	<b>141</b>
5.1	Overview . . . . .	141
5.2	Method . . . . .	143
5.2.1	Procedure . . . . .	144
5.2.2	Evaluation Test Sets . . . . .	145
5.2.3	Evaluation Metrics . . . . .	148
5.2.4	Statistical Tests . . . . .	150
5.2.5	Participants . . . . .	150
5.3	Results . . . . .	151
5.3.1	Comparing 3 <i>Nudge</i> Strategies . . . . .	151
5.3.2	User Variations with a Transparent Strategy . . . . .	154
5.4	Discussion . . . . .	156
5.4.1	Findings Related to Study Specific Hypotheses . . . . .	156

## CONTENTS

---

5.4.2	Findings in the Context of Research Questions . . . . .	157
5.4.3	Lessons Learned . . . . .	159
5.5	Summary . . . . .	159
<b>6</b>	<b>Investigating Nudges in an Online Setting</b>	<b>161</b>
6.1	Overview . . . . .	161
6.2	Method . . . . .	164
6.2.1	Procedure . . . . .	164
6.2.2	Evaluation Test Sets . . . . .	165
6.2.3	Evaluation Metrics . . . . .	167
6.2.4	System Evaluation (Adapting IR Metrics) . . . . .	170
6.2.5	Statistical Tests . . . . .	176
6.2.6	Participants . . . . .	177
6.3	Results . . . . .	178
6.3.1	Annotations for Misinformation . . . . .	178
6.3.2	System Impacts on Privacy and Decisions . . . . .	179
6.3.3	Search Behaviour . . . . .	182
6.3.4	Search System Evaluation . . . . .	184
6.3.5	Strategy Preference . . . . .	191
6.3.6	Transparent Strategy Specific . . . . .	192
6.4	Discussion . . . . .	197
6.4.1	Findings Related to Study Specific Hypotheses . . . . .	197
6.4.2	Findings in the Context of Research Questions . . . . .	202
6.5	Summary . . . . .	205
<b>7</b>	<b>Useful Cues for Harm Prevention</b>	<b>209</b>
7.1	Overview . . . . .	209
7.2	Background, Motivation and Hypotheses . . . . .	211
7.2.1	Motivation and Choice of Features . . . . .	211
7.2.2	Research Question and Hypotheses . . . . .	215
7.3	Method . . . . .	217
7.3.1	Evaluation Test Sets . . . . .	217
7.3.2	Evaluation Metrics . . . . .	218
7.3.3	Statistical Tests . . . . .	219
7.4	Results . . . . .	220
7.4.1	Web Page Annotations . . . . .	220

7.4.2	Effects on Number of Trackers . . . . .	221
7.4.3	Effects on User Search Goals . . . . .	224
7.5	Discussion . . . . .	228
7.5.1	Limitations . . . . .	229
7.5.2	Conclusions . . . . .	231
7.6	Summary . . . . .	232
<b>8</b>	<b>Identifying Effective Boosts</b>	<b>235</b>
8.1	Overview . . . . .	235
8.2	Motivation and Hypotheses . . . . .	236
8.2.1	Motivation for Fact Boxes . . . . .	238
8.2.2	Hypotheses . . . . .	239
8.3	Method . . . . .	240
8.3.1	Procedure . . . . .	240
8.3.2	Evaluation Metrics . . . . .	247
8.3.3	Statistical Tests . . . . .	248
8.3.4	Participants . . . . .	249
8.4	Results . . . . .	249
8.4.1	Task Questions . . . . .	249
8.4.2	Post-Task Estimations . . . . .	250
8.5	Discussion . . . . .	251
8.5.1	Limitations . . . . .	252
8.5.2	Conclusions and Recommendations . . . . .	253
8.6	Summary . . . . .	254
<b>9</b>	<b>Boosting vs. Nudging</b>	<b>257</b>
9.1	Overview . . . . .	257
9.2	Method . . . . .	259
9.2.1	Evaluation Test Set . . . . .	259
9.2.2	Procedure . . . . .	261
9.2.3	Evaluation Metrics . . . . .	267
9.2.4	Statistical Tests . . . . .	269
9.2.5	Participants . . . . .	270
9.3	Results . . . . .	270
9.3.1	Main Study Results . . . . .	271
9.3.2	Evaluation of Knowledge Gained . . . . .	277

# CONTENTS

---

9.4	Discussion . . . . .	279
9.4.1	Findings Related to Study Specific Hypotheses . . . . .	280
9.4.2	Findings in the Context of Research Questions . . . . .	281
9.4.3	Limitations . . . . .	282
9.5	Summary . . . . .	284
<b>10</b>	<b>Conclusions</b>	<b>287</b>
10.1	Summary of Thesis . . . . .	287
10.2	Discussion . . . . .	291
10.2.1	Viability of Strategies ( <b>G-RQ-1</b> ) . . . . .	292
10.2.2	Effectiveness of Strategies ( <b>G-RQ-2</b> ) . . . . .	294
10.2.3	<i>Nudging</i> vs. <i>Boosting</i> ( <b>G-RQ-3</b> ) . . . . .	295
10.2.4	Validity of Studies . . . . .	296
10.2.5	Other Limitations . . . . .	297
10.3	Avenues for Future Work . . . . .	298
10.3.1	<i>Nudging</i> vs. <i>Boosting</i> . . . . .	298
10.3.2	Harms and Strategies . . . . .	299
10.3.3	Different Contexts . . . . .	300
10.4	Closing Remarks . . . . .	301
	<b>References</b>	<b>303</b>
<b>A</b>	<b>General Methodology (Appendices)</b>	<b>319</b>
A.1	Questionnaires and Scales . . . . .	319
A.1.1	Pre/post-task . . . . .	319
A.1.2	Privacy Attitudes . . . . .	320
A.1.3	Privacy Protective Behaviors / Actions . . . . .	320
A.2	Latin and Graeco-Latin Square Design . . . . .	322
A.3	Supplemental to Online Nudge Study . . . . .	324



# List of Tables

4.1	Cochrane Medical Search Tasks Used in Studies . . . . .	108
5.1	Privacy Impacts for the 3 <i>Nudge</i> Strategies in the Offline <i>Nudge</i> Study	152
5.2	Search Task Decisions for the 3 <i>Nudge</i> Strategies in the Offline <i>Nudge</i> Study . . . . .	153
5.3	User Preferred Strategies in Offline <i>Nudge</i> Study . . . . .	153
5.4	Self Report Privacy Attitudes and Protective Behaviours and Stop- light Strategy . . . . .	155
5.5	Summary Findings for the 3 <i>Nudge</i> Strategies in an Offline Setting .	158
6.1	Break Down of Result Misinformation Assessments for Results in On- line <i>Nudge</i> Study . . . . .	179
6.2	Online <i>Nudge</i> Study - Strategies Effects on Privacy Impacts . . . . .	180
6.3	Summary of Results for SERP Interventions on Medical Decisions . .	181
6.4	Results of <i>Nudge</i> Strategies Compared to the Control System on Harmful Medical Decisions . . . . .	181
6.5	Summary of Search Task Decisions by Intervention . . . . .	181
6.6	<i>Nudging</i> Impacts on Search Behaviour . . . . .	183
6.7	Analysis of Document Assessments by Quality of Information (e.g. misinformation) for each search system . . . . .	183
6.8	System Impacts (MRR) Comparing <i>Nudge</i> Strategies with Control System . . . . .	185
6.9	System Impacts (Precision) Comparing <i>Nudge</i> Strategies with the Control System . . . . .	186
6.10	System Impacts (Precision @ k) Comparing <i>Nudge</i> the Control System	187
6.11	Cumulative Probability of Different Click Types Across All Systems .	188
6.12	Model of Cumulative Probability of a <i>Correct</i> Assessment for Each <i>Nudge</i> Strategy Compared to the Control . . . . .	188

## LIST OF TABLES

---

6.13	Model of Cumulative Probability of a <i>Incorrect</i> Assessment for Each <i>Nudge</i> Strategy Compared to the Control . . . . .	189
6.14	Model of Cumulative Probability of a <i>Harmful</i> Assessment for Each <i>Nudge</i> Strategy Compared to the Control . . . . .	191
6.15	User Preferred Strategies in the Online <i>Nudge</i> Study . . . . .	192
6.16	Comparing Interactions with Stoplights in all <i>Nudge</i> Systems with the Control System . . . . .	193
6.17	Perceptions of Risk and Warning Light Colour in Stoplight <i>Nudge</i> . . . . .	194
6.18	Analysis of Self Report Privacy Measures and Compliance to Transparent Stoplight Strategy . . . . .	196
6.19	Summary Findings for the 3 <i>Nudge</i> Strategies in an Online Setting . . . . .	203
7.1	Independent Variables Used to Identify Features Indicative of Reduced Likelihood of Privacy Impact and / or Misinformation . . . . .	219
7.2	Dependent Variables Used to Identify Features Indicative of Reduced Likelihood of Privacy Impact and / or Misinformation . . . . .	220
7.3	Breakdown of Total Number of Web Pages by TLD Type in Test Set to Identify Features . . . . .	224
7.4	Results of Logistic Regression Models to Identify Features Useful for Harm Prevention . . . . .	225
7.5	Results (contingency tables) to Identify Features for Harm Prevention . . . . .	226
8.1	Piloting of <i>Boost</i> Strategies (Task 1) - Multiple Choice Questions . . . . .	246
8.2	Piloting of <i>Boost</i> Strategies (Task 2) - Multiple Choice Questions . . . . .	247
8.3	Questions Asked for Estimating Task to Test Knowledge Gained from <i>Boost</i> . . . . .	247
8.4	Main Analyses of Search Tasks and Fact Boxes . . . . .	250
8.5	Post-Hoc Analyses of Search Tasks and Fact Boxes . . . . .	250
8.6	Main Analyses to Test Fact Box Knowledge Estimation . . . . .	251
8.7	Summary of Fact Box Knowledge Estimation . . . . .	251
9.1	Summary of the ( <i>Boost</i> vs. <i>Nudge</i> Study) Evaluation Test Set . . . . .	261
9.2	Comparison of <i>Boost</i> vs. <i>Nudge</i> Strategies on Harms from Privacy . . . . .	273
9.3	Comparison of <i>Boost</i> vs. <i>Nudge</i> Strategies on Compliance to the Strategy (stick with .org and .gov TLDs) . . . . .	273
9.4	Comparison of <i>Boost</i> vs. <i>Nudge</i> Systems with $nDCB_p$ . . . . .	274
9.5	Comparison of <i>Boost</i> vs. <i>Nudge</i> Systems with $nDCH_p$ . . . . .	275

9.6	Results <i>Boost</i> vs. <i>Nudge</i> Post Experiment Multiple Choice Questions	277
9.7	Results <i>Boost</i> vs. <i>Nudge</i> Post-Experiment Estimation Questions . . .	278
9.8	Summary of Findings for the <i>Boost</i> and <i>Nudge</i> Systems Compared to the Control System . . . . .	279
A.1	Pre-Task and Post-Task Questions . . . . .	319
A.2	Attitudes Towards Privacy (General) . . . . .	321
A.3	Attitudes Towards Privacy (Health) . . . . .	321
A.4	Privacy Protective Behavior Questions (General) . . . . .	321
A.5	Privacy Protective Behavior Questions (Browser) . . . . .	322
A.6	Privacy Protective Behavior Questions (Search Engine) . . . . .	322
A.7	Latin Squares - Helpful Tasks . . . . .	323
A.8	Latin Squares - Not Helpful Tasks . . . . .	323
A.9	Latin Squares - System Treatments . . . . .	323
A.10	Graeco Latin Square Design - Task and Experimental Variants . . . .	324

## LIST OF TABLES

---

# List of Figures

2.1	Wilson’s Original Model of <i>Information Behaviour</i> [251]. . . . .	24
2.2	Dervin’s Sense-making Process . . . . .	27
2.3	Bates’ Berrypicking Model of Search [23] . . . . .	31
2.4	Overview of the Indexing Process of an IR System. . . . .	39
2.5	Overview of the Query Process of an IR System. . . . .	43
2.6	Ingwersen’s Cognitive Model of IR Interaction [102] . . . . .	55
2.7	Baeza-Yates Diagram of the Vicious Cycle of the Search Process [16] . . . . .	57
2.8	Kelly’s Continuum of IIR Research [119]. . . . .	74
3.1	Framework for Harm Prevention in Web Search . . . . .	93
4.1	Control Search System (SERP) with No Manipulation of Results and No Intervention for Harm Prevention. . . . .	111
4.2	Stoplight <i>Nudge</i> Search System (SERP) with Stoplights (best viewed in colour) Warning About Levels of Privacy Risks. . . . .	113
4.3	Filtering <i>Nudge</i> Search System (SERP) for Reduced Privacy Risk. . . . .	114
4.4	Re-Ranking <i>Nudge</i> Search System (SERP) for Reduced Privacy Risk. . . . .	115
4.5	Search Task Medical Decision Page. . . . .	116
4.6	Annotation Instructions as Presented to the Annotators for Classify- ing Web Pages as <i>Correct</i> or <i>Incorrect</i> . . . . .	124
4.7	Overview of All Empirical Studies. . . . .	139
5.1	SERP with Stoplight (best viewed in colour) Strategy Used in Current (offline) <i>Nudge</i> Study. . . . .	145
6.1	Probability of Assessments (@ rank k) of <i>Correct</i> Information by Search System . . . . .	189
6.2	Probability of Assessments (@ rank k) of <i>Incorrect</i> Information by Search System . . . . .	190

## LIST OF FIGURES

---

6.3	Probability of Assessments (@ rank k) of <i>Harmful</i> Information by Search System . . . . .	191
7.1	Examples of Modern SERPs Providing Visual Guidance for HTTPS .	212
8.1	Proposed Information Nutrition Label by Fuhr et al. [74] . . . . .	237
8.2	Two Fact Boxes Designed to Communicating Risk Associated with PSA Resting as Diagnostic for Prostate Cancer . . . . .	239
8.3	Large fact box to <i>boost</i> individuals with a skill to reduce privacy impacts. . . . .	242
8.4	Small fact box to <i>boost</i> individuals with a skill to reduce privacy impacts. . . . .	242
8.5	Prototype SERP to <i>boost</i> users with knowledge and skills to protect personal privacy during Web search. . . . .	243
8.6	Overview of the study design for piloting a fact box for boosting skills for better privacy. . . . .	244
9.1	Re-ranking Search System for <i>Nudging</i> Users Towards .org / .gov TLDs.	263
9.2	Search System with Fact Box to <i>Boost</i> Individuals with the Harms and Benefits of Results (based upon TLDs). . . . .	264
9.3	Overview of the Study Design for Comparing <i>Boost</i> and <i>Nudge</i> Strategies. . . . .	265
9.4	<i>Boost</i> vs. <i>Nudge</i> Scatter Plots for All Dependent Variables Evaluated	276
A.1	Interface Used to Ask Pre-task and Post-task Questions. . . . .	320

Part I

Theory





# Chapter 1

## Motivation

### 1.1 Harms from Web Search

It seems no better place than to begin this thesis with stories that demonstrate the potential harms of Web search.

*Story 1* Brenda is expecting the birth of her first child. In an online search for medical check-ups for children she finds information about vaccines. Exploring the topic in detail, the results she visits suggest a high risk for lifelong disability. She posts this information in a social media group for first-time parents, where many members respond with articles confirming the claim. This convinces Brenda (and others) to not have their children vaccinated. Years later, an outbreak of one of the diseases covered by the vaccine occurs. Unfortunately, Brenda's daughter and many other children become ill with long-term complications, including some who had been vaccinated.

*Story 2* Brian has felt sadness for some time and begins searching online about the problem. An advertisement for a well-being test appears as the first result in the search engine. Brian goes to the test, and as directed in the test, gives honest responses including suicidal thoughts. He gets help, as suggested by the website. This website shared his data with 3rd party companies, including his favourite social media site, whom years earlier allowed his insurance company access for reduced premiums. Now his insurance company will not give him life insurance and is tripling the price of his health insurance, placing him and his family in a challenging financial situation.

## 1. MOTIVATION

---

**Are these stories far fetched?** The fact of the matter is that, though these stories are entirely fictional, there is sufficient evidence demonstrating how they can happen, and in some cases already have happened.

Take *Story 1* for a start. It is established that a now rescinded scientific article about links between autism and the MMR vaccine is still regularly presented as fact in various forms across the Web [218]. Complicating matters further are the biased results of information retrieval platforms and the behaviours of the users themselves [179, 244]. Furthermore, the interactive data collected online allows companies to target individuals based upon their beliefs, including beliefs in pseudo science [195]. Specific to the anti-vaccine movement, there have been repeated cases of outbreaks for diseases once thought to be under control [218].

Turning to *Story 2*, recent evidence demonstrates how 3rd party companies learn about the mental health status of individuals [181]. A court decision in New York state now allows life insurance companies access to social media data [201]. In this example, it was the social media platform and the insurance company partnership that scored Brian as a higher risk. Partnerships like this already exist<sup>1</sup>, and governments (e.g. China) assign risk scores to their citizens [33].

**Common Themes to Harms from Web Search** Looking across these stories, there some common themes that stand out.

First, Web search is not specific to the search engine it self. In the examples, social media platforms play a role. Recommender systems, not mentioned in the stories, but also borne out of the IR community, are another approach to help the individual find information they might find important and / or interesting, but the

---

<sup>1</sup>Facebook partners with insurance at <https://www.theguardian.com/technology/2016/nov/02/admiral-to-price-car-insurance-based-on-facebook-posts> (LA: 2020-10-26)

user does not enter a query in the traditional form of search. The takeaway message is that Web search is broad, however for this thesis, we focus on the traditional query based Web search.

Second, harm from Web search does not necessarily happen to the individual and in many cases can impact social circles (e.g. families) and in some cases communities and broader society (as in *Story 1*).

Predominant in both of these examples is the interaction between the user and the system. In both stories, the searcher interacts with the system, the system collects the interactive data linked to the information they viewed. The underlying algorithms learn something about the searcher from their interactions, resulting in new information being served to the searcher (and potentially other searchers), and ultimately leading to unfortunate outcomes for many.

These stories only demonstrate some of the challenges.

Data collected across various platforms has been shown to radicalise views [230] and to alter outcome of elections [39]. Furthermore, there is some evidence to suggest that information online and the platforms used to find that information can lead to addiction [169]. The addiction problem is not surprising, given that some data scientists are assigned with the task of tuning the ranking and recommender algorithms for maximum engagement [52] and profit [267]. Tuning for engagement on search, video and social media platforms, also increases the carbon footprint for delivery of these services, a footprint that far exceeds air transport [136]<sup>2</sup>.

In summary, the potential harms can become quite dystopian [231], however cataloguing these harms is not the purpose of this thesis.

---

<sup>2</sup>It is currently estimated that delivery of YouTube alone uses the equivalent power of 1.7 million homes in the United States [136]

## 1. MOTIVATION

---

Instead, we wish to focus on methods that may in fact prevent them from occurring, or at the very least reduce the risk of their occurrence.

**How to prevent harms?** On a positive note, there has been growing focus (and pressure) in the IR community to develop algorithmic and scalable solutions to address potential harms, with research in areas including explainable AI and unbiased algorithms. A look at the paper sessions and workshops at SIGIR and SIGCHI over the past couple of years supports this trend. For instance, SIGIR 2019 had 3 workshops, SIGIR 2020 had 2 paper sessions and SIGCHI 2019 had 2 sessions entirely devoted to areas related to harm prevention in Web search environments (e.g. misinformation, privacy, bias, explainability, better decision making). However, we note that the large majority of the strategies published are algorithmically focused and give very limited focus on direct communication with the user about risks.

In commercial platforms (e.g. search engines and social media) where search for information is common, there are occasional actions taken to protect their users, such as content moderation and warning notices. Earlier this year, both Facebook<sup>3</sup> and Twitter<sup>4</sup> began providing notices about information that may be dubious. However, these notices appear to be quite selective given their focus on Covid-19. Furthermore, dependent on the nature of the information, they may employ a take down approach.

Such approaches, though well intentioned, do not provide an *opt-out* (or *opt-in*) mechanism for the user, and therefore have the potential for a ‘chilling effect’ that threatens freedom of expression and gives rise to possibilities of censorship [25, 26] and becomes an issue in the domain of human rights [221, 222]. These approaches

---

<sup>3</sup>Facebook Notice about Misinformation at <https://about.fb.com/news/2020/06/more-context-for-news-articles-and-other-content/> (LA: 2020-10-26)

<sup>4</sup>Twitter Misleading Information Policy at [https://blog.twitter.com/en\\_us/topics/product/2020/updating-our-approach-to-misleading-information.html](https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html) (LA: 2020-10-26)

also raise questions about the evaluation methodologies used, as demonstrated by earlier failed attempts by Facebook to combat misinformation<sup>5</sup>.

One counter example to the take-down approach in modern Web search is related to adult safe-search, where popular search engines in the English speaking countries (e.g. Google, Bing, Yahoo and Duck Duck Go) all include either an *opt-out* (or *opt-in*) mechanism for explicit content (e.g. profane language and pornography). In some cases, such as Bing, the default is to have some level protection from explicit content, and requires the user to *opt-out* if they want explicit results included. As this thesis will define in the later background section, the approach used by Bing is a *nudge*, and falls within the scope of behavioural and cognitive strategies, which are a main focus of the overall research goal.

*Nudging* is just one approach one might consider that gives protection, but still allows the searcher autonomy in their decisions. Another approach that will be the focus is known as *boosting*, which enables individuals with cognitive skills to more safely navigate environments. Behavioural and cognitive strategies such as these, are far different from the earlier mentioned approaches, such as content moderation (and removal) that is flagged by imperfect humans and algorithmic models<sup>6</sup>. They have the benefit of retaining freedom of choice, a matter important for human rights, and simultaneously addressing risks of harm, a matter important for both the individual Web searcher and more broadly to society.

---

<sup>5</sup>Facebook ditches fake news warning at <https://www.bbc.co.uk/news/technology-42438750> (LA: 2020-10-26)

<sup>6</sup>Identifying information that is harmful (and sometimes hateful) is challenging. For example, take the [Facebook Hate Speech Moderation Quiz](https://www.nytimes.com/interactive/2017/10/13/technology/facebook-hate-speech-quiz.html) at <https://www.nytimes.com/interactive/2017/10/13/technology/facebook-hate-speech-quiz.html> (LA: 2020-10-26)

### 1.2 Scope of Harms

Modern computing has enabled great advances in broad areas such as communication, commerce and entertainment. Email, online store fronts, video conferencing, social media platforms, travel booking websites and online banking are but a few examples of services now ubiquitous in our society made possible by computer science. Usage of any of these services, along with the devices they operate on, opens the door to harms to the individual and potentially to the entire globe (e.g. computer viruses).

Several demonstrating examples are provided in this section to highlight the broad scope of harms one might encounter in computing along with common approaches to prevent them. Following this, the harms this thesis aims to address are outlined.

**Computing and its Broad Set of Harms** The advent of email being made available to the broad population through services such as AOL and Outlook has provided the benefit of much faster and more affordable communication around the globe. Emails have the potential to contain malware (e.g. viruses) and/or spam (e.g. phishing scams), which are two common types of harms one may encounter online. Technology has advanced greatly in response to these issues, and problems such as these are regularly addressed by software that scans computers for viruses and machine learning models that classify messages as spam or not spam.

The development of online banking and online store fronts are technological advancements that greatly reduce the need for tedious activities such as writing cheques and travelling to stores to buy goods. Nonetheless, fraud is now a common enough harm that banks and online store fronts are moving towards the requirement

of technologies such as two-factor authentication [44]. Such technologies are also being applied to prevent password theft for accounts for a broad set of online services.

There is also the risk of information being stolen (e.g. via hacking) for a multitude of purposes, such as financial gain, blackmail or espionage. Such harms are seemingly more commonplace as more corporate and government data is housed in large data centres as time progresses. Technologies such as encryption play an important role in reducing the risk of these types of harm.

These are only some examples where harms may occur as a result of the technological environment in which our daily lives operate.

**Harms Relevant to this Thesis** In the context of this thesis, the focus is given to harms related to interactions with information in an information retrieval (IR) system during the process of search. The process of searching for information on the Web itself includes the risk of harms due to the various types of information collected about the searcher and found by the searcher.

Provided a set of results in an IR system a searcher will be faced with a multitude of actions (e.g. visit Web page X vs. Y, read a snippet, enter another query, read the Web page they visit), for which the actions may have different harms associated with them. In this thesis, the action of choosing a Web page to visit is given particular attention. This action is motivated by a need for information, therefore the decisions produced and pathways taken when the need for information has been satisfied are seen as highly important too.

For this thesis, the central goal is the evaluation of novel methods and IR systems designed for better privacy protection that simultaneously result in no increased exposure to misinformation and negative search outcomes. Therefore, we introduce

## 1. MOTIVATION

---

the following definitions of harm for which the underlying causes of these harms are introduced in Section 2.4:

- *Privacy Harm* - A significant loss of privacy for one Web search environment compared to another.
- *Misinformation Harm* - A significant increase to exposure of misinformation as a result of the Web search environment.
- *Search Task Harm* - A significant reduction in positive outcomes for a search task when comparing Web search environments used for the task.

*Privacy harm* is the main harm for which the interventions (IR systems) introduced in later chapters are tested and is therefore the harm most relevant to this thesis. Nonetheless, these harms are not exclusive from one another. For example, a *privacy harm* may result in an individual being served an advertisement containing misinformation (*misinformation harm*) or an entirely different searcher receiving misinformation (*misinformation harm*) in their search results due to the nature of collection search log data (*privacy harm*). There is also the issue of trade-offs to not be overlooked, where an intervention reduces one harm but increases one or more other harms. Ultimately, we view the *search task harm* as the most critical to avoid, and it is assumed that increases in the other two harms will greatly increase this harm.

### 1.3 Research Questions

Taken together, this leads to the main goals of the thesis: *to identify behavioural and cognitive strategies with proven success for harm reduction to individuals and*



*broader society in other domains and to evaluate them in the domain of Web search environment.*

Evaluation is a crucial aspect, given that an intervention designed to reduce risk of one particular type of harm (e.g. loss of privacy) may have the trade-off (or cost if you will) of some side-effect elsewhere in the search process (e.g. increased time to find information). There is always the potential that a harm prevention strategy produces significant negative side-effects (such as with Facebook’s warning sign), and one must therefore question the feasibility of a roll-out of such an approach in real-world applications.

Thus, we formulate the guiding questions this thesis aims to address (denoted as **G-RQ-#**).

**G-RQ-1** “*Are the proposed behavioural and cognitive strategies viable for prevention of harm during Web search?*”

**G-RQ-2** “*Which of the proposed behavioural and cognitive strategies are most effective at harm prevention?*”

As the thesis gives attention to two very different behavioural and cognitive strategies (*nudging* and *boosting*), there are theories that suggest the effects they will produce when compared with one another, but with little research performed to compare these approaches. This motivates a research question important for the application in Web search and for broader applications as well.

**G-RQ-3** “*To what extent do the claims about *boosting* compared to *nudging* exist within the experimental search environment?*”

### 1.4 Thesis Contributions

The work presented in this thesis has produced several novel contributions for the general research community, along with several notable publications.

**Theoretical Contributions** The framework introduced in Chapter 3 is a novel theoretical component that is threaded throughout the thesis. This theoretical framework is provided as a template for future research aimed at reducing harms in Web search.

Findings related to **G-RQ-3** (in Chapter 9) are used to test broader theories (beyond Web search) related to *nudging* and *boosting*, theories which claim that *boosting* will be overall more effective *nudging*. To date, there is no known published work that compares these paradigms to test this theory.

**Methodological Contributions** Though several studies of note (introduced in the background chapter) have evaluated behavioural and cognitive strategies for harm prevention in interactive search environments, none of these studies evaluate multiple strategies within the same study. The general methods used to perform comparisons in this thesis, along with the more detailed specifics for empirical studies, are seen as novel contributions.

Similarly, the behavioural and cognitive intervention strategies, which are based upon the risk prevention paradigms known as *nudging* and *boosting*. Though the designs used in the studies were influenced by previous research, they are novel methods as well that may be useful for future research.

Multiple evaluation measures are introduced throughout the general methods

chapter as well as the subsequent empirical chapters. Some of the novel measures were introduced to measure harms to the individual searcher during the process of Web search. Other novel measures take into account the attitudes and actions of the individual searcher (outside of the lab setting) towards the harm being prevented. Finally, several novel evaluation measures for system performance (which adapt existing IR metrics e.g. precision) are introduced as methods to compare across the search systems.

Methods in one study (Chapter 7) demonstrate how environmental cues for risks of harm can be extracted from log data. The general methods demonstrate how the cues identified here, or for any other type of harm in Web search, can be linked to multiple behavioural and cognitive approaches (Chapter 8 and 9).

**Evaluation Test Sets** Evaluation test sets from studies in Chapters 5 and 7 have been made publicly available in open data repositories mentioned within each publication.

**Published Works** The efforts on this thesis have produced multiple published works listed below. Links to full citations in the bibliography are provided.

The first empirical Chapter resulted in two separate papers.

An overall comparison of strategies in Chapter 5 and findings were presented at the *2019 Conference on Human Interaction and Information Retrieval (CHIIR '19)*. See [266].

Findings specific to user actions and concerns and their interactive behaviours with a warning light *nudge* (Chapter 5) were presented at the

## 1. MOTIVATION

---

*42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*. See [265].

Methods and findings in empirical Chapter 7 were presented at the *2020 Conference on Human Interaction and Information Retrieval (CHIIR '20)*. See [264].

Sections from the thesis background and framework (Chapters 2 and 3) were presented at the *SIGIR '20 Workshop for Bridging the Gap between Information Science, Information Retrieval and Data Science - BIRDS*. See [262].

Research performed in the earlier stages of thesis research resulted in two publications, are included here (but not part of the main thesis):

Much of the theory underlying the behavioural and cognitive approaches and motivations for their use in IR were presented at the *41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '18) Doctoral Consortium*. See [260].

Initial focus for this thesis was on system oriented solutions for detection of hate speech. It was during this research that the major gaps in interactive cognitive approaches were identified, result in a major shift in focus. Nonetheless, this earlier research of my thesis resulted in a neural ensemble approach for the identification of hate speech, and was presented at the *10th International Conference on Language Resources and Evaluation (LREC '18)*. See [261].

### 1.5 Presentation Style

Stylistically, the presentation of the thesis is inspired by David Maxwell's PhD thesis [148]. The following styles were used throughout this thesis to better help

guide the reader.

**Language** The thesis is written in British English, with a mixture of the active and passive voice. Though third person language (e.g. ‘we’, ‘our’) is regularly used, the thesis is single authorship.

**URLs** For some claims and demonstrating examples, external Web resources were utilized. When a web resource is introduced for the first time, such as a [hyperlink with embedded URL](#), a footnote<sup>7</sup> will include the full URL and date last accessed. For any subsequent uses of the same hyperlink, the full details are generally not provided.

### Hypotheses and Questions

**Example-1** is the style used to present hypotheses and research questions

**Document Navigation** References to key points (e.g. research question [Example-1](#) and style section 1.5) in the thesis may be clicked for fast navigation. Bibliographic references are also navigable (see [\[261\]](#) as an example).

**Evaluation Measures** Each time an evaluation measure is introduced for the first time within an empirical chapter, the measure is **emphasised like this**. Some measures are used in multiple empirical chapters, and are given **the same emphasis** when introduced for the first time in the specific empirical chapter.

---

<sup>7</sup>[hyperlink with embedded URL](#) at <http://dummy.url> (LA: 2020-10-31)

## 1. MOTIVATION

---

**Evaluation Test Sets** Evaluation test sets are commonly used in IR and interactive IR studies. Though different test sets are mentioned throughout the thesis, any test set used within this thesis is indicated like this: `example test set`.

**Strategies and Framework Concepts** Strategies introduced in the general methods (Chapter 4), and concepts for the framework introduced in Chapter 3 are associated with alpha numeric characters with a box drawn around them (e.g. strategy `S1` and framework component `FC-Policy`). This style is carried throughout the chapters that follow their introduction.

**Emphasis on Concepts** Important concepts central to this thesis, such as *nudging* and *boosting*, are *emphasised like this* throughout. Other important concepts, but less central to this thesis, are *emphasised like this*.

### 1.6 Thesis Structure

The thesis is divided into two parts.

**Part I - Theory** The background provided in Chapter 2, along with introduction Chapter 1, are used as motivation for a framework for harm prevention in Web search introduced in Chapter 3.

**Part II - Empirical Studies** The empirical studies, five in total (of which four are user based) are iterative in nature. Each study is performed in turn and used to inform later studies and allow for refinement of the general methods.

The general methodology is introduced in Chapter 4, which presents the harm

prevention strategies and introduces some of the search systems evaluated in later studies. General techniques used to evaluate the systems and strategies are introduced in this chapter, including considerations for study design, search task and type of harm to prevent.

*Nudging*, as a behavioural strategy, is the central focus of empirical studies presented in Chapters 5 and 6, of which both are user based. The first *nudge* study is performed in an offline environment, and compares three diverse systems and strategies against a control search system, with all systems connected to a static evaluation test set. The second study compares the same strategies, but is connected to a live dynamic search environment and is therefore online in nature.

Chapter 9 evaluates a set of environmental cues currently available in most commercial search engines. The cues were evaluated with respect to two different harms regularly discussed in the mainstream: (a) loss of privacy and (b) exposure to misinformation common in Web search. The findings from this study informed the direction of the final two studies.

Several *boosting* strategies are introduced and evaluated in Chapter 8 with data collected via a crowd sourced pilot study.

Using *boost* and *nudge* strategies found to be both viable and effective ( **G-RQ-1** and **G-RQ-2** ), Chapter 9 compares *nudging* and *boosting* strategies against a control system to answer **G-RQ-3** . This study is fully interactive, with an offline static test set similar to the first study, and performed with users in a lab like both *nudge* studies.

A discussion comparing the strategies, main study findings as well as limitations and directions to head for future research are detailed in the closing Chapter 10.

## 1. MOTIVATION

---



# Chapter 2

## Background

### 2.1 Overview

*Information retrieval*, as defined by Salton [194], *is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.* This field has produced many ubiquitous applications, such as search engines, that allow individuals to find information easily and for information to flow quickly through society. Though IR research has traditionally focused on the retrieval of text, other types of media fall into play as well, including images and video [51]. IR methods are applied in many different domains, that allow end users to more easily find the most relevant information, with examples including the entire open internet [170], an enterprise (such as a large corporate Intranet) [128] and commercial sites recommending different products (e.g. books to sell, movies to watch) [186].

Indeed, the field of IR and Web search is quite broad, and in general the background in the following sections is presented with this broad scope in mind. However, the empirical chapters presented later, for which this background forms the founda-

## 2. BACKGROUND

---

tion, focus on Web search in a traditional search engine and textual information.

This background chapter will introduce (Section 2.2) some of the core theories and concepts specific to the search processes, of which many are rooted in the research from information science (IS). Additionally, there will be an introduction (Section 2.3) to the basics of an information retrieval (IR) system and establish what is minimally necessary to create a modern search engine for the Web. Also, we will drill down extensively (Section 2.4) into the various sources of harm in Web search, such as system elements and human biases, which influence user decision making processes during Web search. An introduction to interactive information retrieval (IIR) is provided in Section 2.5. IIR a more recent area of research that considers both the IR system and the users of those systems as one together, and is the area of research most specific to this thesis. In Section 2.6, a brief survey of behavioural and cognitive interventions that promote decision making for reduced risk of harm is provided.

Throughout this chapter, a recurrent theme of the decision making processes inherent during Web search is intertwined. This is an important point to make, at the root of many intervention strategies to improve the Web search environment (e.g. a warning label about misinformation, an algorithm to provide faster results, log analysis for query recommendations, changes to snippets in the results page, etc. etc.), the decision making process will almost certainly be impacted. One must understand these impacts, as they certainly have the potential to have both positive and negative impacts, which motivates the attention given to commonly used evaluation approaches for IR and IIR research (in Section 2.7).

## 2.2 The Search Process

Advancements of modern information retrieval (IR) systems, such as commercial search engines and social media news feeds, have been guided by multiple views of the search process. Information seeking [251] and information foraging [177] are two examples; views which aim to model how humans search for and interact with information. These models have inspired the development of many automated methods in the field of IR to help meet an individual's information need [245], including automatic discovery of web documents [170] and automated assessment of relevance [108]. While the advancements inspired by these views have made positive changes for the world we live in, it is worth exploration of the underlying views guiding IR design as they provide insights into where risk may present itself in the search process. The following sub-sections highlight a sample of predominant views guiding IR design, with some examples framed within these views to highlight potential avenues for harm to individuals as well as society as whole.

### 2.2.1 Searching the Library

From a historical standpoint, IR design has been heavily influenced through research of the search process in the library setting. In his review of user interactions with search systems, White [245] states with respect to user *information seeking* behaviour:

...the reference librarian model (of a human search expert trying to satisfy a patron's information needs) remains the prevalent interaction model in many search systems, including commercial Web search engines. The primary difference is that in these systems human librarians have been largely replaced by automation in the retrieval process (including

## 2. BACKGROUND

---

formulating effective queries via tools such as query auto-completion, query suggestion, and backed query alterations), and by the searcher themselves (for example, in decisions regarding the relevance, filtering, and synthesis of the retrieved items) to generate a set of relevant information items, and ultimately, one or more answers to the questions that motivated their search.

**Example** Beginning with the reference librarian, they may be biased towards a subset of reference material, perhaps due to their own views about credibility of the sources. Over the years, the patron has established trust in their reference librarian, and this trust bias results in them believing they have the best information at hand. These biases of the patron and the librarian translate to risk for individuals and society as whole. Most immediately, the risk to the patron is that some relevant piece of information will be missed. Dependent on the information need of the patron, the potential loss (risk) due to this missed piece of information could be high (e.g. the missed piece of information contained the safest treatment for a medical issue). Perhaps this patron had the medical issue and is trying to find the best treatment. Alternatively, the patron could be a medical professional treating 100s of patients, some of whom have this medical issue in question. The patron is now a source of risk to greater society (e.g. some of the patrons patients end up dying) as they are yet aware of the safest treatment available. As a medical professional, the patron's reputation is also potentially at risk (e.g. word might spread about prescription of unsafe treatments).

This toy example is intended to demonstrate the potential harms due to the biases and beliefs (biases and beliefs and their role in search are further detailed in Section [2.4.5](#)) of actors in the process of searching for information, and it further

demonstrates that harms might occur regardless of the search environment (be it a library or search engine). Additionally, it is plausible that neither the librarian or the patron were aware of their biases. In present times, the role of the librarian is more often replaced by automated IR systems. Thus, to minimize risk to the patron (the searcher) and broader society via their contacts, interventions should address the biases of the librarian (IR system) and the patron (the searcher). In summary, this example demonstrates risks of harms in the library, which apply to Web search as well.

### 2.2.2 Information Behaviour and Information Seeking

Multiple models of *information seeking* and *information search* have been proposed, including Maxwell's recent model of stopping behaviour [148]. Wilson's general models are discussed broadly in literature and are quite useful for the demonstration of many of the issues with Web search.

*Information seeking* behaviour is framed within Wilson's nested general model [253] of *information behaviour*, which combined together encapsulate *information search* behaviour. Using this nested model, Wilson [254] defined *information behaviour* as

... the purposive seeking for information as a consequence of a need to satisfy some goal.

and *information search* as

... the 'micro-level' of behaviour employed by the searcher in interacting with information systems of all kinds.



These ‘individuals’ continue to interact with the IR system until their search is successful and information need is satisfied. For individual 1, based on the information viewed, they are now convinced that benzodiazepines are the most effective treatment. Individual 2, based on the information they encountered, are now quite certain that going cold turkey and sitting out the withdrawal is the best treatment. Each of these treatments comes with the possibility for harm, of which the second treatment is quite dangerous (e.g. due to inherent risk of death from convulsions). At this stage, the risk is only to each of these individuals. However, one or both of these individuals could share this information with another individual (‘Other people’ in Wilson’s model), at which point the potential for societal harm becomes a concern.

In addition to highlighting the potential for risk of harm to individuals and society<sup>1</sup>, this example is useful to identify other concerns. First, understanding of a user’s *information need* is not sufficient for harm prevention. Even though both individuals had the exact same *information need*, the outcome of their seeking behaviour was very different. Biases of the information system or the individuals using them are factors [16, 244] that may have played a role in this outcome (see Section 2.4.5). Second, the example suggests the value and importance of evaluation of successes and failures (both highlighted in Wilson’s model) of the seeking process. In the example provided, the information was relevant and satisfactory for both individuals, but this information produced two very different outcomes. Therefore, commonly used approaches (e.g. TREC evaluation and graded relevance) to evaluate modern IR systems and the information seeking process are not enough. A more recent proposal, evaluation of search outcomes (e.g. positive health outcome after

---

<sup>1</sup>Wilson’s later model of *information behaviour* [253], does consider an element of risk / reward in the search process for consideration of information sources, and is a rare example within a model of the search process where risk is explicitly considered. However, this consideration appears to be from the view of risk to the individual and not broader society.

information seeking is complete) as a measure of success [244], is one possible approach to address the discrepancies of success for the two individuals. Evaluation is covered further in Section 3.2.4, and it is an important consideration for IR systems and Web search.

### 2.2.3 How Information is Used

Beyond the views and models of how individuals search for information, it is important to understand how information is used. White [245] breaks down the usage of information into four areas: sense-making, exploratory search, problem solving and creativity. We focus on the first two, as sense-making and exploratory search are often a part of the decision making process [245], which implies the potential for consequences (including harmful outcomes) beyond completion of the search process. Both exploratory search and sense-making are important areas related to search [56, 145, 178, 245, 247], for which we can only scratch the surface of their importance; we begin with Dervin's view of sense-making.

**Sense-making** Dervin's model of sense-making (see Figure 2.2) is an intuitive view of the processes involved to make sense of information [56], for which she uses the metaphor of a bridge being built across gaps. These gaps represent voids in knowledge that an individual has come to recognize [178], for which gap recognition has many possible triggers [56] (e.g. individual does not know an answer to a particular question, uncertainty about dietary choices and health implications). Sense-making applies to many different types of information, including television and virtual environments in gaming [185], however the current concern remains within documents found on the Web.

The process of sense-making is two way, where an individual's reality about a



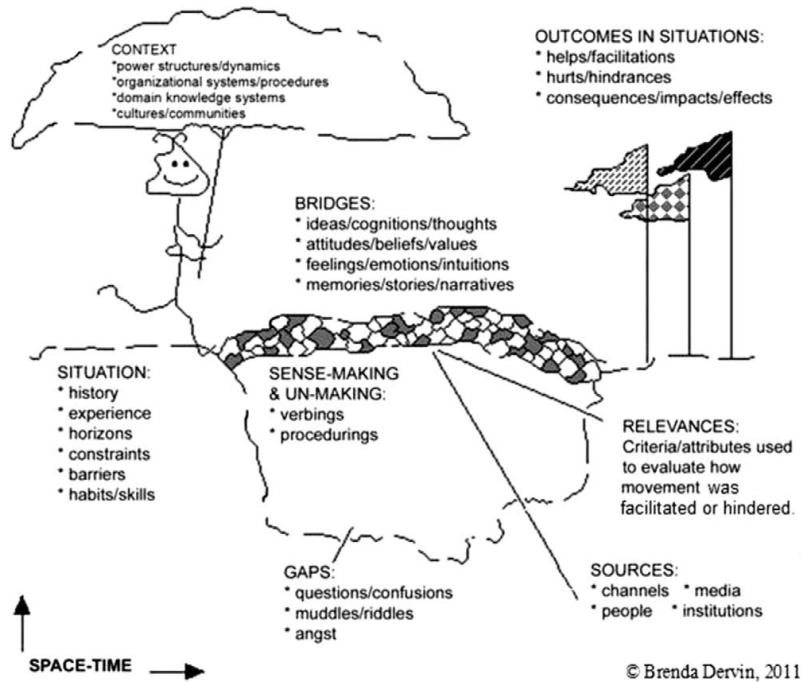


Figure 2.2: Dervin's Sense-making Process - Figure from Reinhard and Dervin [185]

knowledge gap can be bridged and then un-bridged [56] (e.g. through realization that knowledge they had was incorrect). It is a continuous process of framing hypotheses based on the data, and placing appropriate data to fit within a hypothesis [178, 245]. In many cases, a decision will be made using the knowledge gained through this process [56, 178, 245].

Pirolli and Card [178] suggest that ingrained beliefs and biases (e.g. confirmation bias) play an important role in the sense-making process and ultimately the decision making process. They furthermore suggest that biases are one possible avenue for technological innovation. When considering the sensemaking process of connected individuals [245], it is also possible these biases are influenced by our cultural and societal networks

Sense-making as a whole, undoubtedly has elements of decision making for the individual. In Dervin's model (see Figure 2.2), it is clear that outcomes are a part

## 2. BACKGROUND

---

of the sense-making process, outcomes which may impact the user positively or negatively [56, 185]. This view suggests that outcomes of the search process are an important consideration when designing systems to support Web search.

**Exploratory Search** Marchionini, a pioneer in the field of exploratory search, highlights that this type of search is fundamental to human behaviour, yet there are limited tools available to support this important process on the Web [145]. Compared to other views of the search process, there is quite limited research producing formal models of exploratory search, however more formal models may be helpful [245].

White [245] frames exploratory search as a contextual problem that may be open-ended and long lasting as well as a process that can be repeatable and opportunistic, and may be used in the broader decision making process. Overall, the goal of exploratory search is one of two outcomes: a "knowledge product" (e.g. research paper) or underpin a future action (e.g. choosing a holiday destination, choosing a medical treatment). Unlike the knowledge gap in sense-making, which implies a clearly defined goal, the ultimate goal of exploratory search is much less defined. The process of exploratory search may be driven by curiosity and often takes place across multiple search sessions and temporal scales of days, weeks and months. Perhaps most important of all, is the embedded element of learning. On the topic of learning White states:

Exploratory searches can have a profound impact on searchers' personal development, because they reflect the pursuit of higher-level learning objectives.

This potential for profound impacts through the process of exploratory search

[145, 245, 247], is certainly a potential pathway for harm in Web search.

**Implications** Both views of information usage discussed (sense-making and exploratory search) suggest possible implications of the information encountered by the searcher, implications that could be both negative or positive to the individual searcher, their direct social network and / or more broadly to society as whole. Though developing models that identify and mimic such search behaviour (e.g. models that identify search tasks that are exploratory [121]) are beyond the goals of the current research, such models are indeed a promising direction when it comes to potential harm reduction strategies for Web search.

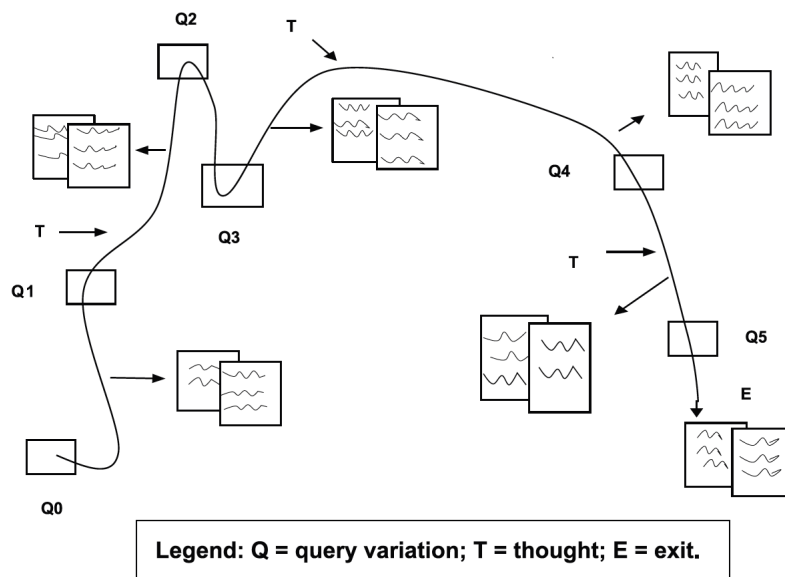
Combining such models with topical models of the search space, an intelligent search system may be able to intervene and steer a user towards content that still gives broad perspectives but having a goal that minimizes potential harm. Topical areas where this may be relevant include those related to health and medicine as well as topics with a high likelihood of addictive behaviour, with other topics one might consider too. For instance, if the search system can identify the search behaviour as sense-making of a particular medical treatment, the system could filter results that are linked to low risk outcomes. Alternatively, imagine a teenager, who is quite curious and susceptible to a "profound impact" to their development. If this teenager is performing exploratory search for information in the topical space of sex and sexuality (or drugs, weapons, etc.), there are many impacts that might occur and it is therefore considering interventions for harm prevention. These are but two illustrative examples, which provide motivation for further development of such models, in addition to motivation for identification and evaluation of more immediate interventions and strategies.

### 2.2.4 Evolutionary and Ecological Views of Search

Ecological and evolutionary models, that model animal behaviour and the search for food and other resources, have been adapted to describe and model searcher behaviour and are used as a guide for researchers to design IR systems. Here we briefly introduce two such frameworks, followed with real world examples where these views break down.

**Berrypicking** Bates' Berrypicking model [23] (see Figure 2.3) present the searcher looking for information as an analogy of someone looking for berries, and the berries as information. Bates' suggests that searching is a non-linear sequence of steps as opposed to a process that results in the perfect set of relevant documents and information. She suggests that search should be modeled by the behaviour of searching for berries in a forest, and includes the notion of information patches as an analogue to the berry patch. The non-linearity of the search is the paths taken between different information patches (just like one would do while picking berries in the wild). Eventually the searcher has all the information necessary for their *information need* (or enough berries to eat) and the search will terminate. As Figure 2.3 demonstrates, the path for information (just as the path through a forest) may be all over the place.

**Foraging** Information Foraging Theory (IFT) introduced by Pirolli and Card [177] presents the searcher as a predator and the information as prey. In addition to the predator prey view, IFT makes important links back to earlier models of rational decision making [205], models demonstrating that much of our decision making processes in our quest for information are rational and adaptive to the environment. The pioneering work by Pirolli and Card also suggests that information environments



**Figure 2.3: Bates' Berrypicking Model of Search [23]** - Bates' provides the following description: "The continuity represented by the line of the arrow is the continuity of a single human being moving through many actions toward a general goal of a satisfactory completion of research related to an information need. The changes in direction of the arrow illustrate the changes of an evolving search as the individual follows up various leads and shifts in thinking. The diagram also shows documents and information being produced from the search at many points along the way."

## 2. BACKGROUND

---

can be further optimized for the searchers that use them. Pirolli and Card focus heavily on the costs of search, particularly time, and introduces the notion of ‘gain’ (information gained) as part of the search process and therefore one key bridge to the economic view of search and IR (see Section 2.2.5).

In a general sense, when considering both Berrypicking and IFT, both are inspired by theories of evolution and ecology and both treat information as prey and the searcher as the predator [177, 198]. However, both neglect the fact that hunter-gathering and foraging for food comes with many additional risks [135], including the possibility of death from toxins in the food or being attacked by predators during the process of searching for food. Looking again at Bates’ model (see Figure 2.3), the model suggests that thoughts can (and will be) changed due to the information picked during the process. When re-framing this model with the addition of toxic information (or berries if you will), the model has a threat of harm. When taken with this view, both views have a lack of consideration for external dangers such as toxic information and predatory behaviour, therefore flaws in models that aim to describe search behaviour. There are well publicized recent events in Web search demonstrating where these models might break down.

**Examples** Two recent real world examples demonstrate the threat of harm both to individuals and society due to foraging on toxic information and being attacked by predators. Toxic information, such as that related to a retracted scientific paper on the links between MMR vaccines and autism, has resulted in measles outbreaks due to loss of herd immunity [218]. Predation occurred in 2016 when Cambridge Analytica [39] made use of personal data collected to target individuals with information and advertisements containing dubious information, and potentially impacted society via outcomes of voting.

These examples are just some to consider as critiques of IFT and Berrypicking and potential avenues for individual and societal harm, there are additional possibilities and shortcomings to touch upon briefly. IFT was originally designed with the assumption that outcomes are certain, and even though uncertainty may be a desirable outcome of the process of information foraging [34], and thus has limited consideration for risk [177]. Furthermore, IFT focuses heavily on information scent [177], which is challenging, as there is plenty of information available that may smell quite good to the searcher but is in fact information that is analogous to junk food (e.g. click bait [42]).

No model is perfect, and the shortcomings discussed are likely due to the idealized views of the search process and search environment used as assumptions within these models. They are nonetheless important views of the search process that have guided advancements in modelling search behaviour, most notably the economic view.

### 2.2.5 Economic View

Recently developed models of search behaviour integrate theories from economics, including microeconomic concepts including consumer and production theory, with the overall aim to better predict user behaviour in the IR environment [11]. Incorporation of the economic view of IR is quite promising, in that it opens the doors for researchers to evaluate systems and their users from the perspective of trade-offs of costs and benefits, in addition to traditional relevance based metrics [11, 14]. Cost is not a new consideration for IR systems (e.g. Pirolli and Card include this in IFT [177]), but there are limited considerations for it in practice [13]. Unfortunately, studies around economic models and evaluation metrics thus far appear to be heavily focused on information gain (as measure of benefit) and the costs of time (see Azzopardi et al [13] for example). In the decision making process,

## 2. BACKGROUND

---

elements of risk and uncertainty are costs as well [135, 143], and thus should also be considered in the economic view. The gain based measures (e.g. [106]) used to test economic frequently use graded relevance (e.g. TREC test collections), which at the end of the day is still relevance and therefore neglects the possibility that a document graded highly relevant may also contain harmful information. There is the added challenge that economic models of IR behaviour are sometimes designed with the assumption that users are rational (e.g. [14]), for which ample evidence exists is not always the case [115, 233].

One can adapt the examples already provided to demonstrate challenges with the existing economic views, therefore we turn our attention to considerations for the cost and gain components for harm prevention which may someday be incorporated into economic models of search. Costs that might be considered in the frame of these models when considering harms are personal data collected (e.g. the information collected about the searcher) as a cost, the monetary costs to access information (e.g. the price a news website or scientific journal charges for access to an article in search results). With respect to the gain component, document relevance and document misinformation are not mutually exclusive. Evaluations that consider relevance, are just one view to consider in economic modelling and it may be wise to improve upon graded relevance to somehow take into account that dangerous information may also be relevant. Health search is one particular area of concerns where dangerous information may also be relevant to the search [179, 244].

Economic models are undoubtedly useful for evaluation of interventions such as query suggestions to understand expected impacts to costs of time and gains of information [13]. As interventions are particularly relevant for addressing harms in Web search (see Sections 3.2), it may be possible to incorporate such approaches into these models.



### 2.2.6 Concepts

In the views considered, several important concepts have emerged which are commonly used, including *information need*, *task*, *search outcome* and *relevance*. Each of these concepts touch upon one or more categories of consideration.

*Relevance* is a fundamental concept in IR research and is a major consideration for the development and evaluation of IR systems. The notion of relevance applies to queries in IR systems for books in a library and documents on the Web (e.g. via Google), and may also be extended to other types of IR systems including news personalization and product recommender systems. When considering definitions, including those that define a document as *relevant* if it contains the information the user was looking for when they submitted a query [51], one may get the sense the concept of *relevance* is somewhat ambiguous. There are different types and categories of *relevance* which make this concept less ambiguous. Types and categorizations of relevance include, but are not limited to the following: topical relevance (e.g. for the query ‘baseball’ documents about baseball are topically relevant, whereas documents about basketball are not), temporal relevance (e.g. documents about last year’s World Series are temporally more relevant than the 1984 world series for a query ‘MLB World Series’) and locality (e.g. Chinese restaurants in London are likely much more relevant than restaurants in Seattle for a search query ‘Chinese food’ sent from a mobile device in London). Many, if not all, of the categories of *relevance* are factors playing into the overall notion to of *relevant* information for the users.

The *search outcome*, as it relates to the decisions made by the user during and after their search for information, may result in good and / or *harmful* implications. This concept of *search outcome* falls within the scope of performance and outcome

## 2. BACKGROUND

---

based evaluation measures [119, 245], which are touched upon in section 2.7.3. The points made in the previous models of search behaviour suggest that *relevance* is not sufficient for evaluation of IR systems with respect to *search outcome*, as *relevance* does not consider what the user actually does with this information, and we note that Dervin’s model of sense-making suggests different categories of *search outcome* during the search process (see upper right of Figure 2.2).

*Task* in and of itself is threaded throughout all of the views above. There are many different *tasks* a user may perform, and the *tasks* may vary by environment (e.g. at work vs. at home).

Kim [121] provides an extensive review of *tasks*, and a thorough taxonomy to categorize different attributes of *task*. Broadly, Kim groups these attributes into Search and Work *task*. She focuses on 3 types of *task*: factual, interpretive and exploratory, which are respectively defined as closed, open-ended but focused, and completely open-ended. Kim’s research is motivated by the need for better models of information seeking.

However, Kim does not provide the only taxonomy. Wildemuth and Freund [248] in their systematic review<sup>2</sup> of task broadly break them into complex versus simple, exploratory versus lookup and navigational, information or transactional (as suggested by Broder. [36]), indicating there are multiple ways one might describe a search task.

*Information need* (and *search intent*) is threaded throughout all of the views and models discussed in the previous sections. Ultimately, it is the need for information, whether the need is clearly defined or not, that drives a user to perform a search.

---

<sup>2</sup>Their review also provides a database of search tasks published in IIR literature before 2017 in the frame of this taxonomy. See [Systematic Review of Assigned Search Tasks](https://ils.unc.edu/searchtasks/index.php) at <https://ils.unc.edu/searchtasks/index.php> (LA: 2020-10-30)

There are many ways one might categorize *information needs* of the user, with the search taxonomy produced by Broder [36] being a quite useful approach to do so. Broder proposes, and provides some evidence, that queries (and the underlying *information needs*) are either navigational (e.g. searching for a particular website or information that is likely or known to exist), informational (e.g. reading about a particular medical issue) or transactional types where the intent is to reach a site for additional interactions (e.g. shopping for books). The Taxonomy introduced by Broder launched additional research in this area to describe and understand a user's search intent.

A more refined description of Broder's three categories was produced by the work of Jansen et al. [105], and validated with a analysis of commercial search logs. In their work, it was found that the large majority of queries are informational in nature, and much less often navigational or transactional (counter to Broder). Jansen et al. also treat intent on three different levels, which is perhaps a more intuitive and specific way to describe searcher intent than Broder's initial definitions.

There are no direct critiques to make about these concepts , however it is worth noting that documents are readily available on the Web (see [76]) that leverage these concepts (such as Broder's categorization of *information need*) that provide advice on the topic of Search Engine Optimization (SEO). Authors of content, especially content with commercial value, know that it is beneficial to have their results rank high and therefore make use SEO techniques [168] to increase the likelihood of monetary conversion (such as display advertising or sale of a product). SEO is a broad topic in and of itself, but is just one example where publishers of Web content take effort to adapt to the ever changing Web search environment. Another example of adaptation was to Google's original Page Rank algorithm [170], where authors intentionally added in-links and out-links to their pages to rank more highly

## 2. BACKGROUND

---

results. Relating these examples back to the earlier discussion on IFT, Pirolli and Card suggest that IFT is intended to predict and understand how individuals adapt to information environments and understand how environments can be adapted to individuals [177]. As such, researchers might consider not just individuals as the only agents interacting with information environments, but other agents as well (e.g. authors, organisations, artificial bots).

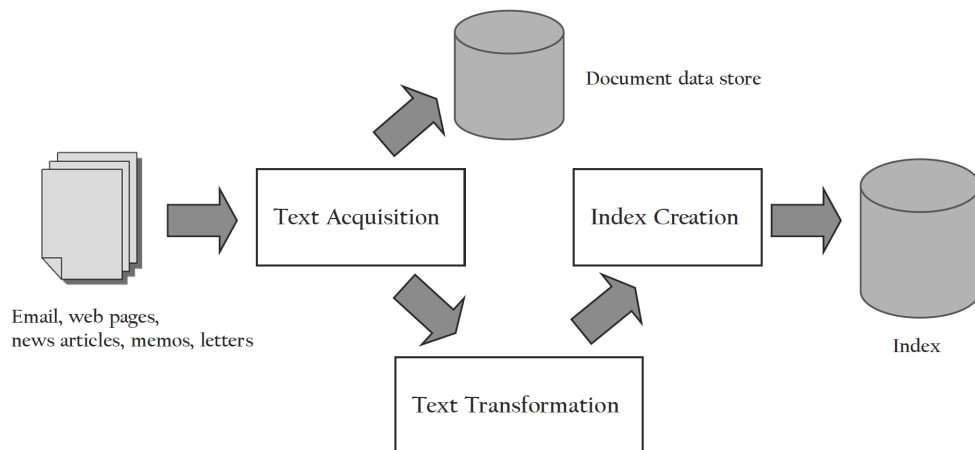
### 2.3 Information Retrieval System

As so much of Web search occurs with search systems such as commercial search engines, it is worth taking a look at the main components of such systems. The field of Information Retrieval (IR) and the systems produced as outputs of the field is a quite large and technical subject in and of itself. Search engines (e.g. library book search, Google) are more traditional flavours of IR systems based upon the lookup model of search [245], with newer IR systems incorporating social information [51]. The current research focus is not about detailing the algorithms used in these systems, however a basic understanding of the main components and concepts is helpful for the subsequent background regarding potential harms.

Fundamental elements for IR systems include document pre-processing, ranking and filtering as well as representing information in a retrieval model to determine relevance. Croft et al. [51] provide a general overview of search engine architecture, and split search systems into two high-level processes that collectively make up a search system. There is the indexing process (see Figure 2.4) and the query process (see Figure 2.5). Splitting the description in this manner is just one way to introduce the key basics of search systems, basics which are common to many introductory IR resources (e.g. [17, 51, 144]).

It is also important to keep in mind that IR systems related to retrieval of semi-structured (e.g. HTML) and unstructured (e.g. raw text files) data are the central focus of this introduction, as opposed to systems that retrieve information from structured data (e.g. relational databases). Also, for this introduction, *content* and *document* are generic terms used to describe the information stored in retrieval systems (e.g. text, images, video), however many popular resources (e.g. Croft et al. [51]) focus on the retrieval of *text*, thus all three terms (*content*, *document* and *text*) are treated interchangeably. Some examples of other document types are provided, however our focus is on *text*.

### 2.3.1 Indexing Process



**Figure 2.4: Overview of the Indexing Process of an IR System.** - As presented by Croft et al. [51]

The primary task of the index process is to ensure query results that are provided to the user are returned efficiently. As users are accustomed to retrieving information in short order <sup>3</sup>, there are indeed very efficient mechanisms underlying any large scale IR system. The output of this process is an inverted index, which is a highly

<sup>3</sup>Google Search typically returns query results in less than half a second on billions of documents

## 2. BACKGROUND

---

efficient data structure that allows queries (via a *retrieval model*) to return candidate documents quickly [51].

The index is a representation of a set of documents (*corpus*), which in the case of modern search systems, is continually updated via a content-acquisition sub-process commonly known as *crawling*. Once a document (e.g. an html Web page) is captured in a crawl, it goes through a transformation sub-process to greatly simplify the document representation, where in the case of textual documents, techniques from *natural language processing (NLP)* are commonly used to cleanse and simplify the text (e.g. removal of common words). Only at the end of this process (or pipeline, if you will), are the documents placed into the inverted index, which is accessed via querying of a *retrieval model*. The sub-sections that follow provide a bit of detail about the acquisition, transformation and models used to retrieve documents in the index.

### 2.3.1.1 Content Acquisition

*Crawling* is the process of discovering new, updated and removal of content across the Web [51, 144], and is a massive task when one considers the amount of information available on the open Web and the factors that must be taken into account during the crawl. Some factors to consider are respect of hosts of Web pages (i.e. a host says I do not want to be included in your crawl) and weighting of higher quality Web sites (e.g. re-crawl high quality news sites regularly vs. spam sites much less so) [144]. There are also factors related to coping with scalability of the problem, such as deciding where to start crawls (often with a seed set) and how often to ensure that documents retrieved are up to date (known as freshness) [51].

It is useful to note that just as websites can be configured to block crawlers, crawlers can be configured to not visit certain sites. At the end of the day, a search

system can only return results for sites that have been crawled and included in the index. This is analogous to a library in that you have no chance of finding a book if the library has made the decision to not purchase it, or the seller refuses to let the library purchase a copy.

### 2.3.1.2 Content Transformation

Content transformation, is the sub-process to convert crawled documents into a condensed representation for the document index. A common approach is to pass textual documents through an NLP pipeline that first tokenises them (i.e. splits up texts into individual words and punctuation) before normalising the tokens with steps such as stemming, lemmatisation and removal of stop words [111, 144].

Stemming, using algorithms such as the Porter stemmer [180], removes suffixes from words (e.g. ‘Likes’, ‘liked’, ‘likely’ and ‘liking’ all become ‘like’).

Similarly, Lemmatisation (e.g. ‘better’ is treated as ‘good’) is also useful for condensing words, and is a linguistic based process known as morphological analysis [144].

Stopword removal is the process of stripping low value terms (e.g. words having little semantic meaning such as ‘be’, ‘not’, ‘or’, ‘to’) [144], of which one must create additional approaches to circumvent issues with phrases such as “to be or not to be” [51].

The examples of content transformation presented here are but a few, and the transformation process and related tasks (such as document classification and cue extraction) will be revisited in sections and Chapters ahead. Content transformation and it’s relation to harms on the Web are introduced Section 2.4.1. The evaluation of document collections is an related and important action introduced in 2.7.1. Extrac-

## 2. BACKGROUND

---

tion of information from content (such as cues) is yet another under the umbrella of content transformation and is central to the harm prevention framework introduced in Chapter 3.

### 2.3.1.3 Content Retrieval Model

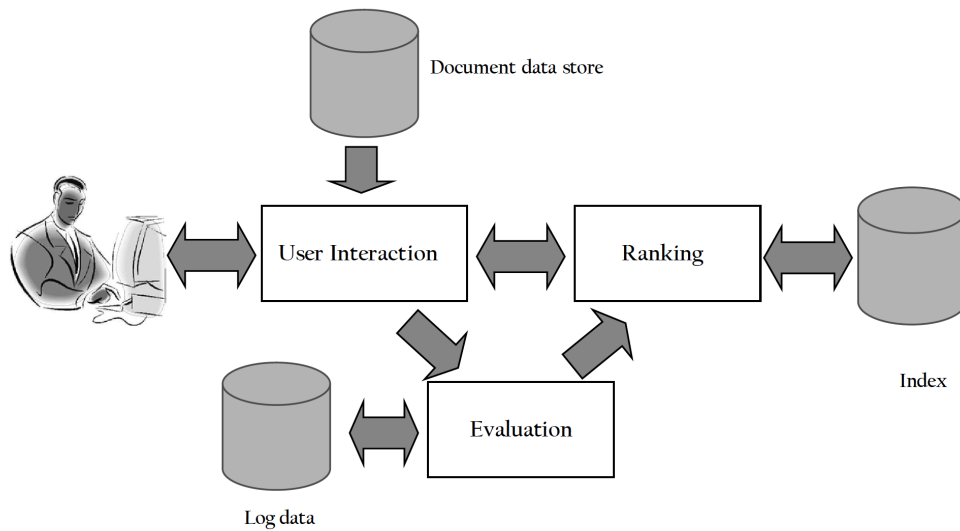
Determination of potentially relevant documents is commonly achieved through development of a retrieval model. A simple vector space model is achieved through term weighting measures such as term frequency and inverse document frequency (TF-IDF), where relevance could be a measure of cosine similarity between documents [51]. BM-25 is a probabilistic approach assuming binary independences (e.g. there is a set of relevant and non-relevant documents) that adds in the term weighting components [144], which is a more recent advance. Usage of representations output from machine learning models, such as topic and embedding models are additional variants one should consider [51]. Recent evidence suggests that neural approaches are quite competitive with a BM-25 model but simultaneously should not be considered as the end all be all solution [155, 156].

### 2.3.2 Query Process

Focus now turns to the query process (Figure 2.5). In the most general sense, the user interaction is the fundamental element of the overall query process and consists of a query input, query transformation to represent the query in a form similar to documents in the index and an interface component that displays output of one more results and captures the user interactions (e.g. clicks on a result).

Ranking, which is the process of determining if document B is more relevant than document A, requires many inputs which might include features of documents and query, time of day, search history of user, topicality of document [51, 144] or





**Figure 2.5: Overview of the Query Process of an IR System.** - As presented by Croft et al. [51]

specifically links between documents [170]. Learning to rank algorithms (as covered in [139]) take different features from the documents as well as logging behaviour of the users to more effectively rank documents over time, and such algorithms are fundamental components of modern search systems. It is worth noting that details of ranking algorithms are often protected secrets of platforms such as Google and Facebook, though it is widely known that Facebook ranks information for user news feeds with an internal algorithm known as EdgeRank, little information has been published about how it works [122].

The evaluation component of the query process, is perhaps most important of all to any IR system, as proper evaluation allows system designers to test new changes, identify problems with a system and ultimately improve the user experience [51, 144]. Much can be said about evaluation approaches and there are many different ways one can evaluate search systems (e.g. precision, recall, mean reciprocal rank (MRR)), therefore evaluation is given particular focus (see Section 3.2.4).

Finally, there is the user interaction itself, and is a topic that is central to this

## 2. BACKGROUND

---

thesis. In fact, the topic has sprung an entire area of research known as interactive information retrieval (IIR) and is given special attention in Section 2.5. This process includes logging of user interactions is performed to serve many purposes (e.g. personalization, query recommendation, spelling correction), and is a topic covered later (Sections 2.4.2 and 2.4.4).

### 2.4 Sources of Harm

The potential harms of Web search have many underlying causes, for which we emphasize the key factors. First, corporate (IR) platform policy can encourage data scientists to create addictive systems [52] and environments that are non-transparent to searchers<sup>4</sup> [125, 140]. There are also biases existent in both the system [16, 244] and the user [108, 164, 244], and the reinforcement factor [16, 18, 230].

Information itself is a factor, which comes in two forms. One form is the information found (e.g. search results, Web pages, videos, social media post), and the other form is information collected about the searcher (e.g. queries, IP address, usernames) – both forms offer many benefits to searchers such as exposure to more relevant information and faster task completion [245], but simultaneously may contain harmful content [244] and cost them their privacy [211, 245]. The centralized nature of search environments [258, 259], which are now commonplace, were built upon IS models of search developed and evaluated in quite different environments, such as the library [245, 259], and is another factor likely playing into concerns around privacy. The motive of profit [267] certainly encourages searchers to view information they might not otherwise do [231]. Moreover, the metrics used to evaluate systems (a focus of data scientists and IR researchers [256]) are quite different to

---

<sup>4</sup>For example, to understand privacy impacts, searchers must read lengthy privacy policy statements, written in an obfuscated manner.

those suggested by IS researchers (e.g. [56]). This issue of the communities (e.g. IR, data science and cognitive science) operating independently, rather than collectively towards a shared goal, is a possible underlying cause of harms in Web search that should not be overlooked.

We now hone in on specific sources of harm relevant to the overall discussion.

### 2.4.1 Content Found on the Web

Web documents and the information contained within are composed of different forms of media, such as text, video and images. Important considerations for the design and evaluation of search environments include the structure and position of information as well as the classification, categorization and semantics of the information and overall document.

Many of the topics and tasks covered here are relevant to the content transformation process discussed earlier (see Section 2.3). These tasks are introduced here because doing them poorly can lead to a degraded search experience and harmful outcomes.

Web documents combine the units of information contained within a document into a structure (typically using a markup language such as HTML), which can be exploited to extract positional details and apply weightings for improved interactive search experiences [127]. Similar techniques may also be applied to other media types such as video and audio [229]. Ultimately, this information is useful for linking of information across the Web [154, 229], and allows for improved search in various forms of media such as social media [152]. It has also been suggested that different document types and structures imply different cognitive meanings and therefore should be treated differently [102].

## 2. BACKGROUND

---

Semantics, for the current discussion, is the meaning (or sense) of an informational unit, including the meaning of a phrase, sentence, paragraph and overall Web document. Matching semantic meaning of the document with the semantic meaning of the user’s information need (e.g. via a query) is central to Web search and is one area where semantics is helpful. Just as a user has a semantic representation (albeit imperfect) in their mind, IR systems also have a semantic representation of the documents and information contained within. There are many automated approaches (often times probabilistic) available to create a semantic representation of the information in a document, such as LDA [31], TF-IDF [210], Doc2Vec [104], BERT [57], ELMo [175] and entity disambiguation [187] with knowledge graphs such as DBpedia [154]. The overall goal of such approaches being a reduction or resolution of ambiguity [111], however this ambiguity is not necessarily resolved (e.g. the role of bias covered in Section 2.4.5).

So far, we have highlighted how content structure and semantic representations of the content on the Web are important for Web search. Performing these tasks reduces ambiguity [111] and ultimately lead the searcher to information more specific to their *information need*. On the converse, poor reduction in ambiguity has the potential to lead the searcher into information that is less relevant and that may result in a higher risk of harm.

There are also many ways to analyse, categorise / classify and describe information found on the Web (often making use of approaches mentioned so far) worth mentioning here as well, as they are more focused to specific harms.

Discourse analysis, which includes the identification of controversy [59] and is a task tightly related to identifying disputable claims [64] and textual entailment (argument detection) [4]. In the same vein, there are ongoing efforts to classify infor-

mation that is counter-factual [3] or entirely fake or misleading [41, 43]. Somewhat related to discourse analysis is the classification and extraction of opinions, valence and / or sentiment of information [137, 171].

Going beyond discourse analysis, one might wish to determine if the information is offensive or hateful [188, 241, 261], the political slant [243] or the possibility that the content is propaganda [20]. Virality [74], that is the likelihood the information will spread, is yet another way to classify information, as is the reading level of language used [74]. There is also motivation to identify content that is addictive [256]. Classification cuts across document and media types, for instance one might need to classify the sentiment of a Tweet [89, 263] or determine how healthy the food is in an image [61]. Measuring the credibility (or trustworthiness) of the information is another important factor [94].

### 2.4.2 Information Collected While Searching

Web search is a two way street with respect to information that is exchanged. To retrieve the information they are looking for the user must submit a request (information), such as entering the domain for the home page of their favourite website in their browser or submitting a query for a set of results on their favourite search engine. The request includes information that may be collected by various parties. Understanding what types of information a user sends in their request and who collects this information is important for the discussion on potential harms.

Beginning with who collects the information, there are *first parties*, including the user's Internet Service Provider (ISP) (e.g. their mobile phone company) and the *information provider* to whom they have made the request (e.g. Google search, Facebook, a mobile weather App, their favourite news website). The *information*

## 2. BACKGROUND

---

*provider*, is the organization that stores the information being requested by the user. Organizations include news websites, social media companies and search engines, just to name a few. Dependent on the design of the information service being provided, the organization may or may not save information about the request made by the user. For example, Google records queries and clicks by users versus Duck Duck Go (a more privacy aware search engine) stores queries in a non-identifiable manner. Search queries are but one example of online behaviour that is collected and the AOL search log blunder highlighted the sensitive nature of search logs and raised anxieties amongst searchers [258], yet the general population perceives this information as much less sensitive than many other factors (many of which can be derived from search logs) [142]. There are also *third parties* for which first party providers may share some or all of the information in the request made by the user. Third parties may be state actors, such as the United States National Security Agency (NSA), where ISPs may share the information being requested by users [9].

There are many different types of raw information in the request sent by the user. As the main focus of our later studies is Web search, some focus is given to the information in the query log and Web pages that one might visit.

Cooper [49], in her survey of privacy enhancing techniques for query logs, identifies the basic components of a query log to include the user's IP address, the time they made the request, the query content (e.g. the query terms used), browser and operating system details, unique cookie ID and any results clicked. Privacy concerns around the query elements highlighted by Cooper have been raised [204, 245], such as de-anonymising users with these logs. Mobile phones have become quite popular in the last 10 years, and more recent methods demonstrate how this de-anonymisation can be performed, simply through the use of IP addresses in query logs, where the location of an individual can be closely approximated.

When a user clicks on the search result, the website (a different information provider) will set cookie IDs unique to its website and will collect much of the same information as the search engine (excluding the query terms). While navigating within the website, the website will record all pages visited by the user, which can be linked to the unique cookie IDs. Technology such as javascript and JQuery allow for collection of information [227] (such as scroll depth of a page visited). For users that try to block cookies and recording of IP (e.g. with a VPN), there are newer methods such as browser fingerprinting [30] which can be used for unique identification. Location of the user may also be collected, either inferred via the IP address or directly via the GPS coordinates provided by the user device [225]. Perhaps most important of all, all of the data collected may be shared with third parties for commercial or legal purposes [49], and therefore replicated in multiple locations.

For optimal IR system performance it is beneficial to not only collect the direct interactions with the system (e.g. queries, result clicks, rank of result clicked) but also secondary interactions known as Web trails (e.g. web pages navigated to beyond the initial search result visited) [245]. Therefore, dependent on the search system used (e.g. Google search) and interface to view information (e.g. Google Chrome), the search system may record the direct IR system interactions as well as the indirect actions that take place within the results visited. This is in addition to the information that is collected by the websites visited and the 3rd party intermediaries through the creation of a Web trail.

The user data collected through Web search has many positive benefits [49, 160, 245], including reduced time for task completion for the collective user group (not just an individual). The data collected is also useful for performing aggregations to learn the preferences [239] and demographics [110] of individual users, which plays

## 2. BACKGROUND

---

an important role to serve commercial purposes [49]. But one must not overlook that collection and usage of data in this centralized manner comes with risks of potential harms [245, 258] (see introduction for some examples).

### 2.4.3 Personalization

The saying is "there is no free lunch"; and the business model of popular modern search engines such as Google do not contradict this. The main source of income for commercial search engines is personalized and targeted advertising based on data collected about an individuals' online behaviour [235, 267]. Dependent on the search system, loss of privacy is one cost of it's use, a cost which is incurred through the collection of personal data during an individuals' search process.

In exchange for personal data and exposure to advertising, search engines are able to provide a more personalized experience [219]. This exchange (and loss of privacy) by itself is not necessarily a bad thing [97], however there are risks of personalization to consider.

For a start, demographics can be inferred from your search and online behaviour [27, 242, 243], leading to the risk of an individual being cornered into a narrow view of online information which may be entirely incorrect [230]. Prediction of demographic information based upon social links [124], and evidence exists that the risks are exacerbated when applied to social media environments [39].

Additionally, search systems are a gateway to webpages, of which many contain embedded tracking scripts to follow an individual across the web, thus exposing them to many privacy risks [63, 117, 149, 257] and to online behavioural advertising [226] for websites, some of which contain unproven and potentially dangerous products [159]. This data collected via an individuals' online behaviour has the potential to be



used in nefarious ways, including sharing of an individual’s mental health disorders [181], controlling access to insurance [201] and policing of citizens by state actors [33].

Interesting behaviours can be exposed when combining document classification predictions with user data. Sentiment analysis and opinion mining can be used to determine user and group beliefs towards products and entities [137] as well as the impacts of outside factors, such as weather, on social media posts [263], amongst other tasks as well. Sentiment analysis has controversially been used as part of an experimental IR task on a social media news feed which resulted in users posting more negative information [126]. Named Entity Recognition and Disambiguation (e.g. [154]) can be used to build a knowledge graph to better understand user interests.

Algorithmic methods to protect an individual’s privacy [5, 75, 257] are proposed solutions that should not be overlooked, however for many proposed approaches data may still be de-anonymized to determine the traits of individual, such as HIV infection [75]. While algorithmic approaches are a vital area of research for ensuring privacy protection during the search process, it is simultaneously important to take into consideration the users and their interactions [50] with search systems.

### 2.4.4 Ethics and Personalization

One side effect of user profiling and algorithms are *filter bubbles* [18, 54, 55, 230], another being biased results for a profile, where for example gender is the only variant [123]. Certainly, ethical questions arise when considering the potential individual and societal impacts [58, 173, 214].

## 2. BACKGROUND

---

*filter bubble* an environment in which content is selected by algorithms according to a user’s previous behaviours. [18].

User consent to the use of personal data is common practice in the web domain (e.g. use of cookies button, consent to terms of service), however, users do not necessarily consent to having implicit information used for *pre-selected personalization* [268].

*pre-selected personalization* implicit in nature and concerns situations where personalization is driven by websites, advertisers, or other actors, often without the user’s deliberate choice, input, knowledge or consent as defined by [268]

Though the literature suggests that users have transparency and explicit control over their profile, such as Google Advertising controls mentioned by [138] or Google Personalized Search [245], these features are not easily discovered. For example, finding controls at [Google Ads Settings](#)<sup>5</sup> is not a straightforward process for users that are not made aware of them, nor may such controls be guaranteed to remain (e.g. functionality of Google Personalized Search, as indicated by [245], is not available any more)<sup>6</sup>.

### 2.4.5 Biases, Behaviours and Beliefs

System and user biases and behaviours increase the likelihood that searchers on the Web will discover information that is inaccurate or entirely incorrect [244, 245], with the potential for negative search outcomes (e.g. harm to personal health)

---

<sup>5</sup>[Google Ads Settings](https://adssettings.google.com/) at <https://adssettings.google.com/> (LA: 2020-10-23)

<sup>6</sup>It will be very interesting to see the findings of an experimental platform developed at MIT which provides users with algorithmic control of social media feeds. See [Gobo](#) at <https://gobo.social/> (LA: 2020-10-28)

[179, 246]. There is a vicious cycle in the current IIR environment [16] that is a result of system biases (e.g. algorithms and models) and user biases (e.g. learned behaviours and cognitive biases).

There is indeed an interplay between the psychology of humans Web searchers and the Web search systems they use, and therefore these biases, behaviours and beliefs are considered in unison.

### 2.4.5.1 Misinformation Specific

One example of this interplay is in regards to exposure to *misinformation*, noting that other definitions of *misinformation* have been proposed, however the definition provided below is the most accurate for this thesis, where it is also assumed that “*misinformation* is created unintentionally” [202], which is not always the case [218].

*misinformation* science and health misinformation as information that is contrary to the epistemic consensus of the scientific community regarding a phenomenon [218].

This interplay between searcher, search system and *misinformation* is rooted in multiple factors, including algorithmic biases [3, 58, 164, 230], web page bias (due to author/editor) [164], multiple user biases [115, 233] including pre-conceived beliefs [6], trust in search system ranking [108, 164, 179] and source bias (e.g. reading your news from one website) [101]. Additional research in psychology demonstrates that different personalities (e.g. individualistic vs. communitarian) will respond differently to information presented [112], and therefore caution must be used as poorly implemented interventions can strengthen incorrect beliefs [113, 206].

### 2.4.5.2 Privacy Specific

Personal privacy threats and potential harms exist due to collection of our personal data and our online behaviour [27, 97, 230, 243], a potential cost of using IR systems (e.g. Google) that allow us to gain information quickly [267]. Another explanation to loss of personal privacy is that individuals place more or less value on privacy [2, 40], of which an individual's attitudes towards privacy do not necessarily match their behaviours to protect privacy [213] (sometimes called the *privacy paradox*). Such behaviour is also partially explained by an individual's lack of awareness of how their data is used [60, 215] and the cost (e.g. time) of setting up privacy tools and application privacy settings. There are a wide set of views on privacy [2], but nonetheless, many users, if not most, are paying too much with their personal data [215].

## 2.5 Interactive Information Retrieval (IIR)

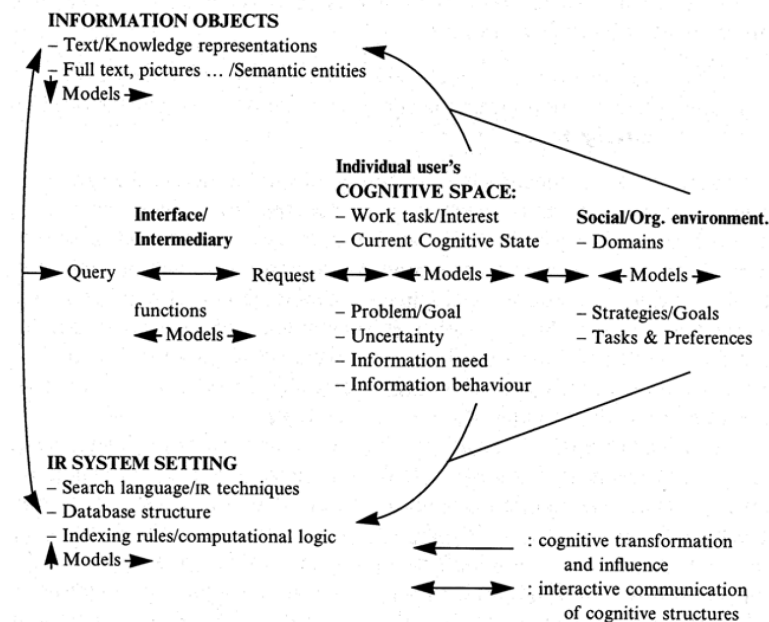
It is nearly 25 years since Ingwersen expressed the concern for the lack of a holistic view of system oriented and user oriented research in IR [102], yet more efforts are needed to achieve this view [50]. Nonetheless, there are growing efforts to investigate the nexus of user and system, as demonstrated by the 2016 inaugural Conference on Human Information Interaction and Retrieval (CHIIR). This holistic view of users and system together is known as interactive information retrieval (IIR) [119, 191, 245], and is the foundation for the experiments covered in later chapters.

We introduce this holistic view here as well as some of the ethical concerns related to IIR. Conversational search, a more recent incarnation of Web search, is a quite hot topic in IIR, that simultaneously has ethical concerns, and therefore given attention here. There is also a growing body of literature linking the cognitive

decision making processes to IIR, and therefore also important.

### 2.5.1 IIR: The Holistic View

There are many approaches to describe the overall search process, for which we find Ingwersen's cognitive model of IR interactions [102] (see Figure 2.6) is most intuitive for understanding and describing the interplay between the IR system, the individual and the societal environment in which they belong and draw links to harms and dangers of Web search. It is worth highlighting Saracevic's stratified model of IIR [196, 197] also suggests the interactions between system, user and the environment beyond, however, his model is arguably less easy to identify and extract the potential harms of Web search.



**Figure 2.6: Ingwersen's Cognitive Model of IR Interaction [102]** - A holistic view of the search process. The user's cognitive space is placed at the center of the model, and the user interacts with the Web search system as well as the broader social environment.

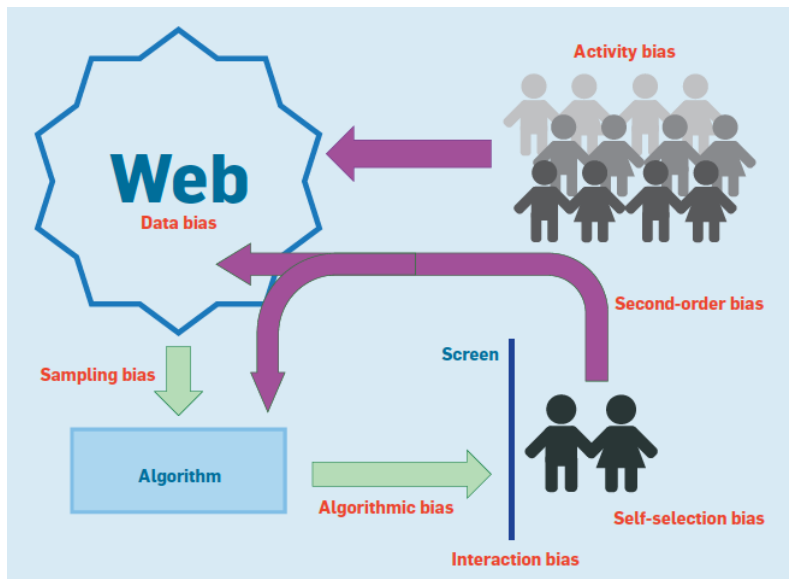
Ingwersen and Järvelin [103] further expand and more clearly define the holistic view of search by combining the cognitive and system oriented approaches, which

## 2. BACKGROUND

---

they refer to as information seeking and retrieval (IS & R). Their review and proposed framework provides useful insights for IIR researchers, some quite important to the current research. They identify that there are not only information seekers (users) in the IIR environment, but other important actors as well, including IR system designers, authors of documents in the IR system index, selectors (i.e. actors that decide which information is made available, such as via a publishing process) and broader communities of individuals. These actors are influenced by their past and current experiences, suggesting an interplay between cognitive changes of the actor and the IR system they interact with. Ultimately, the actors influence the environment, where the environment not only includes the IR system itself, but also the social and cultural environments in which the actors may also be members. Baeza-Yates' recent commentary suggests the interactions between IR systems and users of the systems [16] are a vicious cycle (see Figure 2.7) being a primary cause of problems such as information bias, and is direct real world evidence in line with the holistic (IS & R) view. As such, the cognitive actors and the IR systems they interact with are not separate entities, and therefore should be treated as a whole. Combined together, they are a pathway to identify potential harms of Web search, to prioritize research, and ultimately improve the design of the Web search environment reduce risk of such harms.

It is worth considering the holistic views of IR which integrate cognition (such as [102] and [103]) in combination with Baeza-Yates' recent commentary [16] on the problems and dangers presently associated with commercial IR systems. Combined together, they are a pathway to identify potential harms of Web search, to prioritize research, and ultimately improve the design of the Web search environment reduce risk of such harms.



**Figure 2.7: Baeza-Yates Diagram of the Vicious Cycle of the Search Process [16]** - The diagram provided in the recent commentary takes a holistic view of the search process. It demonstrates that many of the present harms are the result of biases present not only in components of the Web search systems, but also in the users.

## 2.5.2 Ethics of IIR

In her pioneering guide for IIR research [119], Kelly raises identifies two central questions to IR and IIR research. Research in IR asks: *Does this system retrieve relevant documents?*, whereas IIR research asks: *Can people use this system to retrieve relevant documents?* [119]. In a later framework for IIR research, White provides practical considerations for modelling and evaluating IIR systems in the frame of different models of user behaviour [245]. Both Kelly and White suggest ethical considerations for this area of research, and White gives extra consideration around privacy concerns related to IIR systems, and we suggest that all IR and IIR researchers should include such considerations in their research.

As an example, in their highly influential research on implicit feedback in IR systems, Joachims et al. [108] do raise concerns around trust bias in rankings and bias towards results with higher quality snippets relative to others, but do not frame

## 2. BACKGROUND

---

it in an ethical manner.

**Outcomes of IIR** As suggested by many of the views outlined, such as sense-making, decision making and the related outcomes are important considerations in the frame of IIR [244]. Returning to the two central questions identified by Kelly for IR and IIR research [119], we suggest that an additional question should also be asked *What are the potential outcomes to the users of this system?* Outcome is very broad word, for which we regard as measurement of consequences both good and bad. Outcomes encompass existing measures such as relevance of information and costs such as time, but also many other factors such as decisions made with the information provided by the IR system (e.g. medical decisions [179, 244]) as well as societal outcomes due to IR system biases [16]. The potential harmful outcomes of IR systems are a key motivator for the current body of work.

It is quite plausible that the gap between user and system researchers is a source of some of the challenges being raised in modern times. Challenges including biases of algorithms and IR system results, spread of misinformation, as well as concerns related to data collection and privacy. Therefore, we argue studies of IR systems and their users should take the global IIR view.

### 2.5.3 Search Interface and Design

The search user interface (SUI), an umbrella term including search engine results page (SERP) and conversational search assistants, acts as the medium between the human and the IR system, and it would be quite difficult for IR to be interactive without it. One could argue the SUI captures the essence of IIR, and it is therefore an important element to include in this discussion. There are many design principles one must take into account for a SUI [91, 249] and it would be difficult to satisfy all



## 2.5 Interactive Information Retrieval (IIR)

---

principles [91]. Many of the core design principles of SUIs are influenced by the field of human computer interaction (HCI) [91, 249], principles which include consistency (e.g. describing features in a same manner) and handling errors (e.g. error messages) [162]. Principles also suggest the interface should be kept simple and intuitive to address the needs of a broad set of users [91, 162]. The most important principle of all appears to be the response times the user experiences in a SUI [91, 249].

There are many components of a SUI that designers might take into consideration. Wilson [249] introduces four main groups of features common to the SUI: inputs (e.g. query bar) to express search intent, controls (e.g. filters) to modify inputs, information (e.g. URL) about the results and personalisation (e.g. search history) that make the SUI unique to the user. These groups play different roles dependent on the searcher and search task (especially complex tasks), where for instance inputs and controls are important at earlier stages of a search task whereas personalisation becomes more important in later stages [100].

Challenges have arisen through the years that have motivated theories and innovation to predict user behaviour and enable users opportunity to avoid harm, where assessing credibility of online information is one example. For instance, individual differences between searchers as well as differences of online task (e.g. task domain and task purpose) are predictors for what cues an individual makes use of to assess credibility [68]. Concerns around credibility are quite important given that one must consider anyone can publish information online [234], thus providing inspiration for a broad set of interfaces that enable searchers to more easily assess if the information is legitimate or potentially dangerous [234]. There are many different interventions available and considerations to be made to address credibility challenges [116].

User and system biases were already introduced in Section 2.4.5 as a cause of

harm in search. Here again search interfaces are shown to play an important role in addressing such problems [234]. Cognitive science have helped researchers understand biases and address related problems in other domains, and it is certainly a goal to incorporate cognitive approaches into the interface to better assist the searcher in overcoming the related problems (e.g. trust bias) [72]. It is also worth considering motivational affordances (commonly studied in the cognitive sciences) in interface design as they clearly play a positive role in technology that is designed to persuade people towards less harmful actions [90]. Additional focus will be given to the role of cognitive science in search interface design in Chapter 3.

### 2.5.4 IIR and Conversational Assistants

Conversational dialogue systems have, such as Amazon’s Echo, Microsoft’s Cortana and Apple’s Siri, have experienced growth in popularity in recent years [99, 183], for which some forecast growth beyond global human population in only a few years<sup>7</sup>. It has been found the conversation systems are more interactive than traditional search systems [228] (e.g. through a search engine such as Google), and it is possible they will become the future of search [183]

The recently proposed framework for conversational search indicates that trust, ethics and morality are important factors related to conversational search which this framework only introduces as a concern but does not address [183]. Other indications have suggested ethical concerns to conversational assistants, in particular concerns related to privacy [78, 99]. This matter of privacy concerns links back to the earlier introduction to information that is collected during search (see Section 2.4.2, where it has been found that many users of conversational assistants are concerned about

---

<sup>7</sup>for example see [Statista forecasts 8.2 billion conversational assistants by 2024](https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/) at <https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/> (LA: 2020-10-15)

privacy but simultaneously unaware of the risks such assistants present [130].

### 2.5.5 Decision Making Processes

Attempts have been made to better understand the decision making process and rationality components of the searcher in the face of risk and uncertainty during the search process, such as the impact to one's health [179, 246] and the heuristics used in deciding what information is credible [94]. Such approaches suggest directions for alternative search outcome measures, such as measurement of positive health outcomes [179]. There are also recent proposals in the field of IR that would allow individuals to assess quality of information and to make more informed decisions [74], however such methods may have high costs [92] and thus should be assessed with the economic view.

The economic view of the decision making process of search and the costs to reduce risk and uncertainty inherent in the search process is of particular interest. Risk and uncertainty are extensively studied in the area behavioural economics, and thus provide promising pathways for low cost measures to minimize the dangers associated with interactions with modern IR systems and the web of information. From the behavioural economists viewpoint, research suggests that humans are not rational in their decisions [115, 233] or alternatively that our decisions are rational within the bounds of our environment [82, 84]. Such considerations of rationality (or lack thereof) in our decisions suggest that in some way our minds are biased [81, 114] and that such biases potentially cause harm to the individual and / or society as whole [220]. Low cost methods have been developed in recent years, methods shown to address such biases in our economic behaviour [92, 220], with the goal of producing decisions that reduce harmful impacts to the individual and society as whole, and thus having the potential for adaptation to the domain of IR

and web search.

**Heuristics for Decision Making in Search** Evidence exists that users have a set of heuristics (also known as “rules of thumb” [79]) to help them decide which information to visit [207] during Web search, where heuristics are also broadly used to make decisions in everyday life [79, 81]. Research by Novin et al. ([164]) demonstrates how biased heuristics, specifically anchoring, framing, priming and availability heuristics [114], are fundamental to decision making in search<sup>8</sup>. Their work investigates heuristics and their relation to algorithmic bias (ranking) and article bias (bias from author) around a controversial topic [164]. Their methods include use of a mock search engine, placed in a library setting, to study cognitive biases. Anchoring effects are found due to ordering of results in search engine results page (SERP), users skip more challenging content due to the priming effect and lengthier materials are linked to the availability effect [164]. Similar to Joachims et al. [108], Novin et al. [164] finds that users are biased towards higher results than lower ones. Interestingly, [164] argues that the Google approach to finding single answers to complex queries is problematic and overly simplistic.

## 2.6 Behavioural and Cognitive Interventions

If it is yet not apparent from the discussion in section 2.5.5, decision making is fundamental to the search process [191, 245]. Furthermore, as was outlined in the same section, behavioural and cognitive factors (e.g. cognitive biases and be-

---

<sup>8</sup>Biased anchoring and framing heuristics are easily demonstrated through a toy example. Try it! Open two side by side windows of your favourite search engine, in 1 window type "Why is Christianity good?" in the other window type "Why is Christianity bad?". The questions are framed differently (potentially influenced by personal biases), the system provides biased results based on the query formulation, with rank and ordering of results providing an anchor (users are biased towards selection of the first result, which can be the anchor for any subsequent information consumed).

havioural economics) greatly influence the decision making process. Therefore, it is worthwhile to consider strategies from behavioural and cognitive sciences that are developed specifically for minimizing risk and harms.

Three approaches, *nudging* [184, 220], *boosting* [93, 125] and techno-cognition [133], were recently proposed pathways to minimize and address harms in the modern online world [125, 140].

We focus on the first two approaches, *nudging* and *boosting*, as they are quite different in their methodology, yet very similar in their aim of reducing individual and societal risks. Additionally, we introduce nutrition labels and fact boxes as means to communicate potential harms.

### 2.6.1 Nudging

*Nudging* [220] is a popular behavioural-public-policy approach, which has gained notoriety in recent years. *Nudges* aim to push people towards—what the ‘nudger’ believes to be—more beneficial decisions through the ‘choice architecture’ of the environment (e.g., default settings) in which people operate. *Nudging* is theoretically rooted the field of behavioural economics hypothesizing that individuals have two systems of the mind to help them make decisions, where “system 1” is the more automatic and economical side compared to the less economical and automatic “system 2” approach [114, 115, 233]. The theory and empirical body of work suggests that our “system 1” thinking is biased and therefore prone to errors [233], and *nudging* is the proposed solution to address these errors [220]. Thaler and Sunstein [220] provide the following definition of a *nudge*:

A *nudge* ... is any aspect of the choice architecture that alters people’s behaviour in a predictable way without forbidding any options or signif-

## 2. BACKGROUND

---

ificantly changing their economic incentives. To count as a mere *nudge*, the intervention must be easy and cheap to avoid. *Nudges* are not mandates. Putting fruit at eye level counts as a *nudge*. Banning junk food does not.

*Nudging* requires that the ‘choice architecture’ includes an element of libertarian paternalism [220], that is, the *nudge* must allow the individual to opt out (e.g., choose the non-default option); this is different from a purely paternalistic approach such as bans, which have no opt-out mechanism by design and intent. Some *nudges*, such as nutrition labels and warning lights, have educational elements [125], but for the most part, *nudges* aim to directly change behaviour without targeting people’s competences [93]. The political philosophy and claims about human nature underlying *nudging* have been criticized recently [80, 167]; see, for example, [184] for a review of the issues discussed. Self-nudging [184] – people acting as their own “citizen choice architects” – has been proposed as a way to harness the power of *nudging* while largely circumventing its problems.

### 2.6.2 Boosting

*Boosting* (see [93]) is another approach to behaviour change based on evidence from behavioural science. Quoting Hertwig and Grüne-Yanoff (p. 974):

The objective of *boosts* is to improve people’s competence to make their own choices. The focus of *boosting* is on interventions that make it easier for people to exercise their own agency by fostering existing competences or instilling new ones. Examples include the ability to understand statistical health information, the ability to make financial decisions on

the basis of simple accounting rules, and the strategic use of automatic processes (...)

In the context of web search, *boosting* aims to improve people’s skills to effectively and safely search the web. To achieve this, a *boosting* approach combines both IR research on how people search and adapt their search strategies to the environment [94, 131, 207] with general insights on human judgement and decision making online [125, 140] and offline [83] to design and evaluate *boosting* interventions. Quite often, the skills are a simple set of heuristics (or “rules of thumb”) [79], which can be taught in domains such as literature search [131].

The key difference between a *boosting* and *nudging* approach lies in the formers’ assumption that people are not simply “irrational” (and thus need to be *nudged* towards better decisions), but that the human cognitive architecture is malleable and thus new competencies and skills can be instilled [93]—often requiring little time and effort. However, whether *nudging* or *boosting* is the “better” approach for a particular situation depends both on ethical considerations (e.g., how much value is placed on people’s autonomy), but also on pragmatic considerations of which approach will likely be more successful in terms of effectiveness and economic and non-economic costs. For example, since *boosting* needs people’s cooperation to be effective, *boosting* has the advantage that it—by design—cannot be manipulative. But this cooperation requirement also implies that *boosting* will not be successful in situations where people are unwilling or unable to learn or make use of a *boost*. See [92] for a discussion and suggestions for when *nudging* or *boosting* is likely to work better.

### 2.6.3 Nutrition Labels and Fact Boxes

Cognitive science has also provided a large body of evidence on visual approaches for communication of risk in an understandable way. Nutrition labels are one popular visual approach for risk communication, and it is shown that traffic light type approaches produce better outcomes and are more preferred by users [147, 203, 237] than tabular based approaches<sup>9</sup>. Originally designed for medical decision making for doctors and patients, fact boxes are another promising means to provide information in a manner that includes the benefits and harms of the available decisions [150]<sup>10</sup>. Interestingly, both nutrition labels and fact boxes can act as the medium to perform a *nudge* or a *boost*.

### 2.6.4 Comparing Nudging and Boosting

There are some key differences one must be aware of when comparing *nudging* and *boosting*, of which all were established above. Here we directly contrast the most important features that distinguish a *boost* from a *nudge* as they are critical factors for **G-RQ-3**.

A *nudge* is divided into two main sub-types, classical and ‘educative’ [93, 216], of which both can be used in the context of self *nudging* [184]. With respect to this comparison, one distinction is that a classical *nudge* is non-transparent in their intent, whereas an ‘educative’ *nudge* has transparency and provides some insight into the underlying motive [93, 125]. Another variation, is that an ‘educative’ *nudge* may require some motivation on the part of the individual, whereas as the classical

---

<sup>9</sup>Such as those produced by the Food and Drug Administration. See example at [FDA New Nutrition Labels](https://www.fda.gov/food/nutrition-education-resources-materials/new-nutrition-facts-label) at <https://www.fda.gov/food/nutrition-education-resources-materials/new-nutrition-facts-label> (LA: 2020-10-30)

<sup>10</sup>See examples of fact boxes at the [Harding Institute for Risk Literacy](https://hardingcenter.de/en/fact-boxes) at <https://hardingcenter.de/en/fact-boxes> (LA: 2020-10-14)



## 2.6 Behavioural and Cognitive Interventions

---

*nudge* is designed in a way that once the intervention is successfully implemented, momentum is carried forward as long as the *nudge* remains in place[93]. Using a real world example comparing these approaches is perhaps the best way to see this difference. Take the classical save more tomorrow *nudge* [220]<sup>11</sup> and compare it with an ‘educative’ nutrition label *nudge*<sup>12</sup>. In the case of the save more tomorrow *nudge*, no motivation is required and as long as the individual does not *opt-out* of the default, the momentum of the *nudge* will intervene in a manner that individual’s pension continues to grow (assuming no economic catastrophe). In the case of the ‘educative’ *nudge*, the person must be motivated to take a look at the nutrition label (in a sense, they must *opt-in*), and furthermore must take effort each and every time they buy (and eat) food in order to maintain momentum [93, 216].

It was already established that a *boost* is transparent to the individual like an ‘educative’ *nudge* and both require motivation by the individual [93, 125]. Furthermore, both the *boost* and ‘educative’ *nudge* are likely to have a short-term effect with respect to the intervention for which they are designed [93]. However, the main differentiating factor is that a *boost* is designed to transfer a competency for better decision making, which is validated through empirical research, whereas an ‘educative’ *nudge* may have empirical validity for better decision making, but does not have a competency included in the intervention [92, 93, 125].

The point made with respect to the provision of a competency for better decision making leads to the most consequential difference, being the notion of ‘reversibility’ of *boosting* and *nudging* interventions. Hertwig and Grüne-Yanoff ([93] p. 981) define ‘reversibility’ as:

---

<sup>11</sup>The save more tomorrow *nudge* is implemented in the United Kingdom, where the default is that UK employees will contribute a percentage of their salary to a pension plan.

<sup>12</sup>‘Educative’ nutrition label *nudging* is commonly used (and in some cases required) on food packaging in both the United Kingdom and the United States

## 2. BACKGROUND

---

If, *ceteris paribus*, the policy maker eliminates an efficacious (nonmonetary and nonregulatory) behavioral intervention and behavior reverts to its preintervention state, then the policy is likely to be a nudge. If, *ceteris paribus*, behavior persists when an intervention is eliminated, then the policy is more likely to be a boost.

This definition is the dominant characteristic that divides the two strategies, as it says that the behaviour resultant from a successful *nudge*, be it a classical *nudge* [220], an ‘educative’ *nudge* [216] or a self *nudge* [184], will revert to the default after removal, and it can be empirically demonstrated that this will not occur with a *boost*. One must keep this point in mind if they wish to test both classes of interventions in Web search or in any other domain for that matter.

These differences aside, both approaches have merit with respect to promotion of better decision making and therefore both should be considered [140].

### 2.7 Evaluation

Evaluation is a fundamental and necessary process for both pure IR system and fully interactive IR studies. Common approaches include task based (and user focused) [119], batch based [232] and online based [96] evaluation of systems.

Batch based studies have been around the longest, were first used in the well known Cranfield studies in the 1950s (see [45]), and used for offline studies to compare IR systems [51, 144]. They used closed evaluation sets, such as the TREC and CLEF test sets [51], and are evaluated based upon the documents returned for queries associated with a particular information need [144], but do not include human interactions [119].

Online evaluation is typically performed on live systems (e.g. Google search), where a small percentage of users are evaluated with a proposed new (or improved) system to be compared to the existing system [51].

Our studies, as are many IIR studies, are user focused, and therefore evaluation methods for these studies are given special attention in Section 2.7.3. User (IIR) oriented studies regularly make use of evaluation approaches common to system oriented (batch based) studies and therefore some attention is given to these approaches too. As evaluation is such a large topic upon itself, focus is primarily given to approaches utilized in this thesis.

### 2.7.1 Evaluation Test Sets

In many (if not all) IR system evaluations, a document test set is utilized for which each document is assigned an atomic unit of measure to be used in the evaluation, where the most commonly used atomic measure is document relevance [144]. Document *relevance* (a concept first introduced in Section 2.2.6) is defined as the following:

*relevance* taken with respect to a user query (for an information need) and a binary classification of each document returned by the query, the document is either relevant or not relevant [144].

More recently, the notion of *relevance* has extended beyond binary classification, to ordinal ranking of relevance (e.g. 0 not relevant at all to 4 is highly relevant), and is known as graded relevance, which allows for a more user oriented evaluation of the system [118].

Though many standard test set collections are available (e.g. TREC, CLEF,

## 2. BACKGROUND

---

NTCIR) [144], there are many types of evaluations one might perform which requires the development of new test sets. In text classification systems, there are many examples (see some in Section 2.4.1) where new test sets are created to tackle new problems (such as hate speech and misinformation). This is also true for IR specific problems, for example annotating for health misinformation (see [179]). At the end of the day, it is potentially risky to place credence into a system evaluated on a test set for which the annotations themselves have not been evaluated.

**Evaluating Test Set Quality** Evaluation of the “gold standard” document tests sets used for document classification and IR system models is a very critical task to maximize the likelihood that users receive high quality information in the systems they use. These “gold standard” datasets are developed with humans performing annotation tasks dependent on the system and/or model being developed. Many approaches are available to complete the annotation task including usage of on-line ratings [172], a combination of non-domain and domain expertise [240, 241] and crowd sourced annotations with platforms such as Mechanical Turk [157] and CrowdFlower [38, 255]. In many cases, document classification models trained on non-expert annotations via crowd sourcing will perform better than those built with annotations from experts [209].

Multiple measures are available to evaluate reliability of annotations, Kappa statistics are commonly used [144] (e.g. Cohen’s kappa and Fleiss’ kappa). In depth review of these and other measures (such as Krippendorff’s alpha) in the scope of annotations for tasks related to IR and computational linguistics is provided by [10]. It is worth noting that annotation tasks can be quite subjective and challenging for humans, as demonstrated by an earlier introduced example for hate speech

---

annotations<sup>13</sup>.

## 2.7.2 Evaluating the Search System

Evaluation metrics for document retrieval systems can be separated into measures that either do or do not take into account ranking [144]. These measures may also be classified as those that measure system *effectiveness* (e.g. does the system return relevant documents?) or *efficiency* (e.g. does the system return results quickly?) [51]. For studies introduced later in this thesis, we do not investigate *efficiency*, and therefore do not discuss related measures.

Furthermore, the guiding research questions do not use *effectiveness* in the sense of IR text books (e.g. [51]), but in the sense of *effectiveness* of reducing harm (addressed by **G-RQ-2**). With respect to the traditional sense of *effectiveness*, this plays into the overall viability of the search system addressed by **G-RQ-1**.

To complicate matters further with respect to the evaluation of a search system, in some user based studies, relevance assessments (e.g. as in a TREC test set) may not be available for the test sets used, such as in online system evaluation. To circumnavigate this issue, Kelly [119] suggests using the documents assessed by users as an alternative, with the side effect that some evaluation metrics will, such as recall, will be difficult (or not even possible) to calculate.

With that said, we differentiate commonly used system evaluation measures appropriate to studies in this thesis into those that do not account for rank and those that do.

---

<sup>13</sup>See [Facebook Hate Speech Moderation Quiz](#)

## 2. BACKGROUND

---

**Non-Rank Based System Evaluation** The most popular measures that do not take into account the ranking of the system include are *precision* and *recall* [51], which can both be considered jointly to compute the *F-measure* (the weighted harmonic mean of *precision* and *recall*) [144]. All three of these metrics are commonly used in information retrieval as well as other document and text classification problems [144].

**Rank Based System Evaluation** *Precision @ rank k* only evaluates a system based upon the first  $k$  documents returned by an IR system. This in contrast to *precision* which takes into account all documents returned by the IR system. Both measures are used regularly in evaluation [51, 144], and  $k = 10$  is a common choice [51].

*Mean Reciprocal Rank (MRR)* takes into account the rank of documents and the first *relevant* document returned by a system within that rank, simply by taking the 1 divided by the rank of that document [51]. In this manner, systems are heavily penalized when the first document is not *relevant*.

Traditional search evaluation metrics such as precision and recall do not take into consideration that some documents may be more relevant than others. This issue motivated efforts by Järvelin and Kekäläinen [106] to develop *gain based measures* to assess search systems with this consideration at the forefront. The pioneering evaluation methods include *discounted cumulated gain (DCG)* and *normalized discounted cumulative gain (nDCG)*. Their methods can be used in settings with document assessments that are binary or graded (e.g. 4 points for document that is most relevant and 0 points for a document that is least relevant). Furthermore, the metrics are quite versatile and allow for adaptation to measures not limited to graded relevance [106], and therefore should not be overlooked for evaluation of systems aimed at

---

reducing harms. *DCG* and *nDCG* take into account that documents graded high are more useful than those with lower gradings and simultaneously account for the rank [51].

### 2.7.3 Evaluating Interactions

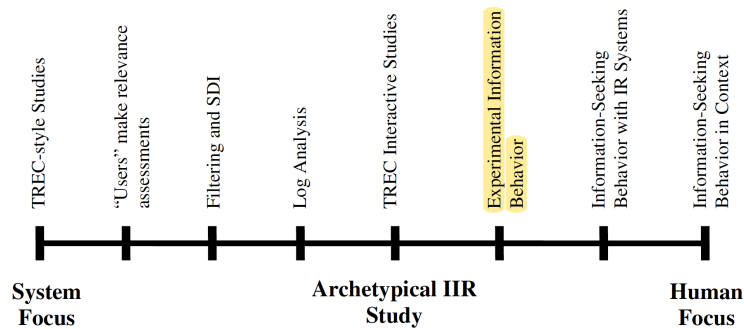
Extensive surveys by Kelly and White (e.g. [119] and [245]) provide the foundations of methods and evaluation of interactive IR studies. Other important considerations are questionnaire design [119, 165], instrumentation such as eye tracking [245], recruitment methods [119] and significance testing of results [192]. Eye tracking instruments which capture gaze and pupil dilation patterns, are a common tool for evaluating impacts to user behaviour [245] with a concrete example provided by [108]. Recruitment of participants can be performed passively [164], actively [179], covertly [126] or via a combination of these options [245].

There are many flavours of IIR studies. Examples include evaluating the usefulness of popular implicit measures [108], evaluation of the re-ranking of results to better fit a personal profile [219], efforts to understand the user and algorithmic biases in the domain of controversial information [164] or gaining understanding the role of user and IR system biases and beliefs which may lead to consumption of misleading or incorrect health information [244]. Studies incorporating economic theory (introduced in Section 2.2.5), such as measuring cost trade-offs for query auto-completion [13], are yet another flavour of IIR research. IIR studies, Kelly suggests [119], fall along a spectrum (see Figure 2.8) of research, from system oriented approaches with minimal user focus (e.g. TREC evaluation tracks) to user centric studies with limited views of the IR system (e.g. [94]).

A highly useful resource for the evaluation of IIR studies is by Kelly in 2009

## 2. BACKGROUND

---



**Figure 2.8: Kelly’s Continuum of IIR Research [119].** - The spectrum is quite broad and is centred on studies that equally consider both the user and the system (e.g. TREC interactive studies). Based on research questions and methods used, studies in the current thesis fall under the highlighted category just to the right of centre.

[119]. Important recommendations from this work include many, for which we highlight those most relevant to this thesis that have not already been mentioned. First, a within-group study design should be used for comparison of systems, which has the benefit of allowing users to give their preferences across the systems. Second, a between-group study design is acceptable when there is the possibility of contaminating result with respect to the research question. To ensure validity of data collected to test hypotheses, methods for rotation and counterbalancing of tasks and randomization of participants should be used. Specific to analyses, Kelly suggests that one use (when possible) parametric statistical tests as they are less sensitive to type II errors.

**Differentiating IIR Evaluation Measures** There are several notable typologies for describing commonly used evaluation measures in IIR studies. White [245] breaks these into two dimensions, search process-oriented and *search outcome* oriented, where the former considers the process employed by the searcher (such as their rationale for picking a result) and the later considers all outcomes (e.g. answer to a search task question, relevance of documents). Contextual, interaction, perfor-



mance, and usability are the four main dimensions suggested by Kelly [119], where performance is for all intents and purposes the same as White’s *search outcome* dimension.

The other three dimensions (contextual, interaction and usability) [119] are quite appropriate for user based evaluation. The interaction dimension, typically requires log data and includes factors such as number of queries entered and total documents assessed. Contextual factors include those differentiating users, of which there can be a multitude of measures, such as age, sex, factors relating to a subjects concern of importance with respect to the search task, etc., of which studies often interrogate participants directly through surveys to collect the necessary data [119]. The dimension of usability similarly ask subjects directly for input about factors such as their preferences about the system and their subjective feelings about the usage of the system and the tasks performed in the system [119].

The *search outcome* dimension is an important one with respect to measurement of harms in Web search. There are indications that newer evaluation metrics are coming on board with respect to this dimension. This claim is demonstrated through recent IR research in the space of health search. In one case, an IIR study in the domain of health search considered *search outcomes* of *harmful* and *helpful* [179, 244] for medical search task, which is a quite different approach to evaluation of *search outcomes* specific to *relevance*. This is again demonstrated by the 2020 TREC Health Misinformation Track<sup>14</sup>, which looked at misinformation. Indeed, it appears evaluation is moving beyond *relevance* towards proxies assumed to be strongly linked to *harmful* and *harmful* outcomes (credibility and correctness of information in addition to *relevance*), at least in the space of health search.

---

<sup>14</sup>2020 TREC Health Misinformation Track at <https://trec-health-misinfo.github.io/> (LA: 2020-10-29)

### 2.7.4 The Science of Evaluation

There were many details covered (and not covered) for the evaluation of IR system and user studies, it is worth stating that many IR system and user studies performed are not following suggested protocol [73, 192] and furthermore often lack statistical rigor [192]. A more recent commentary [46] indicates that many of the problems related to evaluation in IR extend to all areas of computer science, and that studies often times do not take a scientific approach or alternatively are never made available due to publication bias. In many cases there needs to be more effort taken to demonstrate statistical power of studies [192], such as with calculation of effect size [73]. On the other, reviewers need to keep in mind that some studies are more exploratory in nature (in particular human interactive studies) and therefore may not be able to meet the same statistical bar, but are yet important for science as a whole [46].

As indicated, it is important to follow scientific methods to demonstrate that one system either does (or does not) outperform another system, and some simple suggestions are useful to address the concerns raised (but only a starting point). One recommended approach is to calculate the *effectiveness* measures for each query submitted to each system, and then to test for differences with tests such as t-test (parametric) and rank (non-parametric) tests [51]. For user studies, again the use of statistical models is suggested, such as t-tests and ANOVA, to test for differences across groups and systems [119]. And where possible, calculate effect sizes for any tests performed [73, 119].

## 2.8 Summary

Much has been introduced in the previous chapter and the current background chapter, for which it is important to bring together the key points.

Research and theory developed in information science (IS) has inspired many of the algorithms, models and interfaces developed and implemented in modern information retrieval (IR) systems [245]. Even with this influential link between the communities, quite recent commentary [50] suggests there is a gap in research between the two areas and a need for a holistic view of the searcher (the IS focus) and the search system (the IR focus) as one system together [102, 103]. In fact, Ingwersen and Järvelin [103] suggest this view must go beyond just the system and the searcher.

Data science, a profession deemed “the sexiest job” [52], also has agency in the modern Web search environment. The research community, which includes the data science profession, has quite an influential role in modern Web search environments, as members of this community are often tasked with development of models that are optimized to maximize user satisfaction [245], revenue and profits [52, 267] and user engagement [52].

Finally, there is the community of searchers that seek and find information on the Web through a multitude of IR environments such as search engines (e.g. Google), product websites (e.g. Amazon) and social media news feeds (e.g. Facebook) that have their own biases and beliefs [16, 244]. Unfortunately, the biases and beliefs of the searcher and the IR systems they interact with form a feedback loop that not only changes the system [16, 18, 230] but also changes the beliefs of the user [16, 230, 244]. Ultimately, this self-reinforcing cycle exposes individual searchers and

## 2. BACKGROUND

---

broader society to potentially harmful and dangerous outcomes [16, 179, 230, 244]. It is our view that the potential and already-realized harms caused by this reinforcing cycle are a side effect of the non-holistic view, a view which must extend beyond the searcher and the system [103].

Taken together, this motivates the framework for harm prevention in Web search introduced in the next chapter as well as the guiding research questions ( **G-RQ-1** , **G-RQ-2** and **G-RQ-3** ) introduced in Chapter 1, questions which target behavioural and cognitive strategies as part of the solution.

# Chapter 3

## A Harm Intervention Framework for Web Search

### 3.1 Overview

Our framework is aimed at communities which include researchers in IS, IR, data science and the behavioural and cognitive sciences as well as the policy makers in governments and the leadership teams of Web platforms. There is common recognition by these communities of the ethical concerns and potentially grave implications of the technology that are now ubiquitous in our everyday lives. Simultaneously, aside from efforts by IS and IR [24, 50, 103], these communities appear to be working independently of one another. As such, we see a need for a common framework for all of these communities to jointly work towards a common goal of ethical responsibility to the searcher and broader society for which we are a part of. The components of the framework (Section 3.2) are ones we believe are the most essential for these communities to place focus initially. Components include policy updates, cognitive interventions, evaluation methods, and considerations for overall

### 3. A HARM INTERVENTION FRAMEWORK FOR WEB SEARCH

---

search system design. Central to our framework are four themes: collaborative effort by the communities mentioned, greater transparency to the user, greater choice for the user and an ethics-based approach for search system development.

Science, such as the methods and empirical studies introduced in later chapters, plays the role of demonstrating viability and effectiveness ( [G-RQ-1](#) and [G-RQ-2](#) ) of interventions, but rarely (if ever) defines the policy that implements interventions in the real-world. Nonetheless, policy is an important and necessary component, and therefore will be given some focus.

Concurrently, in the space of Web search, there is great focus on new and improved algorithms and artificial intelligence as the solution to many of the problems (e.g. misinformation). The author of this thesis agrees with this focus, however there is seemingly very little in the frame of behavioural and cognitive interventions in Web search, and yet are interventions which show great success in other contexts. Hence, why behavioural and cognitive approaches are included as one of the four pillars of this framework, and ultimately is the main focus of the methods and studies that follow.

This framework serves several purposes. First, it introduces what we suggest are the essential focal points to address potential harms from Web search. Furthermore, in line with the holistic view of IIR (see Section [2.5.1](#), the framework is designed with this holistic view at the forefront. Additionally, the framework positions harm prevention interventions from the behavioural and cognitive sciences as an essential element of Web search systems designed to prevent harm. Finally, it is hoped the framework initiates discussion within the search communities (e.g. amongst researchers and technology platform leadership) as to how we might better design Web search environments to prevent harms to individuals and broader society.

## 3.2 Components

A recent proposal by Smith and Rieh [208] provides important insights for the development of a framework for Web search systems designed to prevent harms to individuals and society. Their proposal suggests that many users have information literacy and critical thinking skills that are useful to reduce the risks of harms from Web search. They highlight that current implementation methods of search systems, such as the Search Engine Results Page (SERP), offer little if any support to utilize these skills. Furthermore, they point out that users (likely due to the high cognitive demands of applying information literacy skills) put too much trust in the results found in the SERP, as has been demonstrated by other research [108, 244]. It appears some vitally important processes of search introduced by the IS community are being inhibited by the current design [208], processes including sense-making and exploratory search, which is unfortunate given the role they play in learning [245]. Their proposal also suggests that current IR systems are optimized for close-ended tasks (e.g. fact-based, question-answering), but should instead be optimized for learning. Ultimately, their proposal being that information in Web search interfaces should offer cues (e.g. topic, author and their affiliations, affective semantics including hate and humor) that enable users to utilize their literacy skills and assist them with critical thinking. However, their framework, as encompassing as it is, does not consider important matters such as privacy of the searcher [211] and platform policies optimized for corporate profit [256, 267]. That said, many elements of their proposal, such as considerations for interface design and informational cues that engage critical thinking for better decision making, are central to our proposal. We therefore see our proposed framework as an extension of their work. The decision-making factor is in fact fundamental to search [191, 245], suggesting a strong need for cognitive interventions, which are proposed as a bridge between

### 3. A HARM INTERVENTION FRAMEWORK FOR WEB SEARCH

---

many of the other framework components.

The proposed framework is segmented into four main components which are seen as core to the development of Web search environments to reduce risk of harm to both the searcher and society.

**FC-Policy** *Policy*, which includes methods of law, education and corporate policy, are suggested.

**FC-Cognitive** A set of *behavioural and cognitive* approaches, developed for the specific purpose of engaging decision making that reduces risk to individuals and society are introduced.

**FC-System** Considerations for the *system design* are provided, for which content enrichment and interface design are the main focus.

**FC-Evaluation** Any framework, and any approach for that matter, needs to be *evaluated*, for which suggestions are also given.

The considerations provided are not exhaustive, and are intended as a foundation for a way forward.

#### 3.2.1 Policy

Policy (**FC-Policy**) is a broad topic, which encompasses relevant areas such as law and education and can be used as a mechanism to prevent harm in Web search. Policies set by the Web search systems are used as a means to leverage their commercial, legal and overall organizational interests. For instance, explicit privacy



policies and data usage policies may be tailored to protect the provider from legal ramifications (e.g. the GDPR), while simultaneously maximizing their commercial profits [267]. Alternatively, a provider may shift their policy to meet social norms and address public outcry from issues such as misinformation [48]. There are also policy decisions around design choices for the core product that searchers interact with, such as how to present information in a SERP, search support tools to include in the product (e.g. query suggestion), and underlying retrieval and ranking models to implement. A recent study (see [67]) recommends that design policy for Web systems to be developed with the key principal of human control to opt-out and have algorithmic control.

Laws can be implemented, locally, nationally, within economic regions and globally, with examples including California, Canada, European Union and human rights law respectively. For Web searchers, the GDPR [66], a law which is designed to better protect their privacy and allow for greater transparency and control over how data collected about them is used, is perhaps the most well known law for harm prevention to date. Laws may also be used to enforce information providers (e.g. social media and search platforms) to take down and / or filter out content that is perceived by law makers to be harmful to individuals and or societies. Examples include the NetzDG Germany [199] that require removal of speech that is hateful (e.g. Nazi imagery) and censoring of Google search results (e.g. websites mentioning the Tienanmen Square massacre) by the Chinese government [224]. Some laws, such as the communications and decency act in the USA [236], place the onus of legal liability on the publisher of the content (e.g. author of news article), but not the provider (e.g. search engine) or user (e.g. searcher). Legal tools achieve harm prevention through penalty (e.g. fines, imprisonment) for non-adherence to the rules stated by law, they also have a sense of authoritarianism and dictating what is good

### 3. A HARM INTERVENTION FRAMEWORK FOR WEB SEARCH

---

or bad. As differentiating between good and bad can be problematic [74], we suggest that law be used as a tool of last resort. Nonetheless, ethical considerations are a critical factor to IIR systems and research and therefore must be taken into account [119, 245]. As part of the framework, basic universal human rights [221, 222, 223] are the recommended lens through which policy is set.

Finally, education approaches and campaigns are a suggested pathway to improve search capabilities that minimize personal harm. There are some efforts by platforms to provide education tools and programs in primary and secondary schooling (see [85]) as well as being broadcast to searchers of any age (see [85, 190]). However, a searcher is not provided such tools directly in the search engines (i.e. there is no link provided)<sup>1</sup>. In our view, education is a promising pathway, as it overlaps with the cognitive interventions discussed in the section that follows.

#### 3.2.2 Behavioural and Cognitive Interventions

We suggest that behavioural and cognitive interventions (**FC-Cognitive**) are the most overlooked component of our framework. Much was already introduced with respect to the possibilities for such interventions (see background Section 2.6), for which *nudging* and *boosting* were given particular focus. The reason for this focus being that both are successful methods in other domains (such as medicine and public well being) and furthermore both are very different in their implementation style and underlying theory driving them. There is furthermore an axis of transparency to consider when implementing such interventions, where they can be quite transparent (in the case of the nutrition labels) or quite non-transparent (e.g. search settings being defaulted to Adult safe search).

---

<sup>1</sup>Query recommendations and spelling corrections may be educational if it can be shown that the user improves their query behaviour or their spelling over time.

This distinction between transparent and non-transparent interventions leads to one key difference between *nudging* and *boosting*, being that *boosts* are always transparent in their goal, whereas this is only sometimes the case for *nudges*, which as pointed out earlier is an ethical concern. Another problem with *nudging* being that skills are not taught to the individual, and therefore the intervention is no longer useful once removed. It is not so straightforward though, one cannot just throw out *nudging* because of these problems, there are cases where one approach may be better than the other [93, 140]. This dichotomy of *nudging* and *boosting* makes them quite interesting, and furthermore strongly suggests that both should be considered in the evaluation of any harm prevention initiative.

### 3.2.3 Search System Design

Much was already said about the interplay of the system and the user in the background on IIR (see Section 2.5). Some specifics, related to the search system, were also introduced (Section 2.3), such as information extraction and retrieval models. Such factors are important for the advancement and development of Web search systems, such as commercial search engines, designed to reduce harm.

Many components are necessary to build a functional systems such as search engine [51] and it is clear that the underlying components of these systems have a tendency to become biased and steer users towards harmful information [16].

In this third component of the framework (FC-System), we focus on content enrichment and the search interface and provide limited discussion on other components, such as data logging and retrieval models as indicators as to where they might play a role within the framework. Based on commonly used implementation methods, such as those leveraging query logs as a primary means to model searchers

### 3. A HARM INTERVENTION FRAMEWORK FOR WEB SEARCH

---

and provide support tools (e.g. query recommendations, collaborative search models) [51, 245], it is conceivable that many of the system biases currently present [16] would abate over time due to the logging of interactions of a subset of users gaining advantage through the system elements discussed in sections below. These system components, combined with a subgroup of users that take the effort to minimize personal harm with well implemented behavioural and cognitive interventions, would quite likely provide additional benefits to all users.

#### 3.2.3.1 Content Enrichment and Informational Cues

Processes that enrich and classify information in Web documents are fundamental to modern search engines and IR systems [51]. As previous research indicates, there are many different cues to consider [74, 131, 208] during this enrichment process to be applied to information that may be useful for minimizing risk to searchers, for which many are important factors for making better decisions in search [94, 131, 207].

For IR researchers and data scientists, the listing of cues provided by Smith and Rieh (see [208]) as well as methods outlined by Fuhr et al (see [74]) are useful guides for development of cue extraction methods. Methods are already available for extracting cues such as the reading level, the virality of the content (i.e. how likely will the information spread), emotionality (e.g. language that is angry, overly positive, etc.), prevalence of factual, opinionated and / or controversial information, trustworthiness of the source (e.g. mechanisms to determine the credibility of a Web page), technicality (e.g. a score for amount of technical jargon in document) and if the document is currently topically relevant [74].

Bibliographic cues (e.g. author affiliation) and inferential cues (e.g. citations to and from document) are also needed for critical thinking and evaluation of information [208]. A lack of transparency exists in affiliations of authors and publishers of

information [208], and therefore there is a clear need for developing methods that evaluate affiliation(s) of authors and publishers of information (e.g. who is funding think tank X that publishes web page Y) [133].

Methods to identify content that is hateful [261], misogynistic [21] or containing vulgar language [53] are also readily available and potentially useful for minimizing exposure to content that some users may find offensive, which Smith and Rieh [208] classify as valence cues. Marking content which is sexually explicit (written, verbally and / or visually) [174], may be useful for developing strategies to minimize harms to minors as well as users that are susceptible to addiction.

As privacy is of paramount concern too, extracting cues that provide greater transparency to the searcher into what data is collected and by whom it is collected and shared are also critical to prevent harms from the collected information. One such task in this space is the identification of 3rd parties that data will be shared with when visiting a Web page [134, 257] and another being the classification of privacy statements on the websites where the content is hosted, a task that could be designed with existing privacy statement corpora [250]. In a similar vein as author affiliation cues proposed by [208] and [133], privacy-based ontologies containing information (e.g. total fines, number of GDPR violations) about 1st-party providers and the 3rd-party affiliations could be developed to present privacy cues.

Many of the cues can be extracted with models produced by machine learning algorithms. Nevertheless, for data scientists that develop these models, it is critical to minimize model bias, such as gender bias [32]. Many solutions in the field of IR are designed with a “find the best” model mindset, as evidenced by the leaderboard approach for shared tasks (e.g. TREC, SemEval), and are a likely cause of some model biases and subsequent poor predictions. There is evidence that ensemble

### 3. A HARM INTERVENTION FRAMEWORK FOR WEB SEARCH

---

approaches are more robust and resilient to bias and more likely to outperform a single model [89, 141, 261], and are one possible alternative. In search spaces with potentially dangerous outcomes (e.g. health), data scientists should also consider interpretable models [189].

#### 3.2.3.2 Interface Design

Informational cues, cognitive interventions and policy are all important for harm reductions [140], but they need a medium for implementation and it is the search interface (such as a SERP) that is this medium.

Extracting cues that allow for the design of better decision making tools (thus enabling users to better tap into their critical thinking skills) and designing interfaces that present such cues and tools in a not-too-disruptive manner are two major challenges for interface design. Commercial SERPs are typically presented as ranked lists [208] and, depending on the query, will contain content such as advertisements, social media posts and news articles [14]. Search support tools are an important IR system component for improving search [245], some of which are available within the SERP including query suggestions and auto-completion as well as spelling correction. Thus, any component that allows the user to minimize the chance of harm, also falls within the scope of search support.

Space is a premium and one challenge is to ensure that the screen is not overloaded [146]. Risk communication tools such as nutrition labels and fact boxes are highly effective and desirable, but may not fit on small mobile devices, where warning lights are likely the better option. Link enrichment is another approach [146], where pop-ups populated with informational cues are included with the results, and is thus especially appealing as it could be applied to both desktop and mobile search. Link enrichment also need not apply only to the SERP, and can be applied as searchers

navigate within [8] or across domains and the Web (Wikipedia desktop offers link enrichment and is one live example).

Alternatively, the SERP could be designed to rank or filter results as to attenuate possible harms from, say, privacy concerns or dangerous medical advice. Indeed, commercial search engines already offer the default of filtering adult content (e.g. content that is classified as sexually explicit), and takes up little space within the interface. Altering results in the SERP in this manner is a *nudge*, so long as the user is given the capability to opt-out [220]. However, we caution against such approaches, as it does not tap into the important critical thinking and literacy skills of searchers [208] and thus likely does not generalize to other contexts without such a *nudge*.

The interface is also where policy can be implemented. It is conceivable that law makers may someday require IR systems to include any number of the approaches already discussed—an information nutrition label [74] is one such possibility. Or platforms, such as Google, could voluntarily set policy that provides links to educational resources in the SERP, simplifying the process for searchers to learn how to better protect themselves during the search process.

### 3.2.3.3 Additional System Approaches

There are many additional system based approaches one might consider for harm reduction in Web search, for which we highlight several examples, but go not further as our focus is on other elements.

**Reducing Information Collected** As stated in previous sections, there are benefits and risks to the information collected about individuals online. Approaches to

### 3. A HARM INTERVENTION FRAMEWORK FOR WEB SEARCH

---

scramble the interactive data collected about users and simultaneously consider the loss of benefits due to this approach appear promising [29]. More recent research, in the domain of search and recommender systems, suggests a lower number of features are necessary for high quality results [28] and is therefore a pathway towards reduced data collection.

Third party companies regularly collect information about users during their search for information resulting in many privacy risks [62, 63, 117, 149]. Methods are available to algorithmically detect 3rd party tracking (e.g. via cookies, embedded scripts and browser fingerprinting) in Web pages [30, 257]. Additionally, many easy to install tools also exist to prevent 3rd party tracking (e.g. Ghostery and Privacy Badger)<sup>2</sup>, which greatly reduce but do not entirely prevent tracking [69].

**De-biasing Language Models** For classification tasks related information encountered on the Web (Sections 2.4.1 and 3.2.3.1), there are risks of discrimination due to underlying language biases (e.g. due to locality and demographic differences by authors) existent in collections used to develop train such models. The emerging area of research referred to as Computational Sociolinguistics addresses such issues. The work of Hovy et al. demonstrates the importance of incorporating such information into document models [98, 109]. Word sense use differences are easily captured in tweets where geo-location of postings [166] and furthermore when underlying embedding models are built upon text from a large twitter corpus [19, 47]. Other factors, such as time, may also play an important factor [176]. New methods to test statistical significance of location, demonstrate that location has an influencing role in language [129]. Intuitively, one would ask if location can be inferred from language used online and recent results show just this [153]. Additional areas

---

<sup>2</sup>See <https://www.ghostery.com/> and <https://www.eff.org/privacybadger>



of computational sociolinguistics are further outlined in [161].

### 3.2.4 Evaluation

Evaluation ([FC-Evaluation](#)) of Web search is a large topic in and of itself, for which much was already covered (see Section 2.7) with respect to traditional approaches for evaluation of systems and interactive behaviours. Such evaluation approaches should not be overlooked with respect to this framework and utilized where appropriate. Attention should also be placed on more recently proposed metrics, including metrics that take an economic view, along with considerations for a newer generation of outcome-based measures for harm.

#### 3.2.4.1 Economic-Based Evaluation

Interventions that reduce risk of harm, such as those suggested in the framework, have costs (e.g. time) for the individual [92, 220] and costs are an important economic consideration for IIR environments [12, 15]<sup>3</sup>. The economic view has inspired a new set of useful evaluation approaches, which integrate theories from economics and have the overall aim to better predict user behaviour in the search environment [11]. Incorporation of the economic view of IIR is potentially useful for evaluating the framework, as it allows evaluation from the perspective of trade-offs of costs and benefits [11, 14], such as the trade-off of costs of time for the benefit of reduced risk of harm as part of the search process. In addition to time, examples of relevant costs one might consider are the money a searcher is willing to pay for information that is of high quality, amount of data they are willing to share with 3rd parties and the effort of searching for information relevant to their task.

---

<sup>3</sup>There are costs with respect to designing and operationalising interventions in a search environment, such as salaries for software engineers. However, for this discussion, we are strictly concerned with the economics of the searcher.

#### 3.2.4.2 Harm Outcome Evaluation

There are many different *search outcomes* that might occur due interactions with the Web search environment, related to factors such as health [179, 244], politics [65] and privacy [1], to name a few. Identifying approaches to measure such outcomes is important, and in our view overlooked, as indicated by the heavy focus on relevance centric measures introduced in Section 2.7. Studies investigating health outcomes [179] and political impacts [65] of search demonstrate how easily one can be manipulated, but also provide insights as to how one might measure such outcomes beyond the lab. For instance, in a similar vein to the longitudinal Harvard nurse study<sup>4</sup>, we might someday link positive and negative outcomes (e.g. health outcomes investigated by Pogacar et al. [179]) to longitudinal behaviour including a broad set of Web searchers.

Sense-making, introduced in section 2.2.3), an area of research within IS that considers the process of filling in gaps of knowledge, also has a strong focus on the ultimate outcome of this process [56], outcomes which may have positive or negative impacts [56, 185].

Such impacts touch upon evaluation of the system effectiveness [51], success [245], usability [119] and performance [119]. Our framework suggests evaluation through the lens of harm as a dimension cutting across the traditional IR evaluation dimensions.

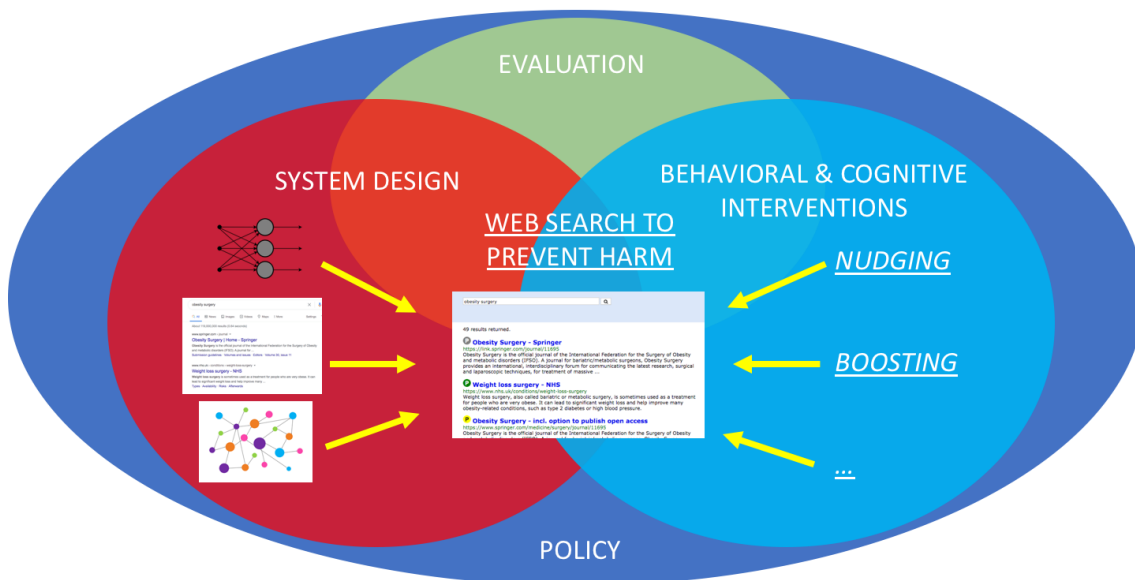
#### 3.2.5 Framework Combined

It is useful to show how these components fit together and Figure 3.1 provides such a view. As stated previously, we view policy as the foundation to design of

---

<sup>4</sup>See [Nurses' Health Study](https://www.nurseshealthstudy.org/) at <https://www.nurseshealthstudy.org/> (LA: 2020-30-10)

Web search environments taking into consideration harm prevention, and hence why it underpins the three other components. The other three components are not independent from one another and have overlap. It is within this overlap of components where Web search environments for reduced harm originate.



**Figure 3.1: Framework for Harm Prevention in Web Search** - Four components are included in the framework for harm prevention: policy, system design, system evaluation and harm interventions from the behavioural and cognitive sciences.

### 3.3 Examples in Practice

Several empirical studies and commercial systems have some (but not all) of the elements of the framework. It is worth noting these to provide a lens into how the framework can be used in practice. To our knowledge, however, no approach addresses all four components. *Behavioural interventions (nudges)* and system-based *content enrichment* were shown to effectively steer users towards healthier food choices [61] and away from Websites that more greatly impact personal privacy [1]. However, all of these lack the *policy* element as they were evaluated in a lab setting.

### 3. A HARM INTERVENTION FRAMEWORK FOR WEB SEARCH

---

Some commercial search engines (e.g. DuckDuckGo), have used *policy*, *system design* and *cognitive* approaches to protect users from adult material, but do not publish their *evaluation* approaches.

Specific to *behavioural and cognitive interventions*, there are additional empirical findings worth noting. A subset of the cues suggested by Smith and Rieh [208] were used to augment search results visually to *nudge* users to more accurately assess credible information [200]. Browser plug-ins can provide a visual *nudge* during Web browsing and exploration, such as the Ghostery<sup>5</sup> 3rd-party blocking tool, which by default blocks data sharing with 3rd parties. However, caution should prevail with 3rd-party blocking tools, as recent findings question their effectiveness [69].

In line with a *boosting* approach, one study evaluated low-cost search tips as a means to provide skills for better searching [158] and another study improved novice searchers skills by feedback based their search behaviour compared to expert searchers [22]; note that neither study was explicitly designed for harm reduction nor explicitly referred to *boosting*.

#### 3.4 Summary

This framework is the lens through which the general methods, introduced in the next chapter, and the five empirical studies that follow (Chapters 5 - 9), were designed and evaluated. The framework allows much room for experimentation, with the core focus being on adaptation and evaluation of behavioural and cognitive interventions (FC-Cognitive) as evidence of their success in other domains and limited use in Web search suggest their untapped potential. Attention is also given to framework components FC-System and FC-Evaluation, as these are integral

---

<sup>5</sup>Ghostery at <https://www.ghostery.com/> (LA: 2020-10-30)

to development of Web search environments. Finally, as stated before, though fundamental, we do not focus on policy ([FC-Policy](#)), as we view this component as beyond the boundaries of scientific research. Scientific research should inform policy, but not set it.

Looking ahead, we initially introduce the general methods, where many of the specifics to the harm prevention strategies are introduced, of which *nudging* and *boosting* are the central focus. We also introduce evaluation metrics ([FC-Evaluation](#)) here, and additionally introduce some of the system design features ([FC-System](#)) which are used as the medium of implementation for [FC-Cognitive](#).

In the subsequent empirical studies, investigation of *nudging* and *boosting* strategies ([FC-Cognitive](#)) and search system design ([FC-System](#)) are the main focus for Chapters 5, 6, 8 and 9. These studies are evaluated with existing and some newly introduced evaluation metrics (component [FC-Evaluation](#)). Chapter 7 shifts the primary focus to [FC-System](#) for the identification of informational cues (as introduced in system design Section 3.2.3.1), which are used in studies that make use of both *boosting* and *nudging*.

As one might begin to see from this summary, these framework components are not standalone.

### 3. A HARM INTERVENTION FRAMEWORK FOR WEB SEARCH

---

## Part II

# Experiments





# Chapter 4

## Methodology

### 4.1 Overview

Task based evaluation was performed to evaluate four harm prevention strategies for Web search. Using both FC-Cognitive and FC-System, three experimental *nudges* and one experimental *boost* to prevent harms due to loss of privacy during the search process are designed. The four strategies are as follows:

- [S1] Filtering *nudge* of to remove results with high privacy risk.
- [S2] Re-ranking *nudge* to place results with lowest privacy risk at the top and higher privacy threats deeper in rank.
- [S3] Stoplight *nudge* with coloured lights indicating levels of privacy risk.
- [S4] Fact box *boost* to teach a skill for reduced privacy risk.

The overall methodology was progressive in nature, that was findings from each study are used to refine and develop methods for later studies. Online and offline

## 4. METHODOLOGY

---

lab based studies are the primary focus of our evaluation and methodology.

The methodology outlined in the sections below most heavily apply to the online studies (Chapter 6) used to test strategies [S1] - [S3], however many of these methods are used (or adapted) to other studies. Offline studies for the same strategies ([S1] - [S3]) were performed as a precursor to the online studies, for which methodology specific to those studies are provided in Chapter 6.

Development of strategy [S4] makes use of test sets and data collected from the online *nudge* studies. Though many of the methods below are useful for this strategy, there are many deviations as well. Methods specific to the development and evaluation of strategy [S4] are detailed in Chapters 7 and 8.

The search systems developed and linked to each of these strategies touch upon both the interface design (Section 3.2.3.2) and informational cue (Section 3.2.3.1) specifics of the system framework component.

Related to the evaluation of the strategies [FC-Evaluation] plays a major part in the studies as well. Existing measures for evaluation, along with many new measures are introduced in the general methods along with additional measures in each study.

### 4.1.1 Hypotheses for General Research Questions

Returning to the introduction of *boosting* and *nudging* in Section 2.6, strategies [S1] and [S2] are classic *nudges*. They are classic in the sense that the environment is modified to reduce risk in a non-transparent manner to achieve their overall goal (reduced privacy impact during Web search). Strategy [S3] is an ‘educative’ *nudge* approach, and is ‘educative’ in that the warning lights and the associated definitions of privacy risk provide some transparency to the individual about the goal of reduced

privacy impact. The distinction between non-transparent and transparent *nudges* is important as it is expected that individual beliefs (and actions) with respect to the harm prevention goal are ultimately linked to motivation for uptake of such strategy (e.g. users that care more about privacy, will more greatly make use of strategy [S3](#)).

Published findings of *nudging* in other domains (e.g. placement of healthy food in a cafeteria to produce healthier eating habits and ultimately improve medical outcomes) demonstrate that classic *nudges* are highly effective. Related to [G-RQ-2](#), it was expected that either strategy [S1](#) or [S2](#) will be most effective at reduction of privacy impacts.

However, there are other important factors to consider, including ethics and user preferences with respect to each strategy. These considerations are important for evaluation of [G-RQ-1](#). Though a strategy may be found to be highly effective (the aim of [G-RQ-2](#)), it may be found to be entirely non-viable (for instance if the strategy was deemed unethical). Ideally, a strategy was both highly effective and viable. Therefore, it was quite possible that strategies [S1](#) - [S3](#) are non-viable even though some or all are found to be highly effective.

*Boosts* were proposed as a response to the challenges associated with *nudges*. The first challenge being that classic *nudges*, such as strategies [S1](#) and [S2](#), are non-transparent and therefore unethical. And a second challenge towards transparent *nudges* (e.g. strategy [S3](#)), being they do not teach a skill for the individual to evaluate risk independently and therefore such a strategy will fail in the instance that it was removed (for example if a user switches from a search environment with warning lights to one without). These challenges provide motivation for the evaluation of [G-RQ-3](#), where it was expected that individuals will not only learn a

skill when treated with strategy [S4](#) but furthermore be able to apply the skill after the treatment was removed. For *nudge* based strategies (e.g. strategies [S1](#) - [S3](#)) no skill was taught and therefore individuals will not be expected to maintain reduced harm.

The most effective and viable *nudge* strategy was chosen for comparison and evaluation against the most effective *boost* approach (see Chapter 9). Development of the *boost* approach was undertaken in studies outlined in Chapter 7 and 8.

### 4.1.2 Adapting Methods from Previous Research

Web search encompasses many domains (e.g. medical, entertainment) and types of tasks (e.g. fact based, exploratory). Our methodology assumes that search in the medical domain was a highly private matter and furthermore can lead to grave outcomes if misinformation was encountered, therefore previously published search tasks from the medical domain were sought out.

Furthermore, as the proposed *nudge* and *boost* strategies were entirely novel to the space of Web search, it was desirable to maintain more control during empirical studies. Thus, we sought out medical search tasks that were also factual and closed-ended (as described in Section 2.2.6). We suggest that control was greatly reduced when using other tasks, such as open-ended tasks or tasks that are exploratory in nature. Obviously, there remains a great amount of opportunity for empirical research of these strategies using less controlled tasks (e.g. exploratory) and settings (e.g. naturalistic), but highly controlled environments are useful for establishing which of our proposed strategies are most effective and viable.

Methodology developed by Pogacar et. al [179] included a set of 10 medical search fact-based closed-ended tasks. The search tasks were all questions with definitive

answers backed by published high quality systematic reviews specific to each question (see Cochrane Systematic Reviews in section below), and therefore treated as ground truth. Their methodology included a study design to test multiple manipulations of the search environment (to understand encounters with misinformation and its impacts on search task outcomes). As our proposed strategies are in essence manipulations to the search environment, we made every effort to borrow as much from their methodology as possible.

However, it was important to highlight several key differences between our methodology and empirical focus with that of Pogacar et al. First, ours was a lens on outcomes related to both privacy and misinformation, not just misinformation. Furthermore, their research focuses on manipulations that increase exposure to misinformation to better understand user behaviour (e.g. trust bias) and what might to expect if search engines were to provide better search ranking (e.g. some user beliefs are so biased that no amount of correct information will produce a healthy search outcome). Conversely, our methodology investigates strategies that reduce privacy impacts as well as exposure to misinformation.

Major differences aside, their methodology was a highly relevant and useful starting point for our methodology and empirical design and was therefore mentioned frequent in the sections below. Specific to the search tasks and *search outcomes* they evaluated, we make some cross-comparisons of our findings with theirs in several of our empirical studies.

### 4.1.3 Cochrane Systematic Medical Reviews

Given the reliance on Cochrane medical reviews for the formulation of the search tasks, it is useful to shed some light on the Cochrane organization, what their reviews

## 4. METHODOLOGY

---

entail and why they can be used as a gold standard for medical search tasks. For a start, we provide the verbatim summary of their organization from [Cochrane About Us](#)<sup>1</sup>. We have *enhanced* the text that is most relevant to the introduction.

“Cochrane is for anyone interested in using high-quality information to *make health decisions*. Whether you are a doctor or nurse, patient or carer, researcher or funder, Cochrane evidence provides a powerful tool to enhance your healthcare knowledge and *decision making*.

Cochrane’s members and supporters come from more than 130 countries, worldwide. Our volunteers and contributors are researchers, health professionals, patients, carers, and people passionate about *improving health outcomes* for everyone, everywhere. Our global independent network gathers and summarizes the best evidence from research to help you make *informed choices* about treatment and we have been doing this for 25 years.

*We do not accept commercial or conflicted funding.* This is vital for us to generate authoritative and reliable information, working freely, unconstrained by commercial and financial interests.”

Cochrane performs independent systematic medical reviews on a multitude of health and medical topics (e.g. examining the effectiveness of melatonin for a jet lag, effectiveness of SSRIs for treating depression). The process of a systematic is highly involved and time consuming. The review involves researchers (e.g. medical professionals in the case of Cochrane) retrieving and examining the available body of literature (i.e. all published studies) on the medical treatment they are evaluating. Once attained, each publication in the body of literature is reviewed for important

---

<sup>1</sup>[Cochrane About Us](https://www.cochrane.org/about-us) at <https://www.cochrane.org/about-us> (LA: 2020-10-04)

factors including quality of methods (e.g. was the study double blind) and statistical analyses (e.g. statistical power was high quality in study X but not in study Y or study Z. Once everything is reviewed, a summary of findings is published stating whether the treatment reviewed is *effective*, *ineffective* or that *insufficient evidence* is available for the evaluation. A good example of the review process is provided in the findings for [the effectiveness of melatonin for jet lag](#)<sup>2</sup>.

To be consistent with previous research using search tasks based upon Cochrane findings (e.g. White et al. [246] and Pogacar et al. [179]), the outputs of Cochrane reviews were translated as follows for our empirical studies:

- An *effective* medical treatment for the purposes of search task is defined as *helpful*.
- An *ineffective* medical treatment for the purposes of search task is defined as *does not help*.
- If *insufficient* evidence is found to determine effectiveness of medical treatment, for the purposes of search task is defined as *inconclusive*.

Returning to the *enhanced* text in the above quote, there are three important points to be extracted.

First, better medical *decision making* is the ultimate goal of the systematic reviews. As already indicated in Section 2.5.5, *decision making* is fundamental to many of our everyday search task, and therefore the goals of Cochrane are nice fit with this process. Interestingly, the strategies we examine (*nudging* and *boosting*)

---

<sup>2</sup>see [Cochrane evaluates the effectiveness of melatonin for jet lag](https://www.cochrane.org/CD001520/DEPRESSN_melatonin-for-the-prevention-and-treatment-of-jet-lag) at [https://www.cochrane.org/CD001520/DEPRESSN\\_melatonin-for-the-prevention-and-treatment-of-jet-lag](https://www.cochrane.org/CD001520/DEPRESSN_melatonin-for-the-prevention-and-treatment-of-jet-lag) (LA: 2020-10-04)

## 4. METHODOLOGY

---

are designed and implemented regularly improve decision making related to health and medicine, and therefore a good fit with the work of Cochrane.

Second, Cochrane has a primary aim to improve *health outcomes*. Throughout the thesis, *search outcome* is a common theme for which improved *health outcomes* sits within. Linking the search tasks to the Cochrane reviews allows for an objective assessment of *search outcome*.

Finally, we turn to the statement regarding *commercial and conflicted funding*, which Cochrane does not accept. This statement indicates gives high confidence in their findings, for example that no pharmaceutical company is influencing Cochrane to suggest a treatment is not, and therefore can be treated as a gold standard. In fact, Cochrane takes their research a step further, where they will publish in their findings specific conflicts of interest about publications included in the review (e.g. publication X was funded by drug company Y). Furthermore, they may reduce the quality “score” of publications where conflicts have influenced the research (or entirely disregard the research). We argue this last point is absolutely crucial with respect to misinformation about medical treatments.

Sadly, recent events within the Cochrane organization raise serious concerns about the future quality of their reviews [86]. Nonetheless, we believe they are still high quality at this time and these three points combined make the reviews highly relevant for our overall research aims.



## 4.2 Procedure

### 4.2.1 Search Tasks

A total of 10 (+ 2 practice tasks) medical search tasks (see Table 4.1) were available for presentation to subjects (the same search tasks used by Pogacar et al. [179]) in an in-lab setting. During each search task, as with Pogacar et al., definitions for the health issue, medical treatment and for the decisions that could be made (*helpful*, *does not help* or *inconclusive*) were provided to subjects. Each search task was complete at the point when subjects submitted their decision (see Figure 4.5 for a visual of the search task decision).

Search tasks were based upon findings from systematic reviews published by the Cochrane medical review panel, with their findings in the rightmost column (see Table 4.1). All Cochrane findings for the medical questions used in our studies were *helpful* or *does not help* and no questions had *inconclusive* findings. Thus, if a participant selected *inconclusive* as the answer their decision was deemed *incorrect*. Participant search task decisions were deemed *harmful* if they chose the converse of the Cochrane finding (e.g. *Helpful* instead of *Does not help*), and deemed *correct* if their decision agreed with Cochrane.

Section 4.2.3 explains how the search tasks were presented to users to ensure a balanced experimental design.

### 4.2.2 Search Systems (and SERPs)

Through the lens of FC-System, the search engine and search engine results page (SERP)<sup>3</sup> used in all offline and online studies was a standard search interface

---

<sup>3</sup>SERP and search system are used interchangeably in the context of our studies

## 4. METHODOLOGY

---

Search Task	Medical Question	Cochrane Finding
Task 1	Do antioxidants help female sub fertility?	Unhelpful
Task 2	Do benzodiazepines help alcohol withdrawal?	Helpful
Task 3	Do insoles help back pain?	Unhelpful
Task 4	Do probiotics help treat eczema?	Unhelpful
Task 5	Do sealants prevent dental decay in the permanent teeth?	Helpful
Task 6	Does caffeine help asthma?	Helpful
Task 7	Does cinnamon help diabetes?	Unhelpful
Task 8	Does melatonin help treat and prevent jet lag?	Helpful
Task 9	Does surgery help obesity?	Helpful
Task 10	Does traction help low back pain?	Unhelpful
Practice 1	Do cranberries prevent urinary tract infections?	Unhelpful
Practice 2	Does Echinacea help treat and prevent the common cold?	Helpful

**Table 4.1:** All 10 search tasks users encountered in the experiment along with 2 additional practice tasks provided before the main experiment, each were based upon findings published by Cochrane.

modelled after popular search engines. A sticky panel was used at the top of the page to allow for search task questions and definitions related to the task to be visible during the scroll of results. A decision button was accessible at the bottom of the SERP for participants to click when ready to submit the answer for the search task.

Raw data collected via interactions with the SERPs included: all queries submitted (for online studies), web page(s) visited, the rank of each result, the pagination of SERP visits (e.g. page 3 of results) and the time stamp of each event. For each result visited the web page was opened in a new window within an i-frame and the time recorded when users returning to the results page (allowing for calculations of time spent on each website). Specific for analyses with respect to the warning light approach, a privacy impact warning light colour was assigned to each website for all SERP variants, however the colour was only visible to the user in the warning light SERP (see Figure 4.2).

A hamburger icon in the upper left of the sticky panel (see Figure 4.1) commonly seen in websites and search engines (e.g. Bing) to change settings was included in the SERP. Traditionally, a *nudge* requires an opt-out mechanism, and the hamburger icon served as this mechanism, which allowed users the ability to turn on and off the privacy settings. The default was set to privacy protection turned on. A minor aim of the current research was a gained understanding of differences between users that change these settings.

The details of the SERPs (Figures 4.1 - 4.4) used in the online studies, which provided a search bar, connected to a commercial search API, to freely enter queries for results from the Web, are included below. We begin with an overview of the control SERP, which was the basis for the 3 experimental SERPS described thereafter<sup>4</sup>. Figures 4.1 - 4.4 are the SERP variants a participant would have experienced on search task 9 (see full list of search tasks in Table 4.1) and had submitted the query "obesity surgery". web pages returned by this query would have been the same for all SERPs, however they were displayed in a different manner dependent on the strategy and associated SERP used to test the strategy.

A mapping of the harm prevention strategies to the corresponding SERP used to test the strategy.

- Strategy [S1](#) was evaluated with the SERP in Figure 4.3
- Strategy [S2](#) was evaluated with the SERP in Figure 4.4
- Strategy [S3](#) was evaluated with the SERP in Figure 4.2
- The SERP used to test strategy [S4](#) was developed and evaluated in empirical studies outlined in Chapters 8 and 9

---

<sup>4</sup>Figures are best seen in colour.

## 4. METHODOLOGY

---

Following the descriptions of the control and experimental SERPs was a description of the search task decision page (Figure 4.5). For all search tasks, users would have landed on this page. In line with methods by Pogacar et al., two search tasks took the user directly to the decision page without a SERP and furthermore provided no additional information to answer the medical question. In this manner, a baseline performance metric could be determined on the search task outcome.

### 4.2.2.1 Control SERP / Search Environment

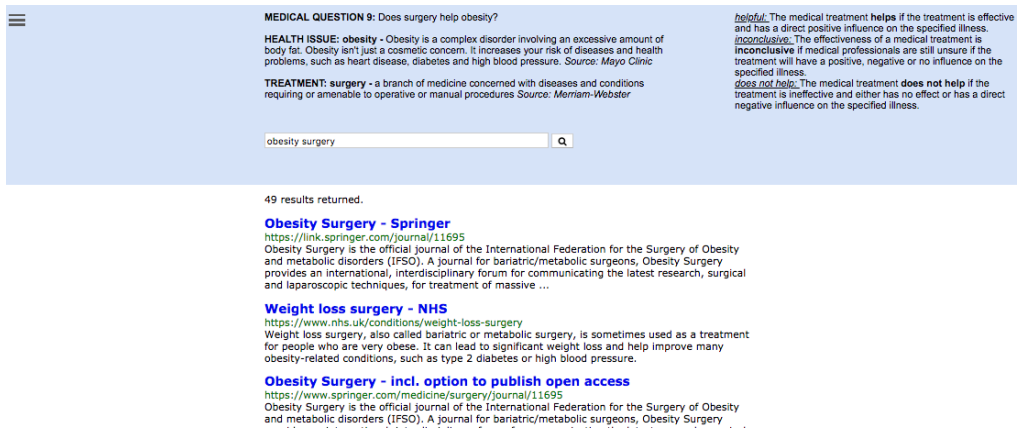
The Control SERP (Figure 4.1) was the basis for the 3 experimental SERPs described further below (Figures 4.2 - 4.4)<sup>5</sup> as well as the SERP introduced in Chapter 9 to test strategy [S4]. Capturing interactions with a control SERP is crucial for the evaluation of all four harm preventions strategies ([S1] - [S4]), as comparison across the four strategies alone was not a sufficient evaluation. The control was necessary to demonstrate if any of the four strategies are a statistically meaningful improvement over the status quo (e.g. current commercial search engines).

In the SERP, the search task, medical issue and treatment definitions are visible in the top panel as well as the decision options available in the top right. For the Control SERP, no manipulations were made to the results ordering as returned from the commercial search API (i.e. the result rankings from the commercial search API were kept intact when presented to the user).

We now turn to introduction of the three SERPs used for evaluating the *nudge* strategies. As it simplifies understanding of the SERPs used to test the non-transparent strategies [S1] and [S2], we begin with describing the Stoplight SERP and search system to test the transparent strategy [S3].

---

<sup>5</sup>Figures are best seen in colour.



**Figure 4.1: Control Search System (SERP) with No Manipulation of Results and No Intervention for Harm Prevention.** - This SERP was designed to simulate current commercial search engines (e.g. Google) and was used for the analysis of the four strategies [S1](#) - [S4](#).

#### 4.2.2.2 Stoplight SERP / Search Environment

Strategy [S3](#) was evaluated with the SERP in Figure 4.2. Recall that this *nudge* strategy is ‘educative’ and therefore gives transparency into the harm it aims to prevent (in this case privacy). The notion of “transparency” as it relates to the strategy being evaluated is an important consideration across participants, as the background (Section 2.6) indicated that one should expect motivation and other factors to play into the effectiveness of such interventions across the population.

The design of the Stoplight strategy is inspired through findings in nutrition studies that encourage healthier eating behaviour [147, 203, 237], which show much better results and satisfaction by users as compared to a nutrition table (such as a proposed approach for use in IR [74]).

The results for the Stoplight warning SERP are ordered in exactly the same manner as the Control SERP. Specific to the online experiments, 3rd party tracking data was only available for a subset of the results returned from the API (see further discussion about linking 3rd party data to test sets in Section 4.3). For any result

## 4. METHODOLOGY

---

where a 3rd party privacy tracker could not be linked, a Gray light was assigned. For the remaining results where trackers were known, quartile values of number of *3rd party trackers* we were calculated. Anything below the median number of trackers was assigned a Green light, above the median but below the upper quartile assigned Yellow, with Red assigned to any result in the upper quartile of trackers.

List below are the definitions of the Stoplights visible to users in when using the SERP.

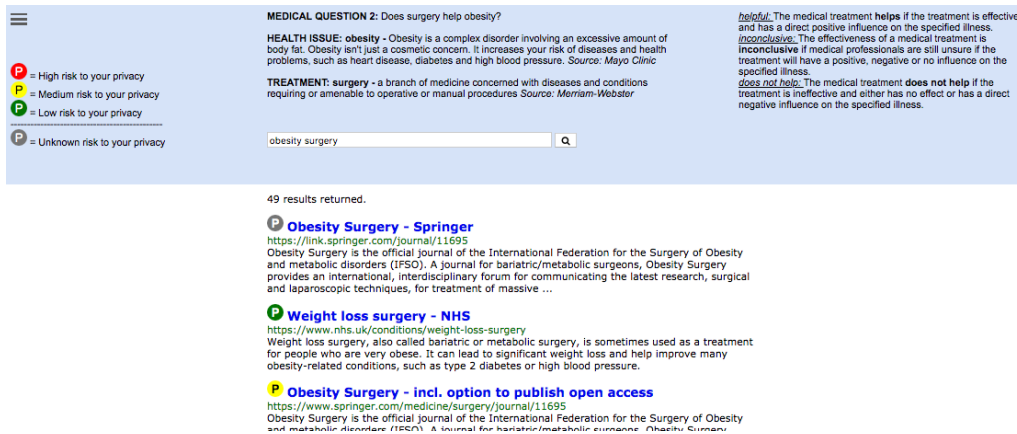
- **Red** = High risk to your privacy
- **Yellow** = Medium risk to your privacy
- **Green** = Low risk to your privacy
- **Gray** = Unknown risk to your privacy

The association of colours to the number of trackers was motivated by the skewed nature of 3rd party tracking data used in the studies, such that a result with a Green Stoplight roughly contains 0 - 4 trackers, Yellow containing 4 - 8 trackers and Red linked to 8 - 30 + trackers. The reader should note these values are rough estimates, which are dependent on factors including the search task at hand, the query submitted, and availability of 3rd party tracking data for a particular result (which are not always available, as explained in Section 4.3). Additionally, though 3rd party tracking data was similarly skewed in the offline studies in Chapter 5, the evaluation test sets used in these studies were static and therefore variations were consistent for all search tasks and strategy combinations.

Finally, we assume that the Gray light will be interpreted as more risky than the Green light, for which analysis was included in Chapter 6 with respect to this

assumption. However to reduce user bias, we placed it below a line that separated the other warning lights.

See Figure 4.2 for additional details and definitions for the Stoplight SERP.



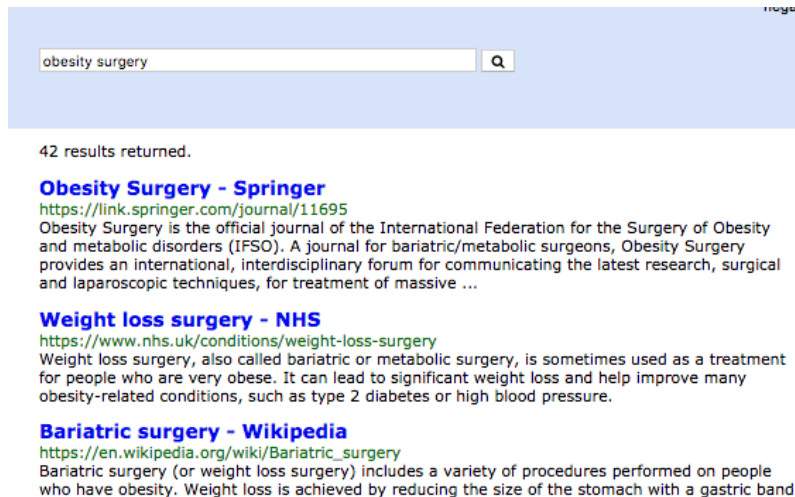
**Figure 4.2: Stoplight *Nudge* Search System (SERP) with Stoplights (best viewed in colour) Warning About Levels of Privacy Risks.** - The Stoplight SERP was used for the evaluation of harm prevention strategy [S3].

#### 4.2.2.3 Filtering SERP / Search Environment

Strategy [S1] was evaluated with the SERP design provided in Figure 4.3. As with the Stoplight and Control SERP, original result ordering was maintained (i.e. the ordering of the commercial search API response). Visually the filtering SERP appears the same as the control SERP. However, any results containing above the median number of 3rd party trackers for the result set (i.e. the set of results returned for the submitted query) were filtered out of the SERP. For comparison to the Stoplight strategy in Figure 4.2, when imagining the warning lights as visible for the Filtering SERP, users would only see results associated with Green and Gray lights.

## 4. METHODOLOGY

---

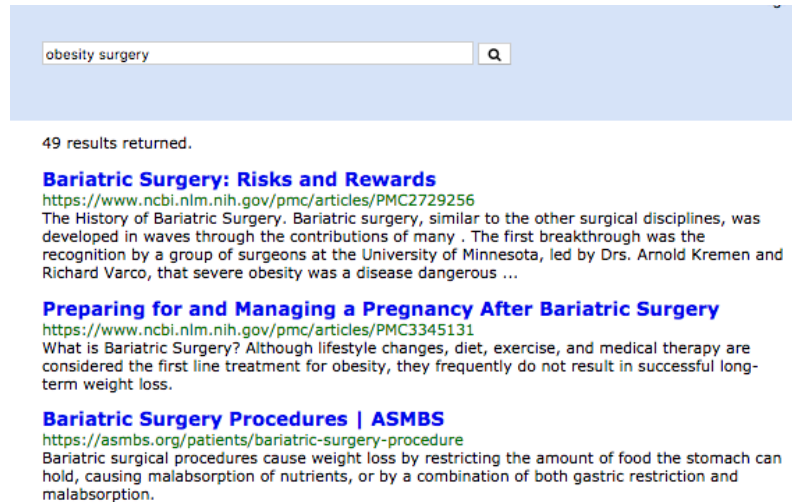


**Figure 4.3: Filtering *Nudge* Search System (SERP) for Reduced Privacy Risk.** - The Filtering SERP was used for the evaluation of harm prevention strategy [S1](#).

### 4.2.2.4 Ranking SERP / Search Environment

The Re-ranking SERP in Figure 4.4 was designed to test strategy [S2](#). Prior to results being displayed in the SERP, results are Re-ranked based on the number of *3rd party trackers* associated with each result. In this manner, the first result always had the lowest number of *3rd party trackers*. For cases where 2 or more websites had equal numbers, the same original ordering was kept. To cope with websites where 3rd party trackers were unknown, such results were placed between results with the median number of trackers and results just above the median number of trackers. For a visual reference, imagine all of the Stoplights being made visible in the ranking SERP, in which Green lights would appear first, followed by the Gray lights, then Yellow lights and finally Red lights. The placement of results was done in this manner to maintain consistency with other variants and because it was assumed that Gray lights would be perceived by individuals as more risky than Green lights, but not as risky as the Yellow or Red lights.





**Figure 4.4: Re-Ranking *Nudge* Search System (SERP) for Reduced Privacy Risk.** - The Re-ranking system used for the evaluation of harm prevention strategy [S2](#). The first result has the lowest number of privacy trackers of all results returned for this query. Note that the top result in this SERP, does not match the top result in the other SERPs, was not the top result in this SERP.

#### 4.2.2.5 Search Task Decision Page

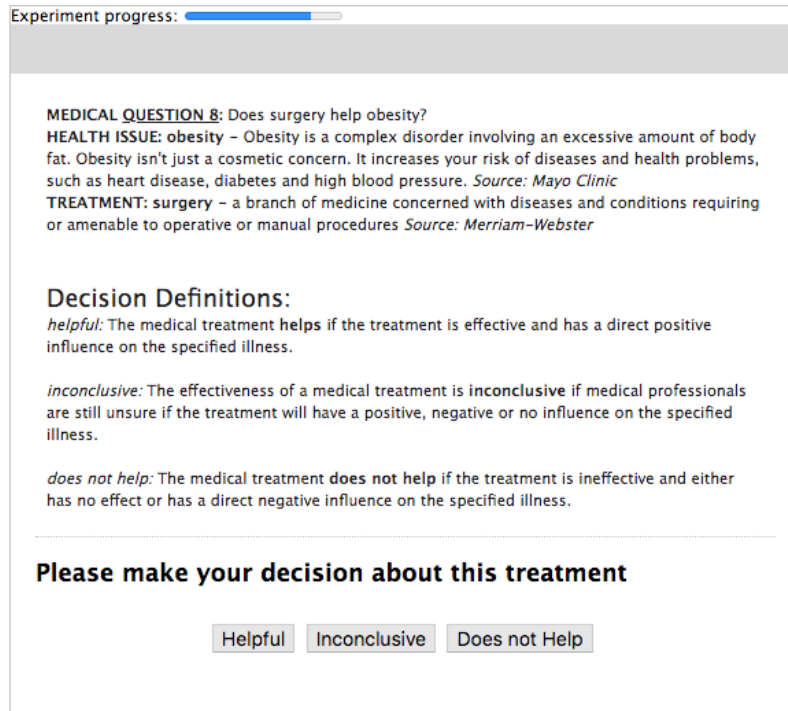
For all search tasks, subjects were asked to make a decision about the medical question at hand before continuing onto the next search task. See Figure 4.5 to see how the question was displayed to the participant along with the answers (*helpful*, *does not help* or *inconclusive*) available to them. Participants could only select one answer and they could not return to the page after submission. After their answer was submitted, they were directed to post-task questions to capture information such as the confidence in their answer.

For two search tasks, participants were taken directly to this page without a SERP to produce a baseline performance metric for the search task. For the remaining eight search tasks, one of the four SERPs already introduced would have been used to assist the participant in the process to best answer the question.

Similar SERPs were used in the offline studies (see Chapters 5 and 9), however

## 4. METHODOLOGY

---



Experiment progress:

**MEDICAL QUESTION 8:** Does surgery help obesity?  
**HEALTH ISSUE: obesity** – Obesity is a complex disorder involving an excessive amount of body fat. Obesity isn't just a cosmetic concern. It increases your risk of diseases and health problems, such as heart disease, diabetes and high blood pressure. *Source: Mayo Clinic*  
**TREATMENT: surgery** – a branch of medicine concerned with diseases and conditions requiring or amenable to operative or manual procedures *Source: Merriam-Webster*

**Decision Definitions:**  
*helpful:* The medical treatment **helps** if the treatment is effective and has a direct positive influence on the specified illness.  
*inconclusive:* The effectiveness of a medical treatment is **inconclusive** if medical professionals are still unsure if the treatment will have a positive, negative or no influence on the specified illness.  
*does not help:* The medical treatment **does not help** if the treatment is ineffective and either has no effect or has a direct negative influence on the specified illness.

---

**Please make your decision about this treatment**

**Figure 4.5: Search Task Medical Decision Page.** - The medical decision was presented to all participants during all search tasks (for in lab experiments).

the SERPs in the offline studies did not provide a search bar for users to submit queries, and instead users were asked to imagine they had entered a query. The offline studies therefore have the benefit of being more controlled, while the online studies are step towards a naturalistic setting.

### 4.2.3 Design for User Studies

**Within and Between Group Design** Motivated by the knowledge that more variants can be evaluated with a minimal number of participants [119], a *within-group* design was used to collect subject data related to evaluation of strategies [S1] - [S3] with respect to strategies **G-RQ-1** and **G-RQ-2** in Chapters 5 and 6. In this manner, all participants encountered the 3 experimental SERPs, the Control SERP and Baseline (i.e. no SERP available for the task) with two search tasks assigned to each (a total of 10 tasks).

A *between-group* design was used to test interactions across self-report measures with respect to the transparent *nudge* and *boost* strategies ([S3] and [S4]). Furthermore, a *within-group* design was used to compare user behaviour in the treatment environment (Stoplight strategy) against the Control environment. This design was used in both offline and online studies covered in Chapters 5, 6 and 9.

A *between-group* design was used for a pilot study of the transparent *boost* strategy (strategy [S4]), which was used to compare three different variations of the same fact box designed to enable users with skills for privacy protection. The design was used as it was expected that one or more variations would produce a better learning effect (an important factor related to G-RQ-1 and G-RQ-2). This design was used in empirical studies covered in Chapter 8, for which additional details are provided.

To answer G-RQ-3 it was also necessary to run a *between-group* as treating a single participant with both *nudge* and *boost* strategies in the same experiment would make testing of hypotheses infeasible. Further details of this design are provided in Chapter 9.

**Presentation of Strategies and Search Tasks** To ensure a balanced design for the participants for offline and online lab based *nudge* studies related to strategies [S1] - [S3], a Graeco-Latin squares design (see details in Appendix A.2) was used for rotation of the SERP and search task variants, each subject had 2 search task question types (helpful / not-helpful) assigned to each SERP variant. For example, a participant encountered a *Helpful* and *Unhelpful* task taken from Table 4.1 in the Stoplight SERP.

In later studies that piloted the *boost* strategy (Chapter 8) and compared the

## 4. METHODOLOGY

---

*boost* and *nudge* strategies (Chapter 9), a combination of Latin squares and randomization was used to ensure a balanced presentation of search tasks and strategies. Additional details are provided within the Chapters specific to these studies.

**Participant Experience** Data collection for the interactive user based offline and online experiments related to *nudge* strategies [S1] - [S3] were performed in a lab and moderated by the same person throughout (the author of this thesis). Each participant came into the lab and was asked to silence and put away their electronic devices. After completing a paper consent form, subjects were logged into the experiment. All experiments were run via the Google Chrome web-browser in Incognito mode. Chrome settings were configured to ensure that no previous queries would appear from previous tasks and users. This browser connected to a web-application server (developed with Python Flask) that was hosted on a university server. Within the browser they were first taken to a set of instructions (with no indication that a main component of the experiment was about privacy) and then asked several questions to ensure their comprehension of the instructions. Only after successfully answering these questions, would they move into the search tasks, for which a practice task was provided. After completion of the search tasks, users were then automatically taken to a post experiment survey run in Qualtrics to capture demographics and other measures relevant to our study.

At the end of the Qualtrics survey, all users were given a debriefing page to provide the answers to search tasks as determined by the Cochrane medical review panel, information was also given regarding the privacy components of the experiment. They were also free to ask the experimenter any questions regarding the experiment. While subjects were informed the experiment would take 60-75 minutes, most finished in under 60 minutes.

The empirical study in Chapter 8 for piloting the *boost* Fact box strategy was instead performed online using a popular Crowdsourcing platform. Details of participant experiment are further covered within the study itself.

In the empirical studies comparing the most effective and viable *nudging* and *boosting* strategies (Chapter 9), an offline lab based experiment was performed in a similar manner to the lab based *nudge* studies described earlier. Therefore, participants had an almost identical experience. Variations to this experience are highlighted in Chapter 9.

### 4.3 Evaluation Test Sets

Evaluation test sets are necessary for the studies related to our research questions. The test sets evolved as the studies progressed, in that the number of documents available and annotations associated with those documents grew over time.

The empirical studies made use of the publicly available test set (see [179]) related to the search tasks used (see search tasks in 4.2.1). However, this test set was not fit for purpose with respect to our research questions and therefore modifications and updates were necessary. The first evolution of the test set, necessary for the in lab offline studies in Chapter 5, occurred prior to any studies being undertaken. The second (and largest) advancement took place before, during and immediately after the in lab online study in Chapter 6. Minor alterations were made to the test sets from this point onwards, as required for studies covered in other studies. For example, the documents were annotated at a less granular level for studies in Chapter 7.

The fundamental elements of the test sets thread across all studies were the annotations related to privacy and misinformation. In the sub-sections that follow,

## 4. METHODOLOGY

---

the annotation methods for privacy and misinformation for the test set used in the online study in Chapter 6 are provided, with deviations from these annotation methods covered as pertinent to individual studies. A separate sub-section describes the meta data associated with test set.

### 4.3.1 Web Page Privacy

As already indicated in background Chapter 2, *3rd party trackers* are one source of potential harm encountered during Web search. Therefore, for all studies, *3rd party trackers* were chosen as a relevant proxy for harm related to privacy. Many *3rd party trackers* bridge across different websites, however to maintain highly controlled experimentation the 3rd-party trackers for each website are treated independently from each other <sup>6</sup>. Two methods were used in our studies to annotate the test set with 3rd party privacy information, for which can be defined as automated and manual.

#### 4.3.1.1 Automated Privacy Annotations

Methods are available to determine the number of trackers once a web page was loaded (e.g. [257]), however no known methods of privacy risk are available for real-time assessment in the SERP itself. Furthermore, modern search engines return results in fractions of a second and thus users expect results to be returned quickly, providing motivation for a privacy risk score being available at runtime of the experiment. As the known web has billions (if not trillions) of web pages, it was infeasible to crawl the Web and determine the number of trackers for each website prior to our experiment. An extensive effort was made to ensure that as many query

---

<sup>6</sup>For example, Google DoubleClick was a tracker on many different websites, and was not independent if you visit two different websites using this tracker. In the experiments in studies that follow, each tracking encounter was treated as separate from encounters on different websites.

results as possible would be linked to 3rd party tracking data, however this was not possible for all web pages and details of how web pages were linked are provided herein.

We contacted the authors of [WhoTracks.me](#)<sup>7</sup> [117] for usage of their tracking data. Their website provides visualizations and tracking statistics for approximately 5000 of the most popular websites<sup>8</sup> visited by users with the [Ghostery](#)<sup>9</sup> 3rd party tracking plug-in installed in their browser. The authors provided us with additional tracking data, composed of the top 10,000 most frequently visited global domains by users that have installed the [Ghostery](#) 3rd party tracking browser plug-in. For the remainder of the thesis, this dataset is referred to as the [WhoTracks.me Test Set](#).

### 4.3.1.2 Manual Privacy Annotations

Across all studies (both online and offline), for any web page that could not be linked to domains in the [WhoTracks.me Test Set](#), the [Ghostery](#) privacy tracking tool was used to determine the number of trackers. The annotation process, a somewhat time consuming and tedious task, was performed by the author of this thesis. For each web page requiring tracker data, the URL was manually pasted into a Google chrome browser. The author would wait for completion of the page load (sometimes 10 seconds dependent on web site), then open the plug-in interface to manually review and sum the *3rd party trackers* collecting data on the web page.

---

<sup>7</sup>[WhoTracks.me](#) at <https://whotracks.me/> (LA: 2020-10-30)

<sup>8</sup>This data is available for researchers an open repository here: [GitHub repository](#) at <https://github.com/cliqz-oss/whotracks.me> (LA: 2020-10-30)

<sup>9</sup>[Ghostery](#) at <https://www.ghostery.com/> (LA: 2020-10-30)

### 4.3.2 Web Page Misinformation

Two post-graduate students, with previous experience annotating documents (not related to this thesis), were recruited to annotate all websites visited by subjects during the in lab online experiment. Each annotator was paid £0.25 for each website annotation. Two rounds of annotations took place. In the first round, each annotator independently judged each website based on the instructions provided (see Figure 4.6). In the second round, the annotators used the same instructions to jointly resolve any discrepancies between their first round annotations. Analysis of the annotations, including agreement measures, are provided in results Section 6.3.1.

**Instructions** Annotators were provided a spreadsheet containing URLs visited by participants and the associated search task (i.e. the Cochrane Medical question) and correct answer. They visited each URL in the list and followed the instructions provided in Figure 4.6. The annotation task was to choose one of the following four options about the web page with respect to the search task (definitions and assumptions of user impacts to search task outcomes are included). Only web pages associated with the 10 main search tasks (see Table 4.1) were annotated, that was URLs linked to practice tasks were not included in the annotation process.

- *Correct*

DEFINITION: Information on web page agrees with Cochrane findings.

ASSUMPTION: User visiting a *correct* web page has increased likelihood of *correct* search outcome.

- *Incorrect - Page Unavailable*

DEFINITION: Page does not load or error returned for result visited<sup>10</sup>.

---

<sup>10</sup>Multiple web pages returned by the search API were unavailable during the experiment



ASSUMPTION: No impacts to search outcomes.

- *Incorrect - Not enough information*

DEFINITION: Information on web page gives mixed signals (both *correct* and *incorrect* information) or no signal (e.g. not enough information, entirely irrelevant information).

ASSUMPTION: User visiting a web page with this label increases the likelihood to choose the incorrect answer.

- *Incorrect - Wrong Information*

DEFINITION: web pages containing the opposite information to Cochrane findings were given this label.

ASSUMPTION: User visiting a web page with this label has greatest increase to likelihood of producing an *incorrect* search outcome as well as *harmful* search outcome.

In line with the terminology used for participant decisions, web page annotations were mapped to the same terminology (*Correct*, *Incorrect* and *Harmful*) for analysis pertaining to the research questions. Any result annotated within the 3 *Inconclusive* sub-categories were classed as *Incorrect*. *Harmful* was assigned to results annotated as *Incorrect - Wrong Information*. Any result annotated as *Correct* remained categorized as *Correct*.

### 4.3.3 Test Set Metadata

In addition to the privacy and misinformation annotations, additional metadata was necessary for all in-lab experiments. As the in-lab experiments were task based and made use of a SERP, metadata collected was based on the SERP requirements

## 4. METHODOLOGY

---

**Is the information on the website 'Correct' or 'Incorrect'?**  
Select the best option (4 options are available in dropdown menu within each cell).

It is 'correct' if it agrees with `cochrane_answer` for the corresponding `cochrane_question`. 3 Incorrect options are available if information not correct.

**example for T6:** 'Correct' if website gives information indicating that Caffeine is helpful for Asthma, 'Incorrect' otherwise

**example for T3:** 'Correct' if website gives information indicating that insoles does not help back pain, 'Incorrect' otherwise

**USEFUL TIPS & ADDITIONAL INSTRUCTIONS:**

- Do not navigate elsewhere within the site, only base it on information in link provided
- Website is 'Correct' if there is information that agrees with `cochrane_answer`
- For sites unavailable or sites that display an error, Select 'Incorrect - Page Unavailable'
- Some results will take you to pages that do not have enough information to determine if it is incorrect/correct, they may contain information that is equally correct and incorrect or the page will not have enough information for you to decide, thus it is 'Incorrect - Not enough information'
- If the information on the website is contrary to the `cochrane_answer`, select 'Incorrect - Wrong Information'
- Use 2 monitors if you have access, it will go faster
- If the hyperlink does not open a webpage, you should paste in the link from `url_raw` column

**Figure 4.6: Annotation Instructions as Presented to the Annotators for Classifying Web Pages as *Correct* or *Incorrect*.** - The instructions also included three sub-categories for *incorrect* results.

to simulate the experience of commercial search engines such as Google. For each web page in the test set, metadata included the title, snippet, URL to actual web page, URL snippet for SERP and rank of the web page for the query used to retrieve the web page.

For the lab based online studies, web pages were loaded in the browser. However, for the offline studies images of the web page were loaded instead of HTML. Using the [FireShot](#)<sup>11</sup> screen grab plug-in images of the entire web page were collected for each URL in the test set.

### 4.3.4 Test Set Progression and Summary

Below is a listing of all test sets used in the empirical studies. The **Waterloo Test Set** and the **WhoTracks.me Test Set** were made available by other researchers, the other four test sets were original contributions as a result of empirical studies covered in this thesis.

---

<sup>11</sup>[FireShot](https://getfireshot.com/) at <https://getfireshot.com/> (LA: 2020-10-10)

The following details of the progression of each test sets and the metadata attached. In addition to the details of each test set listed below, final analyses in each interactive study utilized the interactive log data linked to metadata within each test set.

- I **Waterloo Test Set** = Published by Pogacar et al. and was the starting point for test sets used in current studies
  
- II **Offline Nudging Test Set** = **Waterloo Test Set** + Manually curated web pages
  
- III **WhoTracks.me Test Set** = 3rd Party Tracking Data for top 10,000 most visited Web domains with [Ghostery](#) browser plug-in
  
- IV **Online Nudging Test Set** = **Offline Nudging Test Set** + **WhoTracks.me Test Set** + Annotations of web pages visited
  
- V **Harm Prevention Features Test Set** = **Online Nudging Test Set** + Simplified annotations AND **WhoTracks.me Test Set**
  
- VI **Offline Boosting/Nudging Test Set** = **Online Nudging Test Set** (8 tasks only + only top 10 pages for most common query) AND Metadata + Images

The following details where each test set was used in the empirical studies.

## 4. METHODOLOGY

---

- **Waterloo Test Set** = Not used in any studies (but is a foundation for studies)
- **Offline Nudging Test Set** = Used in offline *nudge* studies Chapter 5
- **WhoTracks.me Test Set** = Used in online *nudge* studies Chapter 6
- **Online Nudging Test Set** = Used in online *nudge* studies Chapter 6
- **Harm Prevention Features Test Set** = Used to identify features for privacy and misinformation threats in Web search (Chapter 7) and pilot study study for the development of *boost* approach (Chapter 8)
- **Offline Boosting/Nudging Test Set** = Used in offline studies to compare *boost* and *nudge* strategies (Chapter 9)

### 4.4 Evaluation Measures

Any strategy to prevent harm (e.g. strategies [S1](#) - [S4](#)) should be evaluated with respect to the specific harm that is being addressed and measures should also take account for the possibility of side effects (both negative and positive) in other areas of Web search. In the subsections that follow, many of the evaluation metrics used for analyses related to [G-RQ-1](#) - [G-RQ-3](#) in the subsequent empirical studies are introduced. As evaluation is a core component ([FC-Evaluation](#)) of the framework we have introduced, everything introduced here falls within this concept.

In the empirical chapters that follow the current methods chapter, the measures introduced here are used at some point, but not necessarily in all studies. Details of their usage, along with study specific evaluation metrics, are provided within each study.

### 4.4.1 Harm Prevention (Search Task Outcome)

#### 4.4.1.1 Privacy Impacts

The privacy metrics listed below, calculated from the log data of participants and associated *3rd party tracker* data, can be viewed as the most fundamental metric linked to the overall research aim: to understand which (if any) behavioural and cognitive strategies are promising avenues for harm prevention in everyday Web search. The measures for privacy below are novel *search outcome* metrics for evaluation of IR systems designed to reduce privacy harms, they fall within **FC-Evaluation** of the framework and are essential for assessing **G-RQ-2**, but are important for **G-RQ-2** and **G-RQ-3** as well.

**Absolute Number of Trackers** Defined as the sum total of 3rd party trackers for web pages assessed by the user.

**Mean (Average) Number of Trackers** - Defined as the mean (average) number of trackers for all web pages assessed by the user.

**Maximum Number of Trackers** - Defined as the maximum number of trackers across all web pages.

**Normalized Number of Trackers** - Defined as the number of trackers for a web page normalized by the maximum number trackers.

## 4. METHODOLOGY

---

One final note about these metrics, if a user did not click on a link for a particular search task, then 0 trackers were recorded for that task.

### 4.4.1.2 Search Task Decisions

Any of the proposed harm prevention strategies have the potential to negatively impact the search task outcome, which is a particularly concerning scenario in certain search domains such as the medical domain considered in our studies. Put another way, the strategies proposed for evaluation may prevent harms due to loss of privacy, but as a side effect of the system designed with such a strategy the participant makes a poor decision or action as a result of the information they encountered. The metrics listed below specifically consider search task outcomes defined in Section 4.2.1 and are most crucial to analyses related [G-RQ-1](#), [G-RQ-2](#) and [G-RQ-3](#).

**Correct Decisions** - Number of *correct* decisions (out of total decisions).

**Incorrect Decisions** - Number of *incorrect* decisions (out of total decisions).

**Harmful Decisions** - Number of *harmful* decisions (out of total decisions).

Though the goal of these metrics is to ensure that poorer decisions were not made as a result of any harm prevention strategy, it is entirely possible that better decisions might also be a side effect of the strategy. For instance, it is plausible that certain Web sites not only have lower costs to personal privacy, but simultaneously have higher quality information with respect to the search tasks.

### 4.4.2 Compliance to Transparent Strategies

Strategies [S3](#) and [S4](#) are both transparent strategies, that is some indication of harm is communicated to the user. With such a strategy, they can choose to comply

with the strategy or not. It is plausible that some users will disregard the strategy, for instance they may still visit results with a Red warning light even though they are told it is more risky than a result with a Green light. Cross comparison of transparent and non-transparent systems are addressed with these metrics and relevant to **G-RQ-2**, **G-RQ-2** and **G-RQ-3**. For example, we assign a hidden html tag with the Stoplight colouration to all results across all search systems introduced in Section 4.2.2, which allows for development of such metrics (covered in later studies) for cross comparisons of systems.

As presented in Chapter 2, there is also evidence of a *Privacy Paradox*, where individuals who express great concern about harms due to loss of privacy in Web search do not make use of privacy protection in the instance when it is offered to them. Such metrics (along with self-report measures in Section 4.4.4.4) are therefore important to better understand this paradox.

### 4.4.3 Search System

Even if a strategy implemented by a search system successfully reduces or prevents the harm for which it was designed, the intervention strategy should ideally will have limited (if any) degradation to the user experience. Nor should the strategy cause an increase to their exposure to web pages that are more risky (e.g. with *incorrect* or *harmful* information with respect to the search task at hand). Combined together, this plays into **G-RQ-1** which looks at the overall viability of an intervention strategy. We now introduce metrics that will be useful to perform analysis with respect to **G-RQ-1** and allow for comparison of the proposed search system and their related strategies.

## 4. METHODOLOGY

---

### 4.4.3.1 Search System (Search Behaviour)

Adoption of commonly used measures for conducting interactive IR experiments [120] were fundamental to the analysis related **G-RQ-1** of strategies. The ideal strategy will have a positive impact on these measures for interactive behaviour, and at the very least have no significant negative impacts, and therefore are deemed important for answering this question.

**Assessments** - Total number of results / web pages assessed by a participant.

**Queries** - Total number of unique queries submitted by a participant.

**Queries w/o SERP clicks** - Total number of unique queries where participants did not click on SERP result.

**Time to Completion** - The amount of time (in seconds) a participant spent completing a search task. Total Time =  $T_1 - T_0$ . Where  $T_0$  = Time when user landed on SERP for new search task and  $T_1$  = Time when user hit decision button to leave SERP.

**Rank** - Average rank of clicks that a user made within SERP.

### 4.4.3.2 Search System (Adapting IR Metrics)

Traditional studies make use of the following metrics which were introduced in the evaluation section, it is therefore useful to make use of such metrics where possible.

The nature of the online study comparing all three *nudge* strategies provided a dataset that allowed for comparison across systems of the quality of information one might encounter. For these comparisons, we modify several traditional IR system



metrics appropriate to this evaluations. Metrics include precision, recall and mean reciprocal rank (MRR). The adaptation of these metrics is with respect to documents that are *correct* or *harmful*, instead of using the traditional approach of a document that is “relevant”. More specifics are provided in Chapter 6 with respect to the usage of these metrics.

In the final empirical chapter investigating **G-RQ-3** we adapt a popular gain based metric, normalized Discounted Cumulative Gain (nDCG), as one approach to compare the *nudge* and *boost* strategies. In this study, harms and benefits of different results with respect to privacy are considered. Using nDCG , the metric is modified for analysis specific to harms and benefits of each system. As with the online studies, more details are provided with respect to these metrics 9.

### 4.4.4 Demographics & Self Report Measures

Pre-task and post-task self report measures were gathered during each search task. Upon completion of the main experiment, all participants completed an in-depth questionnaire after the main study to capture demographic information, of which full questionnaires not included in this section are found in Appendix A.1.3. The evaluation metrics introduced here fall within both the contextual and usability dimensions [119], and are seen as quite important for demonstration of system viability (**G-RQ-1**) across different users perspectives.

#### 4.4.4.1 Demographics

A set of basic demographics were collected in all user based studies, including participant sex, age, education level, and fluency of language (e.g. native vs fluent English speaker). This information was used to rule out problems with the sample (e.g. males behave differently than females). Demographic values are reported for

## 4. METHODOLOGY

---

participants samples across each study.

### 4.4.4.2 Perceptions and Preferences

For the studies investigating the *nudge* strategies, users were asked at in the debriefing segment of the post-experiment survey (to ensure no biases about other questions) about their most and least preferred strategy. These responses results in the following metric.

**Preferred Strategy** - Total count of participants selecting strategy  $X$  as their most / least preferred intervention.

### 4.4.4.3 Pre-task Knowledge and Post-task Confidence

The following pre and post task metrics were captured. However, analysis was only performed on the post-task metrics as the pre-task metrics did not fit with the current high level research questions. The appendix (Section [A.1.1](#)) includes a full description of the questions and how they are presented to participants.

- **Search Task Knowledge** - The average score of the two pre-task questions (Appendix Table [4.4.4.3](#)) to capture user knowledge with respect to the search task (questions adopted from [[179](#)]).
- **Search Task Confidence** - The average score of the two post-task questions (Appendix Table [4.4.4.3](#)) to capture user confidence in the search task outcome (questions adopted from [[179](#)]).

### 4.4.4.4 Actions and Beliefs Towards Harm Prevention Goal (Privacy)

Measures of behaviours to protect privacy and awareness of privacy tools were captured in the post-experiment demographics survey in addition to questions to

capture general attitudes about privacy. The following metrics were calculated.

- **Privacy Attitudes (General)** - The average score for each participant was taken from set of Likert based scale questions to capture attitudes towards privacy. Full details of the questions are found in Appendix [A.2](#).
- **Privacy Attitudes (Health)** - The average score for each participant was taken from set of Likert based scale questions to capture attitudes towards privacy of health information. Full details of the questions are found in Appendix [A.3](#).
- **Action Score (Binary)** - Assigning 1 point to each question response (assigning 1 point to any response between and including "Sometimes" & "Always" and 0 for anything else) in Table [A.4](#), the sum across all 14 questions was taken (maximum possible was 14 points).
- **Action Score (Total)** - Using the raw score of each question response ("Never" was given 0 points) in Table [A.4](#) the sum across all 14 questions was taken (maximum possible was 70 points).
- **Awareness Score** - Assigning 1 point to each question response (assigning 1 point to any response between and including "Never" & "Always") in Table [A.4](#). The sum across all questions was taken (maximum possible was 14 points).
- **Enhancing Browser Score** - The average score was taken for frequency of usage of privacy enhancing browsers (maximum 3). The question asked and options available in Appendix Table [A.5](#).
- **Enhancing Search Engine Score** - The average score was taken for frequency of usage of privacy enhancing search engines (maximum 2). The question asked and options available in Appendix Table [A.6](#).

### 4.4.4.5 Measures Outside of Scope

Economic measures of search utility (time, speed, privacy) were considered, and are certainly an important factor [11, 40]. However, due to concerns around participant fatigue we excluded such measures from our study.

For similar reasons, we did not include a recently developed survey [71] which captures aversion to uptake of new technology (e.g. new technologies such as the transparent *nudge* [S3] and *boost* strategy [S4] search systems).

There are many more things to consider, however they go beyond what we can cover in this thesis.

## 4.5 Statistical Tests

A range of statistical tests were used in the analyses of research questions, and all analyses were performed in the R programming language [182]. Choice of test was predominantly driven by the nature of the dependent variable (e.g. continuous, binary), study design (e.g. between vs. within) and data collected (e.g. repeated measures). Unless otherwise specified we use  $\alpha = .05$  as a level of significance. Where appropriate, Cohen's  $d$  is calculated as a measure of effect size.

Given the nature of the data collected in the studies (e.g. repeated measures of users and tasks), Linear mixed effects regression (LMER) and Analysis of Variance (ANOVA) were appropriate tests for many of the hypotheses.

Analyses using independent t-tests (for between-group comparisons) and paired t-tests (for within-group comparisons) were commonly used as well for comparisons of dependent evaluation metrics across the different systems (e.g. comparing Control vs. Stoplight). Two-sided t-tests were chosen over one-sided t-tests in the analyses

as the directionality of the chosen metrics in our studies was not guaranteed. This choice of the two-sided analysis decreases the chance of rejecting the null hypothesis, but does provide the advantage of detecting unexpected behaviour important for future research.

Logistic regression with log odds ratios were used to test for differences in binary responses, such as the studies run to identify useful features for privacy protection (in Chapter 7).

Chi-squared tests were used for analyses across several factors, such as tests comparing user perceptions of the interventions.

Two reliability measures were employed in the empirical studies. Cohen's  $\kappa$ , a measure of inter-rater reliability, was used to measure the quality of the misinformation annotations (introduced in Section 4.3.2). Chronbach's  $\alpha$  was used to test reliability of the self-report measures introduced in our studies.

## 4.6 Participants and Recruitment

Convenience sampling, recruiting participants based upon convenience of being located nearby the experimenters (a common practice for many IIR studies [119]), was used for recruitment of subjects for all in lab experiments. Three different channels were used for this process. The first channel was via a list maintained by the Psychology department; a list containing a pool of 4000 to 5000 students, university employees and individuals from the surrounding community<sup>12</sup>. Another channel was via recruitment messages sent to a university advertising list server and department messages sent to students in multiple departments including Computer

---

<sup>12</sup>This list was predominantly composed of students, and students are given the option to join this list during registration day at the University. Students are removed from this list after completion (3-4 years). As such, the list was in constant flux.

## 4. METHODOLOGY

---

Science, Law and Sociology. Subjects from the preceding groups were paid £10 for their participation. The remaining participants were recruited via a university website giving course credit in exchange for experiment participation. A master list was maintained for all experiments to ensure that no participant returned for any other study. Demographic breakdowns of participant samples is provided within each empirical study.

For the online and offline *nudge* experiments, recruitment messages included no mention of the privacy aspect of the study. They were told about the privacy aspect only at the end of the post-experiment questionnaire in the experiment debriefing (see Section 4.2.3). However, for the studies that included *boost* strategy [S4], a statement about the privacy aspect was included in the recruitment message. This choice was made intentionally, for several reasons. First, *nudge* strategies are opt-out whereas *boost* strategies are opt-in, and therefore motivation for recruitment in this manner as it simulates the opt-out and opt-in aspect. Second, *boost* strategies require motivation from the individual to learn the skill and we hoped that recruitment of participants in this manner would increase motivation, for which this point is especially important with respect to G-RQ-3. One final note is that all participants in the final study that compares *boost* and *nudge* strategies and addresses G-RQ-3, all participants received the same recruitment message and therefore aware that the study was about privacy protection.

### 4.7 Ethics

Prior to the data collection phase of the experiment, methods went through an ethics review process. Upon sign up for the experiment, participants were provided an information sheet outlining potential risks and how their data would be stored and analysed. Subjects were informed they could exit the experiment at any point;

nonetheless, all participants completed the experiment. Before the start of the experiment, all subjects signed a consent form regarding collection of their data and how long different types of data would be stored (e.g. personally identifiable information would be deleted after all experiments related to the research grant were completed). As the Microsoft Azure Bing Search API (see technical details in Appendix A.3) and Qualtrics survey software were resources used in some of the studies; subjects were informed at the start of the experiment about transmission of their queries and survey responses to these cloud services.

## 4.8 Summary

In this chapter, the general methodology and foundation has been provided for comparison of four behavioural and cognitive strategies (S1 - S4). Many of the methods that cut across the empirical chapters that follow are provided here. Within each specific empirical chapter, deviations from these methods will be included as well as methods that were not appropriate for the general methodology.

An important point about the studies that follow is that they were highly experimental in nature, that is, no one to our knowledge has attempted to compare multiple *nudge* and *boost* strategies in an IIR setting. As such, it was necessary to perform the studies as stages, where the general methods were guided and developed by learnings from each study in sequence. A listing of some of the high level differences is provided below, where the predominant focus is on evaluation (FC-Evaluation) of novel Web search systems (FC-System) incorporating behavioural and cognitive strategies (FC-Cognitive) for harm prevention.

- Chapter 5 is the starting point for development of the general methods. Variations will be presented, such as a more limited set of metrics used in the

## 4. METHODOLOGY

---

analyses of **G-RQ-1** and **G-RQ-2**.

- Chapter 6 uses almost all elements presented in the general methodology. As this study was performed with a live search engine, it had a much richer dataset available for analysis and therefore include the most extensive set of analyses in all empirical chapters.
- Chapter 7 is an important transition study aimed to identify features useful for the remaining two empirical chapters. Most of the methods in this study are quite different from the general methodology, and details specific to this study are found there. Also, the research questions for this study were aimed at identifying features linked to 3rd party tracking and misinformation, these questions and hypotheses are presented within the chapter itself. As already indicated in 4.3, it makes use of data collected and test sets in the online study. From the perspective of the framework, this study demonstrates methods one might use for the development of new informational cues (**FC-System**); cues which might be used to develop new cognitive strategies (**FC-Cognitive**).
- Chapter 8 uses the findings from Chapter 7 for the piloting of a *boost* strategy (a fact box about privacy). Three different variants of the same fact box are evaluated in a user study. However, the study was performed on a popular Crowdsourcing platform and not based in a lab (a key difference from the other 3 user studies).
- Chapter 9 makes use of the findings from Chapter 8 to compare the *boost* and *nudge* strategies. The key difference in this study being that **G-RQ-3** is addressed at this stage and thus requires some variation from the general methodology, which were briefly touched upon in Section 4.2.3.

An overview of the studies is provided in Figure 4.7 to aid the reader in key



differences across each study.

Chapter	Empirical Studies				
	5	6	7	8	9
<b>Brief Summary of Investigation</b>	Nudging in an Offline Search Setting	Nudging in an Online Search Setting	Identifying Useful Environmental Cues for Harm Prevention	Pilot Study to Test Transfer of Skill with Various Boosts	Comparing Boosting and Nudging in an Offline Search Setting
<b>Strategies Tested</b>	S1 - S3	S1 - S3	Features in URL	S4	S2 and S4
<b>Behavioural / Cognitive Approach</b>	Nudge	Nudge	N/A	Boost	Nudge and Boost
<b>High Level Research Question</b>	HL-RQ 1 and 2	HL-RQ 1 and 2	N/A	HL-RQ 1 and 2	HL-RQ 1, 2 and 3
<b>Interactive Study Design</b>	Within Group Graeco-Latin Squares	Within Group Graeco-Latin Squares	N/A	Between Group with Randomization	Between Group with Balanced Helpful/Not-helpful tasks
<b>Cochrane Search Tasks</b>	All (5 helpful/5 not helpful)	All (5 helpful/5 not helpful)	All (5 helpful/5 not helpful)	2 of 10 (1 helpful/1 not helpful)	8 of 10 (4 helpful/4 not helpful)
<b>Search Task Corpus</b>	Static (20-22 Documents per Task)	Dynamic - Full Web - Retrieved via Commercial Search API	N/A	N/A	Static (10 Documents per Task)
<b>Key Dependent Measures of Harm</b>	Privacy - 3rd Party Trackers and Associated Stoplight Colours. Medical Search Task Decisions.	Privacy - 3rd Party Trackers and Associated Stoplight Colours. Medical Search Task Decisions. Information Quality Measures.	Privacy (3rd Party Trackers) & Information Quality	Multiple Choice Answer for Least Privacy Impactful. Estimation Scores of Benefits/Harms for TLD.	3rd Party Trackers and Usage of TLDs
<b>Key Independent Variables</b>	Search Systems (incl. Control) and Self Report Measures Related to Privacy	Search Systems (incl. Control) and Self Report Measures Related to Privacy	URL Features - TLD and HTTP/HTTPS	Control + Factbox - Variations of Presentation (Large, Small, Inline Online)	Search Systems - Re-ranking Nudge, Large Factbox before, Inline Factbox Only + Control
<b>Participants</b>	Recruited from University (N=91)	Recruited from University (N=90)	N/A	Recruited from Prolific Platform (N=212)	Recruited from University (N=70)

**Figure 4.7: Overview of All Empirical Studies.** - Five empirical studies were completed in the thesis. This figure includes an overview to compare differences of each study at a high level. N/A denotes not applicable.

The five empirical chapters are now presented in the order they were undertaken to test and compare strategies [S1](#) - [S4](#).

## 4. METHODOLOGY

---

# Chapter 5

## Investigating Nudges in an Offline Setting

### 5.1 Overview

Three *nudge* strategies ([S1](#)-[S3](#)) introduced in the previous chapter, strategies motivated through [FC-Cognitive](#) of the framework, were investigated in a controlled offline search environment in a lab setting. In the study presented here, participants completed 10 search tasks (Table 4.1) and encountered 3 experimental search systems which mapped to the 3 *nudge* strategies designed to reduce privacy impacts to the individual.

Recall the distinction (in background Section 2.6) between non-transparent classic *nudges* and transparent ‘educational’ *nudges*. Furthermore, it was stated that transparent strategies are more ethically sound than non-transparent strategies. Therefore, from an ethical standpoint, it is desirable that with respect to [G-RQ-1](#) and [G-RQ-2](#) that the Stoplight *nudge* (strategy [S3](#)) will be the most viable and effec-

## 5. INVESTIGATING NUDGES IN AN OFFLINE SETTING

---

tive of the three strategies considered in this study.

As already presented in the background section on biases and behaviours (Section 2.4.5), one should not expect the most ethical strategy under current consideration to perform the best. Presumably, one would expect users to strongly adhere to the transparent strategy if in fact they care about the harm it aims to prevent (i.e. loss of privacy). However, evidence presented in the background also suggests this may not be the case (e.g. the “privacy paradox”). Nonetheless, research in behavioural and cognitive interventions (again in the background) suggests that individuals motivated to prevent the harm will make use of such interventions.

This in turn leads us to the formulation of several hypotheses.

With respect to the high-level research questions **G-RQ-1** and **G-RQ-2** we formulate the following two hypotheses.

**H1a** all 3 *nudge* strategies to significantly outperform the control search environment with respect to the harm being tested (privacy impact).

**H1b** Both non-transparent *nudge* strategies are expected to outperform the transparent *nudge* strategy with respect to privacy impacts.

Specific to the transparent Stoplight *nudge* strategy, we again investigate the high level viability and effectiveness questions (**G-RQ-1** and **G-RQ-2**). However, for this specific strategy the analyses focused on differences in user attitudes and behaviours with respect to privacy (as indications for motivation), and therefore evaluated based on the following hypotheses.

**H2a** For the transparent *nudge* strategy, the “privacy paradox” will not be present.

That is, participants with strong attitudes about privacy will have privacy

impacts significantly reduced more so than participants with weaker attitudes.

**H2b** Similarly, we expect individuals taking privacy protective behaviours in everyday life to make use of the Stoplight *nudge* strategy more so than individuals that take little or no privacy protective action.

In addition to testing the above hypotheses, another aim of studies presented in the current chapter was the refinement of general methods for their use in later studies (e.g. the online *nudge* study in Chapter 6), along with development of new *search outcome* based evaluation metrics (as part of [FC-Evaluation](#)) to measure harm across systems. As no one, to our knowledge, has run a study to compare multiple *nudges* (or multiple harm prevention strategies) in a Web search environment, these studies can also be viewed as pilot studies for the studies covered later. This is one reason why the methods used in the current studies are more limited than those used in later studies. Methods used in the current studies also differ from later studies due to the skills of the thesis author at time of these studies; skills which were somewhat restricted (e.g. knowledge of statistical methods, no previous experience with lab studies). Several important differences from the general methods will be highlighted below.

## 5.2 Method

The current studies acted as precursor studies to inform overall general methods (Chapter 4). As such, the methods used in the current studies are a scaled back form of the general methods. In the sections below, exceptions and deviations from the general methods are provided, such as survey questions provided to the users and evaluation metrics used in analyses of research questions.

An offline study was performed in a lab setting. As described in the general

methods, participants ( $n = 91$  in total) were recruited and were presented a set of search tasks and *nudge* strategies [S1]- [S3] using a *within-group* design for testing hypotheses [H1a] and [H1b]. A *between-group* group design was used in addition to the *within-group* to test differences specific to [H2a] and [H2b], which focuses on strategy [S3] (the transparent Stoplight intervention).

### 5.2.1 Procedure

A key difference in procedures employed for the current studies relate to the SERP used to test the strategies.

**Search Engine Results Page (SERP)** In the current offline studies, which informed modifications to SERPs in later studies, users could not submit queries and therefore no query bar was available in the SERPs for the present studies. This variation, along with several other notable differences, is visible when comparing the Stoplight SERP (Figure 5.1) used in the current offline *nudge* studies with the Stoplight SERP in the general methods (Figure 4.2 used in the online *nudge* studies).

Recalling that 3rd party tracking data was associated with all web pages in the **Offline Nudging Test Set** used in this study, the Gray lights were not include in the Stoplight SERP used in the offline *nudge* study. A tooltip which displayed the definitions of Stoplights (with mild variations in language) upon mouse over by the user (instead of a sticky panel in the online SERP) was another difference. Additionally, in the pre-testing stages of the offline system, two participants reported that it was unclear they could click on the results, therefore a small snippet was added that said “You can click on links below”.

**MEDICAL QUESTION 2:** Does traction help low back pain?

**HEALTH ISSUE:** **back pain** - The spine is a column of bones (vertebrae) held together by muscles, tendons and ligaments and cushioned by shock-absorbing disks. A problem in any part of your spine can cause back pain. Source: *Mayo Clinic*

**TREATMENT:** **traction** - a pulling force exerted on a skeletal structure (as in a fracture) by means of a special device. Source: *Merriam-Webster*

You can click on links below

24 results returned.

**Red Stoplight:** **Lumbar Traction Offers No Benefit for Back Pain**  
<https://www.verywellhealth.com/>  
 Does Traction Really Work for Low Back Pain? A study confirms that using lumbar traction with exercise for low back pain does not offer improved outcomes when compared to physical therapy exercises alone.

**Yellow Stoplight:** **Does Spinal Decompression Really Work in Treating Low Back Pain?**  
<https://www.verywellhealth.com/>  
 A Questionable Treatment Spinal decompression may help treat low back pain, but this popular treatment isn't a sure thing. Advertising for spinal decompression targets people with degenerative disc disease, bulging discs, herniated discs, or spinal stenosis.

**Green Stoplight:** **Traction Therapy for Chronic Low Back Pain**  
<http://www.barcloayphysicaltherapy.com/>  
 The cost of health care is rising every year in the United States. And part of that economic burden is the management of chronic low back pain (CLBP).

**Yellow Stoplight:** **Traction for Low Back Pain With or Without Sciatica: An Updated Systematic Review Within the Framework of the Cochrane Collaboration**  
<https://www.researchgate.net/>  
 Systematic review. To determine if traction is more effective than reference treatments, nlarzho/cham traction, or no treatment for low back pain (LRP).

**helpful:** The medical treatment helps if the treatment is effective and has a direct positive influence on the specified illness.  
**inconclusive:** The effectiveness of a medical treatment is inconclusive if medical professionals are still unsure if the treatment will have a positive, negative or no influence on the specified illness.  
**does not help:** The medical treatment does not help if the treatment is ineffective and either has no effect or has a direct negative influence on the specified illness.

**Privacy risk where:**  
 Red = High privacy risk  
 Yellow = Medium privacy risk  
 Green = Low privacy risk

Figure 5.1: SERP with Stoplight (best viewed in colour) Strategy Used in Current (offline) *Nudge* Study. - A mouseover tooltip was provided with definitions of the lights (as opposed to sticky panel in online study). No query bar is available.

## 5.2.2 Evaluation Test Sets

As stated in the general methods (Chapter 4), the starting point for the **Waterloo Test Set** was used the starting point for development of the **Offline Nudging Test Set** used current study. There are some important points worth highlighting in relation to the **Offline Nudging Test Set** we developed.

In the **Waterloo Test Set** 55 documents were no longer available. These documents were excluded from the study as 3rd party tracking data for would not be retrievable. This reduction in documents produced some challenges. For example, only 8 of the 16 documents associated with task 9 (Table 4.1) remained for our research. Given the desire to have at least 10 results visible for all *nudge* interventions, this challenge was particularly noticeable in regards to the filtering intervention, where for instance only 4 documents would appear to the user for task 9 (recall filtering removes the upper median of trackers). Furthermore, for a consistent test set across tasks, it was desirable to have baseline results closely match the

## 5. INVESTIGATING NUDGES IN AN OFFLINE SETTING

---

findings that approximately 80% of search results for the Cochrane medical search tasks contain *correct* information [246]. These challenges were more than sufficient motivation for expansion of the test set.

For the expansion of the corpus, we used Bing, Yahoo and Google, in line with the authors of **Waterloo Test Set** (see [179]). We also used a VPN set to the Toronto area of Canada (location near original authors) and ran searches using incognito browsing to minimize any potential of personalization of results. We contacted the authors to determine search terms they used and followed their instructions. For each search task (Table 4.1) the instructions were to query “Does search task treatment help search task medical issue?” We recorded the rank of each *correct* result found in the top 30 results. Due to time constraints, anything beyond rank 30 was recorded as 31, as were *incorrect* results. These rankings were assigned to all documents in the original and expanded corpus, we then created a weighted ranking based on the market share of each of the 3 search engines in July 2018<sup>1</sup>.

Additionally, inconsistencies were discovered in the snippets provided in the **Waterloo Test Set**. Though reported that all snippets were length of 2 sentences, many snippets were longer and for Task 9 and Task 10 no snippets were available. To address all issues, we used the first two sentences of the first paragraph for web pages linked to Task 9 and Task 10. For all other tasks, we ran the Python NLTK sentence splitter over the snippets provided to ensure only 2 sentences were visible for each result.

Using the annotation methods for privacy and misinformation already outlined in general methods, all web pages were annotated for privacy using the manual approach. For misinformation annotations, the author of this thesis and his PhD

---

<sup>1</sup>Search engine market share based on desktop market share in Canada. See [StatCounter](http://gs.statcounter.com/search-engine-market-share/) at <http://gs.statcounter.com/search-engine-market-share/> (LA: 2020-08-15).



---

supervisor independently annotated all web pages added to the corpus and then resolved disagreements (only 4) through joint discussions. However, misinformation annotations were only performed at for being *correct* or *incorrect* as related to the search task.

Our updated corpus had 19-24 documents for the 10 search tasks (Table 4.1) and we did not add documents for the practice tasks. Altogether, 95 documents were added (93 correct and 2 incorrect) to the test set, for a grand total of 296 annotated web pages.

**Ranking Search Results** In the process of retrieving documents from the three search engines, we noticed that slight variations occurred in the ranking of documents (e.g. document at rank 3 moves to rank 4 after returning to SERP from landing page on a result visited). Such behaviour occurred in all search engines and was most notable in Bing. We felt this somewhat random behaviour was motivation to add some randomness to our offline environment as it would more closely mimic online environments.

To produce this randomness during the experiment we take the weighted ranking as input to an inverse cumulative distribution function (CDF) to formulate a final rank for the documents related to each task. Use of the inverse CDF in this manner places a much higher probability of a web page found at Google rank 3 to appear for a participant than a web page found at Google rank 27. The inverse CDF was calculated using the Scipy library [238] available in Python. Furthermore, a seed setting was assigned to each participant, so that the randomness only occurred across participants (i.e. the ranking would not change within a specific search task for a specific user). We used the findings that 80.69% of results for medical queries are *correct* in the top 10 results [246] as a guideline to tune the ranking of the results.

In line with result rankings outlined in the general methods, the rankings from the inverse CDF calculation were maintained for the control and two of the strategies (S1 and S3). For Re-ranking strategy S2, all results were re-ranked based upon the number of *3rd party trackers* (least to greatest).

### 5.2.3 Evaluation Metrics

A subset of the metrics introduced in the general methods (Section 4.4) were used as dependent variables for testing the hypotheses in the current studies.

#### 5.2.3.1 Measures Used to Compare *Nudge* Strategies (H1a and H1b)

As the primary goal of the *nudge* strategies was a reduction in privacy impact, the main *search outcome* metric considered is the Mean Number of Trackers. Another important aspect of *search outcome* is the actual decisions made during the search task, for which Correct Decisions and Harmful Decisions were used.

To test negative behavioural differences to the user as a result of the strategy, we use the Time to Completion metric in our analysis.

Critical to G-RQ-1, at the end of the post experiment questionnaire, all participants were asked about their preferred *nudge* strategy. A simple analysis is performed using the Preferred Strategy metric.

#### 5.2.3.2 Measures for (H2a and H2b)

**Harm Prevention (Search Outcome Measures)** For the analyses of the *search outcome* of privacy impact as it relates to specific user attitudes and behaviors (H2a and H2b), the Mean Number of Trackers was not a fair comparison as some tasks had different overall privacy risks associated with them. For instance, one task would

have a lower median number of trackers across all web pages associated with that particular task compared to another task. The thresholds for Stoplight colouration (see Section 4.2.2.2) were based upon the median and upper quartile values. This was problematic, given the *within-group* design used for data collection, in that participant X would see task 1 and 2 with the Stoplight *nudge*, while user Y would see task 4 and 9 with Stoplight *nudge* (recall that 2 tasks were assigned to each strategy). For example, in the instance that both participant X and Y took privacy protective action in the Stoplight SERP (e.g. only visit results with Green lights), the **Mean Number of Trackers** could still be different. As such, for the analysis, the **Normalized Number of Trackers** as described in the general methods was used as a dependent variable for the analyses.

**Self Report Measures** Self report measures were a critical component for testing of the hypotheses related to interactions with the Stoplight *nudge*. Though many different questionnaires are available with respect to privacy, no single questionnaire was fit for purpose for our research questions. As such, using the survey in [132] as a starting point, a new questionnaire was developed. As part of the process, methods provided by [70] and suggestions by authors of [107] were utilized for during the development of the questionnaire.

The following five metrics from the seven introduced in the general methods (Section 4.4.4.4) were used in the analyses.

- **Privacy Attitudes (General)** abbreviated as **Attd-General** in current study
- **Privacy Attitudes (Health)** abbreviated as **Attd-Health** in current study
- **Action Score (Total)** abbreviated as **Bhv-General** in current study
- **Enhancing Browser Score** abbreviated as **Bhv-Browser** in current study

## 5. INVESTIGATING NUDGES IN AN OFFLINE SETTING

---

- **Enhancing Search Engine Score** abbreviated as **Bhv-Search** in current study

One important note regarding **Bhv-General**. The current study used items 1 - 10 in Table A.4 and a simple summation was used for the options selected by the user (minimum score of 0 and maximum score of 10 possible).

### 5.2.4 Statistical Tests

Statistical tests used for analyses in the current studies are divergent from statistical tests outlined in the overall methodology (Section 4.5).

For analyses related to **H1a** and **H1b**, logistic regression was used for analysis of **Search Task Outcomes** variables (*correct* and *harmful*). As with lab based studies covered in later the chapters, all models were controlled for repeated entries of participants and search tasks, however analysis of variance (ANOVA) was used for privacy and time impacts instead of GLM and LMER approaches. The choice of ANOVA was due to the overall progressive nature of the thesis and the author being unaware of the newer methods at the time of the study.

For analyses related **H2a** and **H2b**, linear regression and t-tests were utilized as introduced in Section 4.5.

### 5.2.5 Participants

A total of  $N = 91$  subjects participated in the offline *nudge* study. Of the sample ( $N = 91$ ), 89 were students, of which 77 of those were undergraduates. The average age of participants was 23.2 years and 63 identified as Female. English was reported as the native language for 27 students. STEM departments provided 53 participants, with 17 participants having a background in computer science. 42

participants reported using a privacy-protective browser at least once a month, while only 10 reported usage of a privacy-enhancing search engine.

A technical error with the Graeco-Latin Square design resulted in  $n = 51$  participants receiving imbalanced exposure to the Re-ranking and Filtering SERPs. To address the data imbalances, a subset of  $n = 40$  participants was used in the analysis for **H1a** and **H1b**. The study sample included 39 students (all undergraduates) and 1 non-student participated, of which 29 were female and 11 were male. The average reported age was 21.3 years. The average time for experiment completion was 40 minutes. 14 reported English as the primary language at home, noting the majority of students were non-native English speakers.

The full participant sample ( $N = 91$ ) was used for analyses related **H2a** and **H2b**.

## 5.3 Results

### 5.3.1 Comparing 3 *Nudge* Strategies

To test the hypotheses, the effects of the harm prevention systems on the four dependent variables (Mean Number of Trackers, Correct Decisions, Harmful Decisions and Time to Completion) were tested. A Control SERP and baseline search task were also included in the analyses. Preferences of the strategies were also considered (Preferred Strategy).

The participants interacted with a controlled SERP (either a baseline or privacy Nudge variant) or a control question with no search results available to assist with their decision.

## 5. INVESTIGATING NUDGES IN AN OFFLINE SETTING

---

With respect to the impacts of the 3 *nudge* strategies on reductions of encounters with privacy trackers, Table 5.1 shows significant effects for both the Re-ranking and Filtering approaches as compared with the baseline. However, there were no significant differences when comparing the Stoplight approach with baseline in our sample ( $n = 40$ ).

Independent Variables	Mean # 3rd Party Trackers Encountered	Lower 95%	Upper 95%
Control SERP	6.11	5.25	6.97
Filtering SERP	2.27	1.41	3.13
Ranking SERP	1.93	1.07	2.79
Stoplight SERP	6.64	5.78	7.50

**Table 5.1:** Confidence intervals were calculated with one-way ANOVA. Interfaces have a significant effect on the average number of 3rd party trackers encountered during a search task,  $F(3, 316) = 32.30$ ,  $p < .0001$ . When comparing the individual privacy *nudge* strategies with the Control SERP, non-overlapping confidence intervals for the Ranking and Filtering approaches confirm  $p < .05$ .

Related to the *search outcome* metrics for user decisions, Table 5.2 provides the percentage of *correct* and *harmful* decisions for 40 participants. A direct comparison of the user decisions to [179] shows the control performs worse for our participants. Comparisons of the user decisions for the 3 *nudge* strategies against the Control SERP indicate that *harmful* decisions are reduced for all approaches, however the effects on decision making are not found to be significant for any *nudge* strategy compared to the Control.

Analyses related to the time taken for each task as a measure of possible cognitive impacts [120] were also performed with one-way Anova as a possible indicator for negative implications with the experimental strategies. The mean time in seconds to complete a search task for the 4 SERPs with the sample (Control = 117 sec., Filtering = 99 sec., Re-ranking = 99 sec. and Stoplight = 108 sec.) and any of the

Independent Variables	Correct	Harmful
Control SERP	49%	18%
Filtering SERP	50%	14%
Ranking SERP	49%	16%
Stoplight SERP	54%	15%
Baseline Task	29%	26%

**Table 5.2:** Impacts on decisions. While minor variations exist when comparing the decisions of Control SERP with decisions of the three privacy SERPs, none of the differences indicate statistical significance. For the Baseline Task (i.e. no search results available), task outcomes perform the worst overall (as expected), a finding in line with previous research (see [179]).

variation was found to be non-significant.

Data indicate not a single participant made use of the opt-out mechanism (a necessary element of *nudging*) introduced in the general methods. That is, no participant made use of the switch available in the hamburger icon (see Figure 5.1). In other words, the default privacy protection strategy was left on by all participants.

Finally, data was collected for the participant preferences across the three strategies. Table 5.3 provides the results of the most and least preferred strategies. From the data, no strategy stands out as most or least preferred.

Intervention	Most Preferred	Least Preferred
Filtering SERP	17	15
Ranking SERP	7	12
Stoplight SERP	16	13

**Table 5.3:** Raw counts of the user preferences ( $n = 40$ ) are included for the most and least preferred *nudge* strategies for harm prevention.

### 5.3.2 User Variations with a Transparent Strategy

Data collected from all participants ( $N = 91$ ) in the offline *nudge* study were used to test hypotheses [H2a](#) and [H2b](#).

First and foremost, when comparing interactive data across all participants ( $N = 91$ ) with the Stoplight *nudge* strategy and the interactions from the Control SERP, no significant differences were found with respect to the *search outcome* measures (Mean Number of Trackers, Correct Decisions and Harmful Decisions). These findings are in line with results on the smaller sub-sample used to test [H1a](#) and [H1b](#), which are contrary to expectations for this strategy.

Turning to results related to [H2a](#) and [H2b](#) and the five self-report metrics (Attd-General, Attd-Health, Bhv-General, Bhv-Browser and Bhv-Search). As the questions used to formulate the in the metrics were new, we tested the reliability of these questions. The two attitude groups demonstrated excellent internal consistency, with Cronbach's  $\alpha = .76$  for the 7 general attitudes questions and Cronbach's  $\alpha = .89$  for the health questions. We considered combining the all privacy behaviour questions as one measure, but given the unsatisfactory reliability of the scale (Cronbach's  $\alpha = .53$ ), we kept the 3 measures (Bhv-General, Bhv-Browser and Bhv-Search) separate.

Our findings related to the metrics used (Attd-General, Attd-Health and Bhv-Browser) are provided in Table 5.4. The two remaining metrics were not included in the table as no significant differences were found. The privacy attitude measures (Attd-General and Attd-Health), while reliable appear to have limited predictive validity related to the Stoplight strategy. No evidence was found to suggest it is a useful measure to predict how individuals will behave when presented with a Stoplight privacy protection strategy. However, there are significant links between



the two attitude measures and browser usage.

<b>Bhv-Browser</b>	<i>b</i>	<i>F</i>	<i>df</i>	<i>SE</i>	<i>p</i>
Attd-General	0.30	10.96	89	0.09	** .001
Attd-Health	0.14	8.98	89	0.05	** .004
<b>Normalized Number of Trackers</b>					
Bhv-Browser <sup>M1</sup>	-0.29	5.57	180	0.03	* .020
Bhv-Browser <sup>M2</sup>					
Bhv-Browser	-0.23	-2.64	360	0.09	** .009
SERP	0.03	1.34	360	0.02	.182
Bhv-Browser × SERP	0.12	0.70	360	0.17	.490

**Table 5.4:** Results of 4 predictive models with significant effects are included in the table. In bold are dependent variables of linear regression models, with the lines below separating each model. The first two results are for self-reported attitudes [Attd-General and Attd-Health], each on the self reported behaviour metric [Bhv-Browser]. [Bhv-Browser] is regressed on the [Normalized Number of Trackers] in the SERP for across search tasks. *M1* denotes model on data from the Stoplight SERP only (with no moderation). Model *M2* is moderated by the SERP type (Control or Stoplight), all other models had one input variable. \*\* and \* are used for  $p < .01$  and  $p < .05$  respectively.

Conversely, self-reported privacy behaviours are show some predictive power as to how users will behave in the Stoplight strategy, where small effects when considering reported usage of browsers that enhance privacy were found (again in Table 5.4). In the analysis, a distinct split between subjects (those who reported usage of privacy-enhancing browsers, and those who do not) was noted. Using this split, we performed post-hoc analysis with a Welch’s two-sided independent t-test comparing % of max trackers encountered in the Stoplight SERP for participants that use privacy-enhancing browsers ( $n = 42, M = 0.22, SD = 0.19$ ) with those that do not ( $n = 49, M = 0.28, SD = 0.24$ ) and find these differences to be significant  $t(179.79) = 2.11, p = .036, d = .31$ . The same split is used for within-group analysis to compare interactions against the Control SERP. For the group not using privacy-enhancing browsers, no differences were found for comparison of the Stoplight SERP ( $M = 0.30, SD = 0.15$ ) with the baseline SERP

( $M = .31, SD = 0.22$ ),  $t(47) = -0.46, p = .646, d = -0.07$ . For the group reporting usage of privacy-enhancing browsers there is a trend towards reduction in privacy encounters with the Stoplight SERP ( $M = 0.23, SD = 0.17$ ) compared to the Control SERP ( $M = 0.31, SD = 0.18$ ),  $t(36) = 1.91, p = .064, d = 0.31$ .

As already stated, similar analyses were performed for the Bhv-General and Bhv-Search metrics. For both Bhv-General and Bhv-Search, there were tendencies towards an interaction, but again no findings were statistically significant. We note that a limited number of participants ( $< 10$  reported any usage of privacy-protective "search" engines and  $< 20$  for any usage of "general" internet privacy protection methods) reported ANY behaviours contributing to these constructs.

### 5.4 Discussion

A total of four hypotheses (**H1a** - **H2b**) were considered in the analyses of the three *nudge* strategies for harm prevention in Web search. It is useful to discuss findings related to each of these hypotheses and the overall questions related to the viability and effectiveness of the approaches.

#### 5.4.1 Findings Related to Study Specific Hypotheses

Though significant findings were found across all *nudge* strategies, results indicate the differences compared to the Control were only significant with respect to the Filtering and Re-ranking systems (strategies **S1** and **S2**). Based on this finding, hypothesis **H1a** was not confirmed. However, given both Re-ranking and Filtering *nudge* strategies were non-transparent and their systems outperformed the transparent Stoplight *nudging* system hypothesis **H1b** was confirmed.

One can only speculate as to why overall the Stoplight system did not perform

significantly better overall compared to the Control. Nonetheless, the study was designed with the expectation that a subset of users will in fact make use of the intervention being provided by the Stoplight *nudge*. In the case of the current study, the subset of users was based upon attitudes towards privacy (**H2a**) and privacy protective behaviours (**H2b**). With respect to **H2a**, the findings are in line with existing literature on the "privacy paradox" (e.g. [163]), and the hypothesis is therefore not confirmed. Related to self-reported behaviours, with respect to the usage of a more privacy protective browser hypothesis **H2b** is confirmed. There were non-significant signals found for the other measures for privacy protective behaviour, however it is noted that very few participants take action in their normal lives to protect privacy.

Finally, based on the results, we conclude that the self-reported browser type is a useful pre-screening metric to identify participants more likely to be concerned about privacy, which is an important finding from the analysis.

#### 5.4.2 Findings in the Context of Research Questions

It is important to consider the findings in the context of the broader research questions (**G-RQ-1** and **G-RQ-2**). Table summarises the key findings from the offline study. Given the overall aim of these strategies was privacy prevention, it is clear that all 3 strategies are effective at reducing impacts from *3rd party trackers*. However, as highlighted in the findings, the Stoplight strategy is only partially effective and therefore the least effective. Related to **G-RQ-2**, this suggests that all 3 strategies are effective, just that strategies **S1** and **S2** are more so.

Additionally, the findings suggest there are no impacts on the overall decision making capabilities for any of the strategies. In other words, users of systems with

## 5. INVESTIGATING NUDGES IN AN OFFLINE SETTING

Strategy	Privacy Impact <i>(Harm)</i>	Task Decision <i>(Harm)</i>	Task Time <i>(Search Behaviour)</i>
Filtering <a href="#">S1</a>	Reduced	No Impact	No Impact
Ranking <a href="#">S2</a>	Reduced	No Impact	No Impact
Stoplight <a href="#">S3</a>	Reduced for a subset	No Impact	No Impact

**Table 5.5:** A summary of key findings critical for determining the effectiveness and viability ([G-RQ-1](#) and [G-RQ-2](#)) of each strategy. (*Italicized*) is the metric grouping from general methods, for example (*Harm*) is tied to Section 4.4.1

these strategies can have reduced privacy impacts (a positive *search outcome*) and simultaneously have no negative impacts on their medical decisions (a positive *search outcome* which ultimately may lead to a positive health outcome).

Furthermore, the search behaviour (as measured by task time) was not negatively impacted. As utility of a Web search environment (e.g. time to find information) and the possibility that good medical outcomes are more important than personal protection of privacy, a viable system should not negatively have any negative impacts with respect to these factors.

A remaining, but important factor with respect to user preferences was considered as well. While the data suggests that Re-ranking is not the most preferred approach by users, there are no clear winners and losers with respect to the most and least preferred approach (Table 5.3), which is an important factor in the context of viability ([G-RQ-1](#)).

Putting everything together including the findings in Table and [G-RQ-1](#), it appears that all three strategies are viable. However these findings were determined through a highly controlled lab setting, providing motivation for investigation in other settings.

### 5.4.3 Lessons Learned

Recall the current study was highly experimental and simultaneously the first lab based study for the author of this thesis, there were many learnings which allowed for improvements to the overall general methods.

For example, the behavioural questions related to privacy actions used in the current study did not consider the possibility that users may have never even heard of the action (e.g. use a VPN), the frequency of use, only allowed for yes / no responses and overlooked behaviours such as usage of privacy statements. As a result, the questionnaire was updated to include 4 additional items and used a 6 point Likert scale (see Table A.4. This is one example where improvements were made to the general methods, which ultimately allowed for richer analysis in later studies, in particular with the online study.

Another example relates to the Graeco-Latin square method for interleaving search tasks and strategies in a balanced manner. The current study demonstrated the challenges of Graeco-Latin square design. A minor glitch incorrectly mapped two of the strategies for  $n = 51$  participants. This glitch was thankfully discovered part way through. However it forced analyses related to hypotheses [H1a](#) and [H1b](#) to take place on a much smaller sample  $n = 40$ . For the future, when using Graeco-Latin square, one should triple check the expected presentation for participants (double-checking was not sufficient).

## 5.5 Summary

The preceding studies considered three *nudge* based interventions for harm prevention in Web search. All interventions (strategies [S1](#)- [S3](#)), evaluated with an

## 5. INVESTIGATING NUDGES IN AN OFFLINE SETTING

---

offline search system in a lab setting, appear to be effective at reducing potential harms related privacy impacts from *3rd party trackers*. The Stoplight *nudge* (strategy [S3](#)) is only effective for a specific type of user (users that take privacy protective actions in their daily lives). The remaining two strategies (Filtering and Re-ranking) designed for reduced privacy impacts are highly effective at achieving there goal. Therefore, we are confident that at least two of the strategies, if not all three are effective approaches for harm prevention in Web search, an important finding related to [G-RQ-2](#).

In addition to the *nudge* strategies evaluated, which touch upon framework components ([FC-Cognitive](#) and [FC-System](#)), several new evaluation metrics ([FC-Evaluation](#)) were introduced and important for comparison of the systems.

The offline study was highly controlled (e.g. a static corpus of documents presented in a SERP where users could not submit queries) and simulated a real-world search environment. Therefore, the findings supporting the viability ([G-RQ-1](#)) of the strategies is limited. This shortcoming of the current study is one motivation for the study presented in the next Chapter 6, which considers *nudge* strategies [S1](#)-[S3](#) in an interactive setting that is connected to the live Web.

# Chapter 6

## Investigating Nudges in an Online Setting

### 6.1 Overview

The previous chapter introduced and investigated three *nudge* strategies ([S1](#)-[S3](#)) as pathways to reduce harms related to loss of privacy as a result data shared with 3rd parties during Web search. One drawback of these studies was a highly controlled design that made use of a static test set. To better assess the robustness of these strategies it is useful to run experiments in a more naturalistic environment.

Recall from our previous chapter that many of the learnings in that study were used to formulate the overall general methodology. As such, much of the general methodology (Chapter 4) applies to the current study. There are some specifics to this study highlighted in the forthcoming methods section.

The main foci of the previous study was evaluating the viability and effectiveness ([G-RQ-1](#) and [G-RQ-2](#)) of the *nudge* strategies. Furthermore, the hypotheses in

## 6. INVESTIGATING NUDGES IN AN ONLINE SETTING

---

the previous study tested for variations between the three *nudge* strategies and the control as well as the contrast between the transparent Stoplight strategy [S3] and the non-transparent Filtering and Re-ranking strategies ([S1] and [S2]). Additional hypotheses considered user variations related to the transparent strategy. All of these factors are under consideration for this study as well. And like the previous study, the current study is designed with the cognitive and system components (FC-Cognitive and FC-System) at its core.

One benefit of the naturalistic online study is allowance for much more extensive analysis thanks to a much richer dataset. The dataset (outlined in more detail below) includes more detailed annotations with respect to the search task along with a broader set of search behaviour metrics. Given this, we were able to evaluate the systems (FC-Evaluation) from additional and more insightful angles when compared to the earlier study. Combined together, this allows the formulation for additional hypotheses to those used in the the offline study.

With respect to the high-level research questions G-RQ-1 and G-RQ-2 the following two hypotheses were tested in the offline study and are tested again in the current online study.

**H1a** All 3 *nudge* strategies will significantly outperform the Control search environment with respect to the harm being tested (privacy impact).

**H1b** Both non-transparent *nudge* strategies are expected to outperform the transparent *nudge* strategy with respect to privacy impacts.

In the offline study, the factor of time was one measure considered which was not significantly impacted. The online study captures data to consider behavioural changes such as increased queries, allowing for analyses on the user experience.



**H1c** All three *nudge* strategies are expected to have no significant degradation with respect to commonly used search behaviour measures.

The offline study was a controlled corpus with data skewed towards correct information for the search task, therefore an unreliable for comparison of systems with respect to information quality. However, the more naturalistic online study will expose users to information that is highly dynamic.

**H1d** With respect to information quality, all three experimental systems are expected to perform on par with the control system.

Specific to the transparent Stoplight *nudge* strategy, we repeat the tests on attitudes and behaviours related to privacy as addressed in the offline study.

**H2a** For the transparent *nudge* strategy, the “privacy paradox” will not be present. That is, participants with strong attitudes about privacy will have privacy impacts significantly reduced more so than participants with weaker attitudes.

**H2b** Similarly, we expect individuals taking privacy protective behaviours in everyday life to make use of the Stoplight *nudge* strategy more so than individuals that take little or no privacy protective action.

Additionally, the offline *nudge* study showed limited reductions in privacy impacts for the transparent Stoplight strategy. However, one should not overlook the nutrition literature (introduced in Chapter 2 demonstrating that Stoplight approaches are effective in general. Therefore, one must ask “Does search behaviour suggest users prefer results with colouration deemed to be more safe (e.g. Green is more safe than Red?”), motivating the following hypothesis.

**H2c** When compared to the non-transparent Control system, the transparent Stoplight approach will result in more users visiting results with colours deemed safe compared to colours deemed less safe.

Finally, specific to the transparent Stoplight strategy, challenges with linking real time search results were introduced in the general methods section. It was stated that results where no privacy data was available, a Gray light would be used to indicate unavailable privacy risk.

**H2d** It is expected that users will be (more / less) averse to web pages assigned an uncertain risk compared to web pages with a (low/high) risk.

We now turn to the methods used to test hypotheses **H1a** - **H2d**, followed by an extensive analysis.

## 6.2 Method

Many of the methods used in the current online study were introduced in Chapter 4 outlining overall general methods. In the following sections, we summarise the general methods in the context of the current online study. Additionally, there are details specific to this study that are introduced (e.g. evaluation metrics unique to this study).

### 6.2.1 Procedure

In general, the search tasks, search systems and study design are in alignment with those introduced in Chapter 4. All 10 Cochrane search tasks (Table 4.1) were used in the study, along with all three *nudge* strategies (**S1**- **S3**) which are introduced as three distinct search systems along with a control search system (Figures

4.2.2.1 - 4.2.2.4). Overall a *within-group* design utilizing a Graeco-Latin square balanced design presented participants the strategies and search tasks. Specific to the transparent Stoplight system, a *between-group* design was used for the analysis.

The main procedural difference for the current online studies compared to offline studies in other chapters is the usage of live search data. This difference is to more closely align with a naturalistic search task. In the current study, the SERPs included a query bar (Figures 4.2.2.1 - 4.2.2.4) which was connected to a commercial search API (see Appendix A.3 for technical details of the API ). The addition of the query bar and results from a live search environment is one key difference between the current online study and our more limited offline studies. As was done in the offline studies, results were linked to 3rd party trackers (as a measure of privacy impact). Details of how the queries were processed as well as how results were linked to the *3rd party trackers* are detailed in Appendix A.3. Methods used to collect 3rd party tracking data for web pages visited in the experiment follow the general methods in Section 4.3.1 with specifics to the current study in Section 6.2.2.

### 6.2.2 Evaluation Test Sets

Revisiting the general methods, the output of the current study was the **Online Nudging Test Set** which was used in the analyses covered in the forthcoming results Section 6.3.

The annotation methods related to misinformation (outlined in Section 4.3.2) were performed after all participants completed the experiment, and there are no deviations from these methods.

Similarly, there were no deviations with respect to the privacy annotation methodology (Section 4.3.1). However, a recap of this method with a running example (pre,

## 6. INVESTIGATING NUDGES IN AN ONLINE SETTING

---

during and post experiment) along with summary statistics are provided, as this data is fundamental to the overall harm prevention strategies under investigation.

**Pre-experiment** The pre-experiment privacy annotations are referred to as *Privacy Tracker Set A*. This set includes *3rd party trackers* from the **Offline Nudging Test Set** and the **WhoTracks.me Test Set**.

**During Experiment** At experiment runtime, multiple results were returned that could not be linked to *Privacy Tracker Set A*. This set of results is defined as *Privacy Tracker Set B*. For web pages in *Privacy Tracker Set B*, results were displayed in a different manner (see Section 4.2.2) from those in *Privacy Tracker Set A*. For example, results in *Privacy Tracker Set B* appeared with Gray lights in the Stoplight SERP in Figure 4.2.

**Post-experiment** After completing data collection for all subjects, a list of domains was compiled that were in *Privacy Tracker Set B*. Tracking data was linked using the same manual approach introduced in methods Section 4.3.1 using the **Ghostery** web browser plug-in. For our analysis, both *Privacy Tracker Sets A* and *Privacy Tracker Set B* were combined together along with the misinformation annotations.

**Summary of 3rd Party Tracking Linkage** In the experiment, 523 unique web pages were visited by participants, for which tracking data was linked to various sources. In instances where available, 3rd party tracking data in the **WhoTracks.me Test Set** was linked to 29.3% of the web pages. In instances where a link was not found in the **WhoTracks.me Test Set**, the *3rd party tracker* data in the **Offline Nudging Test Set** was used (linked to 31.2% of web pages). For web pages (39.5%) in which no links could be made to either dataset mentioned, 3rd party tracking data

was collected after all subjects completed the experiment. In summary, 60.5% of Web Pages were in *Privacy Tracker Set A* and 39.5% of Web Pages were in *Privacy Tracker Set B*.

### 6.2.3 Evaluation Metrics

An extensive set of metrics were used in the analyses provided in the results section for the current. Most of the metrics were already detailed in the general methods section. Here we highlight which of these metrics are brought into this study and how they are used in the later analyses. There are some additional metrics introduced that were not appropriate in the general methods, which are detailed in sub-sections that follow. One shortcoming of the offline study was the limited focus on system viability ( **G-RQ-1** ), which can be given much more scrutiny in the online study given the richness of data in the **Online Nudging Test Set**. Therefore, unless noted otherwise, all metrics are used for better understanding of system viability.

#### 6.2.3.1 Harm Prevention (Search Task Outcome)

Privacy metrics (Absolute Number of Trackers, Mean Number of Trackers and Normalized Number of Trackers) as a measure of the harm which the strategies aim to prevent, were used as dependent variables. These metrics cut across high level research questions ( **G-RQ-1** and **G-RQ-2** ) and study specific hypotheses ( **H1a** - **H2d** ) and therefore appear throughout most of the results.

Search outcome metrics specific to the search task decisions (Correct Decisions, Incorrect Decisions and Harmful Decisions) to ensure information encountered ( **H1d** ) does negatively impact search task decisions.

### 6.2.3.2 Compliance to Transparent Strategy

For the transparent *nudge* strategy, warning lights were assigned to levels of privacy risk and the colouration was determined by the number of trackers for each web page (see methodology in Section 4.2.2.2).

Though measures of privacy impacts (Section 6.2.3.1) may not be significantly reduced for this strategy (e.g. due to the method of linking 3rd party tracking data to results) it is still possible that some users make attempts to comply with the transparent strategy which are not apparent in the harm prevention metrics.

As such, we introduce two metrics that allow for analyses from a categorical perspective of adherence to the strategy and a secondary signal for the harm being prevented. These metrics are useful for **H2a** - **H2d** as well as better understanding viability and effectiveness (**G-RQ-1** and **G-RQ-2**).

**Assessments of Clicks by Colour** - Defined as the total number of clicks on each warning light colour

**Mean Assessments by Colour** - Defined as the total number of clicks on a specific warning light colour divided by the total number of clicks (e.g. user A visited 5 web pages, 3 of which were assigned Green lights and 2 assigned Gray. Thus Green = .60, Gray = .40, Yellow = .00 and Red = .00).

### 6.2.3.3 Self Report Measures

**Privacy Attitudes and Actions** We repeat the examination of self-report attitudes and behaviours with respect to privacy. Full details of the surveys used in this study are as presented in the general methods Section 4.4.4. The metrics (**Action Score (Binary)**, **Action Score (Total)**, **Awareness Score**, **Enhancing Browser Score**

and Privacy Attitudes (General)) are selected from the full list (Section 4.4.4.4) and used predominantly for G-RQ-1 and G-RQ-2 and hypotheses (H2a and H2b) related to the transparent Stoplight *nudge* strategy.

Some readers will note the metrics used in offline studies related to health attitudes and everyday search engine usage were excluded from the current study. For the health attitudes, we did not collect user responses for this metric in the current study (to shorten the post-experiment survey). For questions related search engine use, as with the previous study, usage of privacy enhancing search engines was so low (i.e.  $n = 4$  of  $N = 90$  participants reported use of any privacy enhancing search engine) that analysis was not possible with the participant sample size.

**Perceptions** Recall the Stoplight strategy in the current study has an element of uncertainty. To better understand user perceptions around uncertainty (H2d), we introduce the following measure.

**Perception of Warning Light** - Total count of participants for warning light colour  $X$  perceived as "most privacy risk" and "least privacy risk", for example 35 subjects perceive  $X = Green$  and 10 perceive  $X = Gray$  as "least privacy risk".

The metric was calculated with responses collected in the post experiment survey. The participants were asked “Which Stoplight indicates the most amount privacy risk if you click on the associated link” and “Which Stoplight indicates the least amount privacy risk if you click on the associated link?”. To limit the burden on participants, the questions did not attempt to understand the relationship between the Gray light and Yellow light, but rather to confirm the Gray light was not perceived as least or most risky.

**Preferences** Finally, as with the offline *nudge* study, we use the Preferred Strategy metric (see Section 4.4.4.2 to understand preferences of the participants.

### 6.2.3.4 Search System (Search Behaviour)

Unlike the offline *nudge* study where search behaviour impacts were limited to time impacts, the current study makes use of the full set of behavioural metrics introduced in the general methods Section 4.4.3.1 (Assessments, Queries, Queries w/o SERP clicks, Rank and Time to Completion). We also include in the search behaviour group is the self-reported Search Task Confidence metric; though subjective compared to the other objective measures, this metric is an indicator of perceived impact to search behaviour. All behavioural metrics are relevant to hypothesis **H1c**.

### 6.2.4 System Evaluation (Adapting IR Metrics)

Motivation for usage of IR system metrics (when appropriate) was provided in the general methods Sections 4.4.3 and Section 4.4.3.2, as these are seen as useful (and potentially the most valuable approach) for understanding the viability (**G-RQ-1**) of strategies. Thought introduced in the general methods as part of evaluation (for which these are most certainly part of), the details of the adapted IR metrics used in the current study are extensive enough to be given a section upon itself.

For the analyses we employ traditional IR metrics (Mean Reciprocal Rank (MRR) and Precision) for our analyses. Furthermore, motivated by more recently developed evaluation approaches which take a probabilistic view of the system, we create a custom Cumulative Click Probability metric. As IR system evaluations predominantly consider relevance or graded relevance of information retrieved, it was necessary to make adaptations. Recall our evaluation test set (**Online Nudging Test Set**) does



not use relevance assessments, however it does contain annotations related to misinformation. It is these annotations ((*Correct*, *Incorrect* and *Harmful* introduced in the annotation procedures in general methods Section 4.3.2) which are used for the metrics we introduce below. Therefore the adaptation being these metrics consider the misinformation aspect of each document retrieved (as opposed to the commonly used relevance of a document).

It must not be overlooked that MRR is a somewhat controversial metric [73, 193], but nonetheless is useful [193] and is certainly an appropriate metric for the evaluation test set (**Online Nudging Test Set**) used in current analyses. The main challenge to usage of less controversial metrics (e.g. nDCG) being that misinformation annotations were only performed on results visited by participants (we did not annotate results that users skipped over).

One might ask why metrics such as *recall*, another traditional IR metric, and other metrics that take into account false negatives (e.g. *correct* documents that a participant skipped over) were not included in our analyses. Again, this has to do with the live environment used in our study where an unfortunate challenge prevented annotations on results returned by the commercial API that participants skipped over. The challenge being that API end user agreements had limitations on length of time that the results could be held, therefore we could not permit annotators to review these results that were not visited.

With that said, we now introduce how the three metrics were calculated.

### 6.2.4.1 Mean reciprocal rank (MRR)

For MRR, the approach suggested by [96] is used as the basis for adaptation in our analysis.

## 6. INVESTIGATING NUDGES IN AN ONLINE SETTING

---

MRR usually considers the rank of the first ‘relevant’ document for a set of given queries [96]. However, for our analysis, we consider the rank of the first *correct*, *incorrect* and *harmful* documents across a set of user queries. In this manner, we can produce MRR metrics for all 3 document types, where higher MRR is desirable for *correct* documents and lower MRR is desired for *incorrect* and *harmful* documents.

For the current study, MRR is based on the user queries and their clicks. All queries are treated as unique, that is duplicate queries (the same query submitted by different users) are treated separately. We also treat the rank of the first click by each user of a duplicate query as unique.

Considering the first click across all user queries for each search task, the MRR is calculated for each result type (*Correct*, *Incorrect* and *Harmful*), with the three following three metrics used in analyses.

**MRR Correct** - Mean Reciprocal Rank for *Correct* documents

**MRR Incorrect** - Mean Reciprocal Rank for *Incorrect* documents

**MRR Harmful** - Mean Reciprocal Rank for *Harmful* documents

**Example** As an example of how MRR was calculated in our evaluation, consider user A and user B who both issued the query ‘obesity surgery’ for the same search task X. User A clicks on a *correct* document at rank 2 and user B clicks on a *correct* document at rank 4. In this toy example,  $MRR = 1/2 * (1/2 + 1/4) = 3/8$ . User B then entered a different query ‘obesity reduction surgery’ for search task X, and click on the first document at rank 1 which is also a *correct* document. The final MRR for *correct* documents for all 3 queries submitted by users A and B for search task X becomes  $MRR = 1/3 * (1/2 + 1/4 + 1) = 7/12$ . MRR is calculated in the

same manner for all 10 search tasks and for each document type (*correct*, *incorrect* and *harmful*).

Our analysis also considers MRR at two different levels with respect to queries and clicks and make the distinguishment of includes vs. excludes queries with no clicks. The above example is how MRR will be calculated for test set data excludes queries with no clicks.

Continuing with the example above, we demonstrate how MRR which includes queries with no clicks is calculated. Using this example, an additional user (user C) is included in the analysis. User C submitted ‘surgery helps obesity’ for search task X, but did not click on any documents. For analysis that considers queries without clicks, MRR is updated to  $MRR = 1/4 * (1/2 + 1/4 + 1 + 0) = 7/16$ . Analysis in this manner took into account query abandonments, which is not necessarily bad [96].

### 6.2.4.2 Precision

Our data collection method allowed us to calculate precision metrics based on the websites visited by users and the annotations of those websites. Unlike traditional IR system evaluations where a true positive is defined as a relevant document, analysis herein defines the true positive as a click on a *correct* document and also considers clicks on *non-harmful* documents, where *non-harmful* documents are defined as any document that was not annotated as *harmful*, translating to the following two metrics.

**Precision Correct** - defined as the total number of clicks on a *correct* web page divided by (number of *correct* web page clicks + number of *incorrect* web page clicks).

## 6. INVESTIGATING NUDGES IN AN ONLINE SETTING

---

**Precision Not Harmful** - defined as the total number of clicks on a *correct* web page divided by (number of *correct* web page clicks + number of *harmful* web page clicks).

For both metrics, analyses include calculations at rank  $K = (1, 2, 10, 20, 50)$ . 50 is chosen as the maximum due to the maximum possible returned by the commercial search API.

### 6.2.4.3 Cumulative Click Probability

Motivated by more recently developed metrics which take probability into account, which unfortunately could not be adapted to our test set based on reasons stated above.

As such, we introduce a custom metric as a probabilistic signal for the likelihood of encounters with results containing misinformation (performed at rank  $k$ ).

To perform such an evaluation, cumulative probability sums were calculated for each system at each rank for each annotation type (*correct*, *incorrect* and *harmful*).

**Calculating cumulative click probabilities** - Cumulative probabilities were calculated in the following manner.

Let  $K$  be the max rank of clicks to consider in the analysis and  $k$  be any rank between 1 and  $K$ .

Let  $s$  be a search system amongst a set of  $S$  search systems to evaluate.

Let  $C$  be the set of all clicks that occurred in  $S$  for  $k = 1$  to  $K$ , and  $c_{class}$  be the number of clicks of the document credibility classification that occurred at rank

$k$ .

Let  $cprob_{class}@k = c_{class}/C$  be the probability of a click at rank  $k$ .

Let  $cumprob_{class}@i = \sum_{k=1}^{k=i} cprob_{class}@k$

1. For each  $k = 1$  to  $K$ ,  $s$  in  $S[control, filtering, re - ranking, stoplight]$  and  $class$  in  $classes[correct, incorrect, harmful]$
2. Calculate  $cprob_{class}@k$
3. Calculate  $cumprob_{class}@i$  at  $i = k$

**Calculation in practice** First, the number of clicks that occurred between  $k = 1$  and  $k = K$ , where  $K = 10$  for each interface is determined. Second, for each  $k$ , the probability of a click occurring of each web page classification (*Correct*, *Incorrect* and *Harmful*) is calculated. Finally, the probabilities from  $k = 1$  to  $k = K$  are accumulated. This process produces a cumulative probability of a click occurring for a each SERP result classification (*Correct*, *Incorrect* and *Harmful*) by rank  $k$ .

This approach results in the following three metrics and is calculated for the 3 *nudge* based systems and the Control system,s.

**Cumulative Click Probability Correct** - Cumulative Click Probability of a click on *Correct* document at rank  $k$

**Cumulative Click Probability Incorrect** - Cumulative Click Probability of a click on *Incorrect* documents at rank  $k$

**Cumulative Click Probability Harmful** - Cumulative Click Probability of a click on *Harmful* documents at rank  $k$

### 6.2.5 Statistical Tests

There are some important statistical approaches to introduce here not already covered in general methods Section 4.5.

There were multiple instances where test assumptions were not satisfied (e.g. non-normality of data) with the data collected in the current study and therefore the Kruskal-Wallis test (a non-parametric test) were used in addition to LMER. Violation of parametric tests increases the likelihood of a type I error, and non-parametric tests have an increased likelihood of a type II error and the possibility for such errors were one justification for including results from both tests. Additional challenges arise with Kruskal-Wallis due to tests being performed across all experimental variants (i.e. any significant findings provide no further insight as to which approach was best). This motivates the combination of reporting Kruskal-Wallis and LMER models. LMER is treated as a type of post-hoc analyses to allow for cross-comparisons between the experimental strategies and the Control. In summary, for many of the analyses, Kruskal-Wallis was used to test overall effects and linear mixed effects models (with LMER in R) were used to test comparisons of strategies [S1](#)-[S3](#) against the Control.

Using a conservative approach (Bonferroni adjustment),  $\alpha_{adjusted} = (\alpha = .05)/3 = .017$  was used for these post-hoc analyses to test significance of individual strategies against the Control system (e.g. Filtering *nudge* vs. Control system, Stoplight *nudge* vs. Control system). This value is based on 3 experimental systems (strategies [S1](#)-[S3](#)) being evaluated against a control system.

Other tests were also leveraged in our analyses worth highlighting.

In line with authors who developed and analysed the **Waterloo Test Set** [\[179\]](#)

to test impacts on search task medical decisions, we also make use of Generalized linear mixed (GLM) effects with likelihood ratio tests. GLM is a robust approach to test for difference with categorical dependent variables (e.g. search tasks being *correct*, *incorrect* or *harmful*). Though LMER could also have been used, we chose the other method as it may be useful for future research that compares our findings with theirs.

Multiple linear regression was used to test for system differences with regards to the cumulative probability of a click on type of information. Multiple regression was chosen to test for potential interactions between the rank and the strategy.

Correlation analyses was also performed in some analyses. Spearman rank correlation was used to test for links between the transparent Stoplight strategy and self-reported metrics related to privacy attitudes and actions. Chi-squared ( $\chi^2$ ) was used to test for difference with count variables (e.g. user perceptions of warning lights).

### 6.2.6 Participants

Approximately 84% of the subjects were recruited from the Psychology department participant pool. Another 16% were recruited via list servers and department emails at the University.

For the current study, a participant sample ( $N = 90$ ) was used. Thirty-six participants were female, fifty-three were male, and one reported as other. Mean age (with standard deviations) was 27.8 (8.2). The university (based in an English speaking country) considers itself to be quite international and the data support that claim, with 49.4% reporting Caucasian ethnicity and 38.9% reporting English as the primary language spoken in their home. Of all participants, 4.4% were non-members

of the university (i.e. not students or staff), 88.9% were students of which  $n = 44$  were undergraduates and  $n = 36$  were post-graduates. A total of 18 participants were studying Computer Science at the university.

### 6.3 Results

All  $N = 90$  participants completed the main experiment, that is all tasks were completed for each participant and no data was missing from the interactive component. With the Graeco-Latin balanced design in the study, 9 rotations in total were completed (10 participants per rotation). Therefore each experimental (and Control) search systems was visited 180 times in total (with 90 *helpful* and 90 *unhelpful* search tasks associated to each system).

Analyses for all research questions are provided in the sections below. Annotations of web pages visited in the experiment were foundational to many of the analyses, therefore an analytical overview of the annotations is provided first.

#### 6.3.1 Annotations for Misinformation

A total of 529 web pages were annotated by two annotators using the methods in Section 4.3.2. Each annotator independently assessed all 529 result search task pairs. Reliability of the annotations was measured using Cohen's  $\kappa$ . Across all web pages and the 4 classifications used, annotator 1 had weak reliability  $\kappa = 0.59$ , ( $p < .001$ ) with annotator 2. The annotators then jointly resolved any discrepancies with the 4 classification types (see general methods Section 4.3.2 for additional details of the annotations). The 4 classifications were then mapped to the categories (*Correct*, *Incorrect* and *Harmful*) used in the analyses of research questions. Table 6.1 provides an overview of the total counts with each categorical mapping. Note also, these



annotations were analysed at a less granular level (*Correct* and *Incorrect* only) in empirical Chapter 7, with findings that suggest the annotations are of very high quality.

Annotation	Categories for Analysis	Total
<i>Correct</i>	<i>Correct</i>	286
<i>Incorrect - Page Unavailable</i>	<i>Incorrect</i>	16
<i>Incorrect - Not enough information</i>	<i>Incorrect</i>	171
<i>Incorrect - Wrong Information</i>	<i>Incorrect + Harmful</i>	56

**Table 6.1:** Break down (Total by Type) of result misinformation assessments for results in the current study

### 6.3.2 System Impacts on Privacy and Decisions

Metrics (see 6.2.3.1) used for the analysis of privacy impacts were aggregated by the search system (Control + 3 *Nudge* systems) and search task for each user. For example, the average number of trackers a user encountered is aggregated at the task level. The same aggregation method was used for analysis of the search task outcome of the user’s search task medical decision (*Correct*, *Incorrect* or *Harmful*). The user could only make one decision per task, therefore aggregation had no impact to this dependent variable.

Results are first provided for the harm (privacy) impacts to being prevented, followed by overall impacts to search task decisions.

#### 6.3.2.1 Harm Prevention (Privacy)

Both Kruskal-Wallis and LMER were used to test the predictions of user privacy impacts. Across all privacy metrics considered, significant overall effects were found due to the SERP when using Kruskal-Wallis. LMER results indicate that the Stoplight strategy is only marginally effective at reducing privacy impacts across

## 6. INVESTIGATING NUDGES IN AN ONLINE SETTING

the subject pool, however Filtering and Re-ranking appear to be quite robust in reducing impacts to privacy. Results are detailed in Table 6.2, which are used for discussions with respect to hypotheses **H1a** and **H1b**.

	Control	Filter	Rank	Stoplight	Kruskal-Wallis (df = 3)	
DV	$M(SD)$	$M(SD)$	$M(SD)$	$M(SD)$	$\chi^2$	$p$
Trackers [% of Max Possible]	0.32 (0.22)	0.22 (0.14)***	0.15 (0.08)***	0.30 (0.21)	100.92	< .0001
Trackers [Average]	6.78 (4.79)	4.61 (3.23)***	3.19 (1.80)***	6.08 (4.28) <sup>-</sup>	92.866	< .0001
Trackers [Total]	18.64 (17.67)	14.29 (14.55)***	9.79 (8.42)***	17.15 (15.25)	32.321	< .0001

**Table 6.2:** Privacy - Included in the table are analyses of privacy metrics introduced in Section 6.2.3.1. LMER results comparing experimental SERPs to the control SERP are indicated by <sup>-</sup> =  $p < .05$ , \* =  $p < .05$  (with Bonferroni correction), \*\* =  $p < .001$ , \*\*\* =  $p < .0001$ . Means (M) and standard deviations (SD) are also included.

### 6.3.2.2 Search Task Outcome (Medical Decision)

Analysis of impacts to user search decisions was performed in the same manner as Pogacar et al. (see [179]) which allows for a more simplified cross-comparison with their findings. A GLM (with mixed and fixed effects) was used to test impacts to the search task medical decision outcomes (each outcome measure was binary) for the subjects. A likelihood ratio test (with  $\chi^2$  test) was used for significance testing.

Results of the analysis of impacts to user decisions are provided in Table 6.3, which include comparisons of the *nudge* strategy based systems with the Control system. A summary of means and standard errors of the user decisions are provided in Table 6.5. The overall model indicates no overall significant impacts to user decisions, however the model results clearly indicate tendencies for differences with respect to *Harmful* decisions.

Results in Tables 6.3 and 6.4 provide evidence related to hypotheses **H1c** and **H1d**.

The differences with respect to *harmful* decisions was motivation for post-hoc

IV	DV (Decision)	Pr(> $\chi^2$ )
SERP	Correct	.4566
SERP	Harmful	.0828

**Table 6.3:** Decisions - Likelihood Ratio Tests from GLM (mixed effects) on the impacts of experimental SERPs on user medical decisions. Overall the experimental SERPs have no significant negative impacts compared to the control.

analyses with findings (Table 6.4) that indicate *Harmful* decisions are significantly reduced for the Stoplight strategy (S3) compared to the Control system. There are similar (though non-significant) tendencies for the other two strategies.

Variable	Estimate	Std Error	Pr(> z )
Stoplight (SERP)	-.743	.308	.0159*
Filtering (SERP)	-.380	.289	.1886
Re-Ranking (SERP)	-.576	.298	.0535

**Table 6.4:** Decisions (Harmful) - Summary of GLM (mixed effects model) of the fixed effects of each *nudge* system compared to the Control system. In all cases, the negative estimates indicate a reduction in *harmful* decisions compared to the Control, with \* =  $p < .05$  (with Bonferroni correction) for the Stoplight *nudge*.

SERP	Total Decisions	Total Correct	Total Harmful	Mean Correct	Mean Harmful
No SERP (Baseline)	180	67	33	.37 +/- .04	.18 +/- .03
Control	180	94	37	.52 +/- .04	.21 +/- .03
Stoplight	180	103	21	.57 +/- .04	.12 +/- .02
Filter	180	90	28	.50 +/- .04	.16 +/- .03
Rank	180	96	24	.53 +/- .04	.13 +/- .03

**Table 6.5:** Decisions - In line with previous methods of reporting for task decisions for the 10 search tasks (see [179]), totals, means and standard errors are provided for the *correct* and *harmful* decisions made by subjects for each SERP. Total counts of decisions are included (*incorrect* decisions are excluded as they can be calculated from values in table).

### 6.3.3 Search Behaviour

Prior to performing the analysis to determine potential impacts to search behaviour, post-experimental tests for reliability of the two questions to capture post-task confidence (introduced in general methods Section 4.4.4.3) was performed, for which Cronbach's  $\alpha = .91$  was found, demonstrating strong internal consistency. These questions were used to formulate the **Search Task Confidence** metric.

For the analyses of user search behaviour, metrics analyzed were aggregated by the search system and search task for each user. Analyses was then completed with Kruskal-Wallace and LMER models.

Results of general behavioural (Table 6.6) impacts based upon the six dependent search behaviour metrics (outlined in Section 6.2.3 are given first. Separate analyses (Table 6.7) are then provided with respect to the **Assessments** metric to assess potential concerns related misinformation exposure. All results provided below are critical for hypothesis **H1c** and to some extent **H1d**.

#### 6.3.3.1 General Impacts

Analysis with Kruskal-Wallace indicates no impacts due to the SERP on user search behaviour. Linear mixed effects models indicate trends towards increased total assessments for the Re-ranking strategy **S2** compared to the Control system. Means and standard deviations for each behavioural metric across all search systems are included in Table 6.6 in addition to statistical findings.

#### 6.3.3.2 Assessments and Misinformation

Analysis with Kruskal-Wallace and LMER models were used to test interactions with information quality with respect to the three *Nudge* systems and Control sys-

## 6.3 Results

	Control	Filter	Rank	Stoplight	Kruskal-Wallace (df = 3)	
DV	$M(SD)$	$M(SD)$	$M(SD)$	$M(SD)$	$\chi^2$	$p$
Assessments [Total]	2.52 (1.98)	2.64 (2.02)	2.80 (1.94) <sup>-</sup>	2.66 (1.86)	2.9224	.4037
Queries [Total]	1.62 (1.48)	1.49 (0.87)	1.58 (1.04)	1.63 (1.41)	1.3324	.7215
Queries w/o SERP Click	0.09 (0.29)	0.12 (0.32)	0.08 (0.27)	0.09 (0.29)	1.7584	.6240
Rank [Average]	3.57 (3.39)	3.52 (3.23)	3.90 (3.55)	3.92 (3.49)	5.2180	.1565
Time [Seconds]	160 (136)	141 (102)	159 (133)	158 (127)	2.3657	.5001
Search Task Confidence	5.58 (1.26)	5.53 (1.30)	5.66 (1.20)	5.65 (1.19)	0.8210	.8444

**Table 6.6:** System impacts on measures for search behaviour. Means  $M$  and standard deviations ( $SD$ ) are provided for each search system. Linear mixed effects results comparing experimental SERPs to control SERP are indicated by <sup>-</sup> =  $p < .05$ , <sup>\*</sup> =  $p < .05$  (with Bonferroni correction), <sup>\*\*</sup> =  $p < .001$ , <sup>\*\*\*</sup> =  $p < .0001$ .

tem. The dependent variable in this case is the information quality of the result assessed by the participant (i.e. was the information in the website visited by the user *correct*, *incorrect* or *harmful* as it relates to the search task).

Results are detailed in Table 6.7. Analysis with Kruskal-Wallace indicates no significant overall effects. However, when considering results from the LMER mixed effects models, indications are the Re-ranking and Filtering *nudge* strategies may in fact steer users toward documents that are *incorrect*. On the contrary, the models provide weak signals that the Stoplight *nudge* increases exposure to *correct* results when compared to the Control system.

	Control	Filter	Rank	Stoplight	Kruskal-Wallace (df = 3)	
DV	$M(SD)$	$M(SD)$	$M(SD)$	$M(SD)$	$\chi^2$	$p$
Assessments [Correct]	1.66 (1.39)	1.59 (1.27)	1.72 (1.38)	1.86 (1.38) <sup>-</sup>	4.1351	.2472
Assessments [Incorrect]	0.86 (1.27)	1.06 (1.38) <sup>-</sup>	1.08 (1.36) <sup>*</sup>	0.80 (1.07)	5.8284	.1203
Assessments [Harmful]	0.16 (0.42)	0.26 (0.70) <sup>*</sup>	0.18 (0.48)	0.14 (0.43)	2.3428	.5044

**Table 6.7:** Assessments by information type for each search system. Included in the table are the mean and standard deviations of user assessments by result type (*correct*, *incorrect* or *harmful*). Linear mixed effects results comparing experimental SERPs to control SERP are indicated by <sup>-</sup> =  $p < .10$ , <sup>\*</sup> =  $p < .05$ .

### 6.3.4 Search System Evaluation

In the following sub-sections, all search systems the Control search system and three *Nudge* based search systems designed for prevention of harm from privacy impacts are evaluated from the view of the credibility (i.e. quality) of information provided to participants during the experiment. Evaluation of the systems are performed with the IR system evaluation metrics introduced in Section 6.2.4. The evaluation is performed for the purpose of testing hypothesis **H1d** and responding to overall research questions (in particular **G-RQ-1**).

#### 6.3.4.1 Comparisons with MRR

Results of the comparison of systems with MRR are provided in Table 6.8. Means and standard errors were calculated across the 10 search tasks and search system pairs, which provides an aggregate summary of quality of information one might expect with each search system with MRR as the evaluation metric.

Linear mixed effects models (search task being the mixed effect) were created (with LMER packager in R). Though no significant differences were found, the Re-ranking strategy notably stands out as worst performer with respect to MRR for *correct* information.

	SERP	Correct	Incorrect	Harmful
MRR <u>includes</u> queries with no clicks	Control	.34 +/- .08	.15 +/- .04	.02 +/- .01
	Stoplight	.34 +/- .06	.13 +/- .03	.02 +/- .01
	Filter	.34 +/- .07	.13 +/- .03	.02 +/- .02
	Rank	.25 +/- .06	.13 +/- .03	.02 +/- .01
MRR <u>excludes</u> queries with no clicks	Control	.68 +/- .05	.62 +/- .10	.21 +/- .13
	Stoplight	.60 +/- .05	.65 +/- .12	.23 +/- .12
	Filter	.70 +/- .05	.62 +/- .10	.17 +/- .12
	Rank	.56 +/- .07	.67 +/- .05	.23 +/- .12

**Table 6.8:** IR Metric (MRR) - Means and standard errors of MRR across all 10 search tasks based upon *correct*, *incorrect* and *harmful* assessments with each system. Analysis considers two views, inclusion and exclusion of queries with no clicks (i.e. user abandonments). For *correct* MRR, higher numbers are better. For *incorrect* and *harmful* MRR, lower numbers are better.

### 6.3.4.2 Comparisons with Precision

Evaluations were performed across all systems with the precision metrics **Precision Correct** and **Precision Not Harmful** (see Section 6.2.4.2). Results in Table 6.9 for precision ( $rank@k, k = 50$ ), where  $k = 50$  is chosen due to the maximum possible results. Similar to the MRR analyses, means and standard errors were based upon precision metrics across the 10 search tasks for each search system (3 *Nudge* systems + Control system). Mixed effects models (with LMER) were also produced to test for significant differences between the systems. Though no differences were significant, with respect to both precision metrics in the Control system, the Filter and Re-ranking (strategies **S1** and **S2**) under perform and the Stoplight strategy **S3** slightly over-performs.

## 6. INVESTIGATING NUDGES IN AN ONLINE SETTING

---

SERP	Total Tasks	Precision (Correct)	Precision (Not Harmful)
<b>Control</b>	10	.68 +/- .07	.91 +/- .04
<b>Stoplight</b>	10	.71 +/- .05	.93 +/- .03
<b>Filter</b>	10	.62 +/- .06	.87 +/- .06
<b>Re-ranking</b>	10	.62 +/- .05	.90 +/- .03

**Table 6.9:** Precision of *Correct* and *Not Harmful* assessments for each search system. Similar to MRR, precision was calculated for each task, and then means and standard errors were calculated. Higher means indicate better system performance.

Table 6.10 includes precision @ rank  $k$  ( $k = 1, 2, 10, 20$  and  $50$ ), with metrics aggregated at the global level across all search tasks. Calculating precision in this manner prevents testing for statistical significance, however it does provide some signals as to which interventions may be problematic when compared to the Control system (highlighted further in the discussion).

Finally, the precision metrics allow for calculations of the false discovery rates (FDR) of *incorrect* and *harmful* assessments for each system. To do so, take the precision values provided in Tables 6.9 and 6.10 and subtract one from the value.

$$FDR = 1 - Precision$$



P @ k	System	Total Clicks	Total Correct	Total Incorrect	Total Harmful	Precision Correct	Precision No-Harm
1	Stoplight	128	79	49	7	.62	.92
	Control (Bing)	133	81	52	9	.61	.90
	Rank	128	61	67	8	.48	.88
	Filter	143	91	52	8	.64	.92
2	Stoplight	204	132	72	9	.65	.94
	Control (Bing)	215	139	76	12	.65	.92
	Rank	214	120	94	14	.56	.90
	Filter	224	143	81	14	.64	.91
10	Stoplight	431	299	132	21	.69	.93
	Control (Bing)	424	276	148	26	.65	.91
	Rank	475	291	184	32	.61	.90
	Filter	436	256	180	43	.59	.86
20	Stoplight	468	327	141	24	.70	.93
	Control (Bing)	446	293	153	27	.66	.92
	Rank	495	304	191	33	.61	.90
	Filter	465	278	187	46	.60	.86
50	Stoplight	478	334	144	25	.70	.93
	Control (Bing)	454	299	155	28	.66	.91
	Rank	504	309	195	33	.61	.90
	Filter	476	286	190	47	.60	.86

**Table 6.10:** Precision @ k of *Correct* and *Not Harmful* assessments for each search system calculated at  $k = 1, 2, 10, 20$  and  $50$  ( $K = 50$ ). Calculations are performed globally across all search tasks (therefore no means and standard errors could be calculated). Total clicks are also included at each level of  $k$ .

### 6.3.4.3 Comparisons with Cumulative Probability

Using the Cumulative Click Probability metric introduced in Section 6.2.4.3 statistical tests (multiple linear regression including the interactive effects of rank and the search system) were performed to compare the systems. As 95% of all user clicks collected in the experiment occurred between  $k = 1$  and  $k = 10$ , data became quite sparse beyond this rank and resulted in violation of statistical tests for  $k = 20$  and  $k = 50$ , therefore  $k = 10$  was used in the analyses. Results in Table 6.11 indicate significant differences for all information types (*Correct*, *Incorrect* and *Harmful*).

These significant differences motivated post-hoc analyses (with Bonferroni correction) for each information type for comparison of *nudge* strategies against the

## 6. INVESTIGATING NUDGES IN AN ONLINE SETTING

Control system. Presented in order by *Correct*, *Incorrect* and *Harmful* information, results Tables (6.12 - 6.14) are paired with corresponding visualisations (Figures 6.1 - 6.3) of the cumulative probability of click at rank  $k$ .

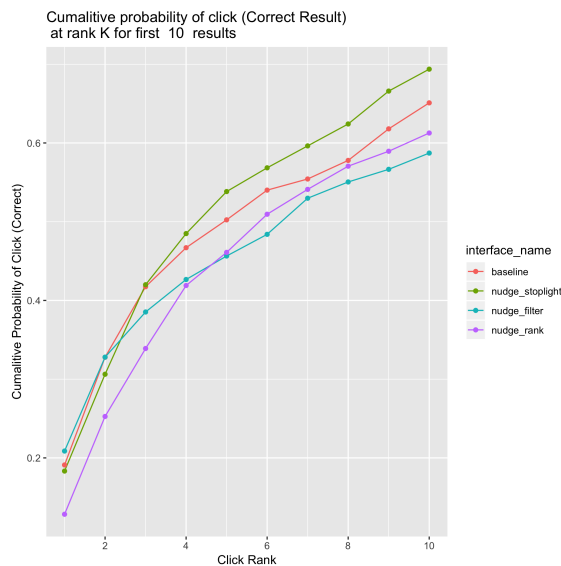
Click Type (DV)	Mult. $R^2$	p
Correct	.915	<.0001***
Incorrect	.924	<.0001***
Harmful	.947	<.0001***

**Table 6.11:** Click Probabilities: Overall  $R^2$  and  $p$  values for multiple linear regression models to test the effects of rank and search system against the dependent variable (cumulative probability of click on information type).

**Results: Cumulative Probability of Click on *Correct* Information** Results Table 6.12 and Figure 6.1 indicate no significant differences (nor any interactive effects with rank @  $k$ ) between the harm prevention systems and the Control system with respect to the probability of clicking on *correct* information. Nonetheless, there are indications that the quality of *correct* information in both the Filtering and Re-ranking strategies becomes less likely compared to the Control as rank becomes deeper.

Variable	Estimate	Std Error	Pr(> t )
Result Rank	.044	.005	<.0001***
Stoplight (SERP)	.023	.021	.2632
Filtering (SERP)	-.032	.021	.1257
Re-Ranking (SERP)	-.042	.021	.0479 <sup>-</sup>
Result Rank : Stoplight (SERP)	.007	.007	.3081
Result Rank : Filtering (SERP)	-.006	.007	.3892
Result Rank : Re-ranking (SERP)	.006	.007	.3962

**Table 6.12:** Model for predicting the cumulative probability of a click on *correct* information between ranks  $k = 1$  to  $k = 10$  for each experimental system against the control SERP. <sup>-</sup> =  $p < .05$ , \* =  $p < .05$  (with Bonferroni correction), \*\* =  $p < .001$ , \*\*\* =  $p < .0001$



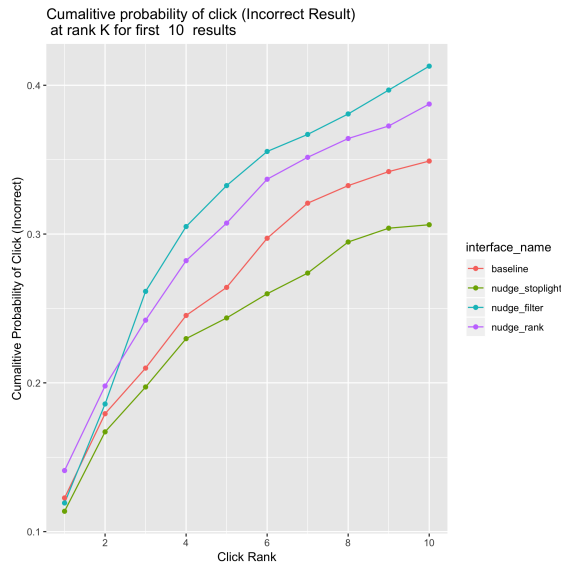
**Figure 6.1:** Comparison of cumulative probability of clicks on *correct* information compared to rank  $k$ , for  $k \leq 10$ .

**Results: Cumulative Probability of Click on *Incorrect* Information** Results Table 6.13 and Figure 6.2 indicate the Re-ranking and Filtering strategies a significant increase in probability of clicking on *incorrect* information compared to the Control system. Though above the threshold for significance, the Stoplight approach reduces the number of clicks on *incorrect* web pages, suggesting the Stoplight *nudge* (strategy [S3]) may be better than the Control with respect to returning results with higher quality information. No interactive effects were found between the systems and rank @  $k$ .

Variable	Estimate	Std Error	Pr(> t )
Result Rank	.025	.003	<.0001***
Stoplight (SERP)	-.027	.011	.0204 <sup>-</sup>
Filtering (SERP)	.045	.011	.0003***
Re-Ranking (SERP)	.032	.011	.0073*
Result Rank : Stoplight (SERP)	-.004	.004	.2694
Result Rank : Filtering (SERP)	.005	.004	.1830
Result Rank : Re-ranking (SERP)	.001	.004	.7126

**Table 6.13:** Model for predicting the cumulative probability of a click on *incorrect* information between ranks  $k = 1$  to  $k = 10$  for each experimental system against the control SERP. <sup>-</sup> =  $p < .05$ , \* =  $p < .05$  (with Bonferroni correction), \*\* =  $p < .001$ , \*\*\* =  $p < .0001$

## 6. INVESTIGATING NUDGES IN AN ONLINE SETTING

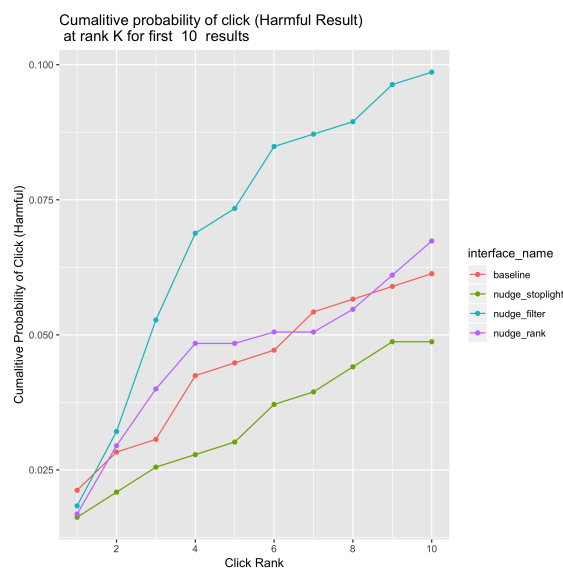


**Figure 6.2:** Comparison of cumulative probability of clicks on *incorrect* information compared to rank  $k$ , for  $k \leq 10$ .

**Results: Cumulative Probability of Click on *Harmful* Information** Results Table 6.14 and Figure 6.3 indicate the Stoplight *nudge* results in a significant reduction in the probability of clicking on a *harmful* website, which suggests that the Stoplight approach is better than any other system (including Control) for information quality. The Filtering *nudge* is found to have a significant increase in the probability of clicking on *harmful* information compared to the Control. Significant interactive effects were found between the Filtering system and rank @  $k$  suggesting the probability of encountering *harmful* information increases significantly as one assesses documents deeper in rank.

Variable	Estimate	Std Error	Pr(> t )
Result Rank	.005	.001	<.0001***
Stoplight (SERP)	-.011	.003	.0002***
Filtering (SERP)	.026	.003	<.0001***
Re-Ranking (SERP)	.002	.003	0.3938
Result Rank : Stoplight (SERP)	-.001	.001	.4054
Result Rank : Filtering (SERP)	.004	.001	<.0001***
Result Rank : Re-ranking (SERP)	.000	.001	.9176

**Table 6.14:** Model for predicting the cumulative probability of a click on *harmful* information between ranks  $k = 1$  to  $k = 10$  for each experimental system against the control SERP.  $^{\circ}$  =  $p < .05$ ,  $^*$  =  $p < .05$  (with Bonferroni correction),  $^{**}$  =  $p < .001$ ,  $^{***}$  =  $p < .0001$



**Figure 6.3:** Comparison of cumulative probability of clicks on *harmful* information compared to rank  $k$ , for  $k \leq 10$ .

### 6.3.5 Strategy Preference

As with the offline study, data was collected for the participant preferences across the three strategies. Table 6.15 provides the results of the most and least preferred strategies, which uses  $\chi^2$  analysis to test for significant differences and important for **G-RQ-1**.

Intervention	Most Preferred	Least Preferred
Filtering SERP	15	54
Ranking SERP	10	22
Stoplight SERP	65	14

**Table 6.15:** Raw counts of the user preferences ( $N = 90$ ) are included for the most and least preferred *nudge* strategies for harm prevention and  $\chi^2$  analysis is performed to test for significant differences resulting in  $\chi^2(3, N = 90) = 59.47, p < .0001$

### 6.3.6 Transparent Strategy Specific

The results so far have been related to the hypotheses (**H1a** - **H1d**) aimed at comparisons of all 3 *nudge* strategies. We now turn to the final set of results which are related to hypotheses (**H2a** - **H2d**) and mainly specific to the transparent Stoplight strategy **S3**.

Results for hypotheses **H2a** and **H2b** are covered last, and begin with results related to the behavioural interactions related to the Stoplight colours (**H2c**).

#### 6.3.6.1 Compliance with Stoplights

The Stoplight search system made visible a warning about privacy risks. However, regardless of the search system (*nudge* strategy or Control), a hidden html tag for the light colouration was assigned to each result and recorded whenever a user visited the web page. This allowed for cross-comparisons of behaviour of the transparent Stoplight strategy with the Control as well as comparisons for the Re-ranking and Filtering strategies with the Control.

As stated in **H2c**, one should expect users to visit results with Green lights significantly more so than the Control system. Furthermore, behaviour should indicate a significant reduction in visits to results with Red, Yellow and Gray lights compared

to the Control as these were expected to be perceived as more risky (see analyses on perceptions in Section 6.3.6.2). Given the design of the two other *nudge* systems (see general methods Section 4.2.2), the same behaviour should also be expected for the Green light (related to **H1a**).

Using LMER and Kruskal-Wallace to test for differences, results included in Table 6.16 indicate that the Stoplight strategy significantly increases encounters with results tied to Green lights when compared to the Control, the same holds true for the Re-ranking *nudge*, with strong tendencies for the Filtering *nudge*. Analyses also indicate the Stoplight *nudge* significantly reduces encounters with Gray lights (those with privacy risks that are unknown). With respect to Re-ranking, the findings are significant across all colours, suggesting it outperforms the Control system at all times with respect to colouration. Interestingly, the Filtering approach significantly increases exposure to results with the html tag for Gray lights (Filtering removed any result above median number of *3rd party trackers* and therefore filtering results tagged with Red and Yellow).

DV	Control	Filter	Rank	Stoplight	Kruskal-Wallace (df = 3)	
	<i>M(SD)</i>	<i>M(SD)</i>	<i>M(SD)</i>	<i>M(SD)</i>	$\chi^2$	<i>p</i>
Assessments [Green Light]	1.11 (1.30)	1.39 (1.23) <sup>-</sup>	2.56 (1.74) <sup>***</sup>	1.51 (1.40) <sup>***</sup>	96.069	< .0001
Assessments [Gray Light]	0.74 (1.02)	1.16 (1.35) <sup>***</sup>	0.17 (0.57) <sup>***</sup>	0.50 (0.97) <sup>*</sup>	105.74	< .0001
Assessments [Yellow Light]	0.32 (0.56)	-	0.06 (0.23) <sup>***</sup>	0.34 (0.58)	55.264	< .0001
Assessments [Red Light]	0.36 (0.66)	-	0.01 (0.11) <sup>***</sup>	0.30 (0.55)	104.38	< .0001

**Table 6.16:** Stoplights - Included in the table are  $M = \text{mean}$  number of clicks and standard deviations ( $SD$ ) for each SERP. Significant Linear mixed effects (with LMER) results comparing *nudge* systems to the Control are indicated by <sup>-</sup> =  $p < .05$ , <sup>\*</sup> =  $p < .05$  (with Bonferroni correction), <sup>\*\*</sup> =  $p < .001$ , <sup>\*\*\*</sup> =  $p < .0001$ .

### 6.3.6.2 Perceptions of Uncertainty

A prediction of the experiment ([H2d](#)) was that users would perceive the Gray lights as more risky than the Green light and less risky than the Red light due to the underlying uncertainty with respect to privacy risk. During the experiment, a tooltip (Figure 4.2) informed users of the level of risk associated and that the Gray light (defined as an ‘unknown’ risk level).

To test the hypothesis ([H2d](#)), a contingency table (Table 6.17) including all four light values (Green, Yellow, Red and Gray) and the two aforementioned questions was built (see Section 6.2.3.3). Values were aggregated on the number of subjects selecting the colour response for each question (N.B. 1 participant did not provide a response). With this table,  $\chi^2$  analysis was performed across all light colours and the two risk factors. Results strongly indicate that users perceive the Red light as most risky and the Green light as least risky.

	Green	Yellow	Red	Gray
Most Risk	4	1	81	3
Least Risk	78	2	3	6

**Table 6.17:** A contingency table containing the number of user responses for each risk level and the perception of the light colour.  $\chi^2$  analysis was performed to test for differences.  $\chi^2(3, n = 89) = 140.54, p < .0001$   $n = 89$ , noting that one participant did not respond to the perception questions.



### 6.3.6.3 Privacy Attitudes, Awareness and Actions

A comparison of the metrics for privacy attitudes, actions and awareness was made against user interactive privacy metrics to test hypotheses **H2a** and **H2b**. The expectation was individuals reporting strong attitudes about privacy protection, taking strong actions to protect privacy and / or having strong awareness of privacy protection methods would utilize the warning light approach more strongly than individuals reporting weaker attitudes, actions and awareness. Self-report measures, as outlined in Sections 4.4.4.4 and 6.2.3.3, were used as independent variables in linear models to predict privacy measures (i.e. **Absolute Number of Trackers**, **Mean Number of Trackers** and **Normalized Number of Trackers**). In all cases, no significant links were found between the self-report measures and dependent privacy variables, therefore no results table is provided.

However, unlike the offline study, as an additional indicator of compliance (introduced in Section 4.4.2) with the transparent strategy, we performed additional analyses related to the behavioural interactions with the Stoplights. Analysis with Spearman rank correlation (Table 6.18) was used to test for differences related to compliance of the strategy based upon Stoplight colouration.

For the analyses, we aggregated the total number of clicks (denoted as **Sum** in Table 6.18) on results with Green lights in the Stoplight system for each participant in the experiment. Recall each SERP had two tasks assigned to it (Graeco-Latin Square design), which allowed us to calculate the average number of clicks on Green lights (**Avg** in Table 6.18). The analysis indicates that users taking actions to protect privacy in their daily lives tend to make use of the Stoplights (specifically visiting more web pages with Green lights) significantly more so than other users. There are also signals the same is true for users reporting stronger attitudes towards

## 6. INVESTIGATING NUDGES IN AN ONLINE SETTING

---

privacy and more frequent usage of privacy protective browsers. Similar analyses were performed for the other colours, but no significant differences were found.

Variable	Sum (cor)	Sum (p)	Avg (cor)	Avg (p)
Action Score (Binary Sum)	.23	.0266*	.21	.0454*
Action Score (Total)	.26	.0137*	.25	.0166*
Awareness Score	.12	.2495	.02	.8253
Enhancing Browser Score	.23	.0272*	.08	.4619
Privacy Attitudes (General)	.23	.0282*	.17	.1162

**Table 6.18:** Correlation (Spearman) analyses comparing user clicks on Green lights (sums[Sum] and averages[Avg]) in the Stoplight system with self reported metrics for actions, awareness and attitudes related to privacy. Correlation values (cor) and p-values (p) are provided with \* =  $p < .05$ .

## 6.4 Discussion

The online study had the overall aim of demonstrating robustness of the three harm prevention strategies (S1- S3) in a Web search environment that was more naturalistic than the earlier offline study (Chapter 5). Given the experimental environment was fully interactive (e.g. users could submit queries to a live index), a much broader and more in depth analysis compared to the offline study was completed, and additional hypotheses were tested (8 in total). Additionally, a higher level attention was placed on the transparent Stoplight *nudge* in the current study to better understand what scenarios such an approach might work or fail. The main findings of the study are presented below as they relate to the research questions (G-RQ-1 and G-RQ-2) and hypotheses.

### 6.4.1 Findings Related to Study Specific Hypotheses

A total of eight hypotheses were investigated the in current study, the first four (H1b - H1d) cut across all 3 *nudge* strategies and are given attention first. The remaining four hypotheses (H2a - H2d) specific to the transparent Stoplight strategy are then given focus.

#### 6.4.1.1 Across Strategies

Though significant findings were found across all *nudge* strategies with respect to the harm being prevented (due to loss of privacy) for the sample ( $N = 90$ ), post-hoc analyses indicate the Stoplight approach provides only weak signals of reduction in privacy impacts compared to the other two *nudge* strategies. Based on this finding, both H1a and H1b are confirmed. However, given the weak findings, caution must be used with the Stoplight strategy if one wishes to strongly reduce privacy

## 6. INVESTIGATING NUDGES IN AN ONLINE SETTING

---

impacts.

The same hypotheses were relevant to the analyses related to the compliance with Stoplight colour compared to the Control (see Section 6.3.6.1 where it is clear that all three strategies steer users to results identified as less risk with respect to privacy. However in this case, only **H1a** was confirmed. There is no evidence supporting **H1b** in this case as the transparent Stoplight strategy performs on par with the Filter strategy. This finding provides evidence that the Stoplight strategy is only weak with respect to the overall privacy measure, but not what it is trying to achieve (steering users towards less risky results i.e. those with Green lights).

Turning to **H1c** and **H1d**, a very in depth analysis was performed, first with respect to impacts to search behaviour (Section 6.3.3), followed by evaluations of the search systems (Section 6.3.4).

Most indications (results Table 6.6) for the general measures for impacts to search behaviour indicate that **H1c** is confirmed, that is all strategies perform on par with the Control. However, there are tendencies for the Re-ranking strategy to increase assessments of results, which is an indication that in general users are having more difficulty finding the information necessary to complete the search task. Though non-significant, the finding is an important consideration related to the viability **G-RQ-1** of the Re-ranking strategy.

The remaining analyses for **H1c** or **H1d** leveraged the annotations of results visited by the participants.

Analyses for the total assessments specific to each information type (*Correct*, *Incorrect* and *Harmful*) as a more detailed measure for impacts to search behaviour does not provide support for **H1c** or **H1d**. Though assessments of results with

*correct* information performs on par (or better for the Stoplight *nudge*), strategy [S2](#) (Re-ranking) significantly increases assessments of *incorrect* information and strategy [S1](#) (Filtering) significantly increases exposure to *Harmful* information.

Additional analyses were performed with system based IR metrics (MRR, Precision, Cumulative Probability) in Section 6.3.4 to provide a different lens on the proposed strategies. These results are partially mixed, but in general we find indications that the Stoplight strategy ([S3](#)) performs better than the Control system and the Re-ranking and Filtering approaches perform worse than the Control. These findings are surprising and certainly do not provide support for [H1d](#).

The findings with respect to [H1c](#) or [H1d](#) highlight the matter that trade-offs are a potential cost of some interventions, most notably the Filtering and Re-ranking strategies. That is, they are highly effective at preventing the harm for which they were designed, but potentially non-viable due to factors such as increased exposure of low-quality and potentially dangerous information related to the search task. This matter of non-viability and information quality is most-notable with the Filtering approach as this approach significantly increases exposure to *Harmful* and potentially dangerous information. Nonetheless, in our sample, there are no indications that these negative impacts with respect to misinformation resulted in negative impacts to the medical decision, thus positive search outcomes were still found with respect to harms due to loss of privacy and medical decision making.

#### 6.4.1.2 Transparent Stoplight Strategy

As results indicate, the transparent Stoplight *nudge* was most preferred by a large majority of participants. Furthermore, this approach is transparent in its goal (privacy protection) and therefore more ethically sound than the other two interventions. These points taken together make this approach a highly desirable

## 6. INVESTIGATING NUDGES IN AN ONLINE SETTING

---

*nudging* strategy for harm prevention. As such, one would expect the bulk of the users to take advantage of this strategy when compared to the Control, given the importance placed on personal privacy (e.g. responses to survey questions), however findings for the Stoplight *nudge* were not so cut and dry when compared to other findings.

Restating the previous findings, overall privacy impacts using the measures related to *3rd party trackers* demonstrates this approach is only marginally effective in the entire sample. However, as was performed with the offline studies, one must consider the possibility that certain individuals with greater concerns (and those taking actions) with respect to privacy on the Web and their uptake of privacy protections provided by the Stoplight strategy. From a reproducibility standpoint of testing the hypotheses (**H1a** and **H1b**) used in both the offline studies and current study, it is clear these measures do not support these hypotheses in the current study. Nevertheless, additional analysis were performed in the current study (Table 6.18) that does provide significant evidence supporting the hypotheses (in particular **H2b**). These findings indicate that overall our sample is potentially too small to detect the reduction in privacy impacts (using the chosen measures) for individuals that have a bona fide belief in privacy protection (as measured through their attitudes and in particular their actions in daily life).

However, additional analysis shows that users genuinely taking action in their daily lives to protect privacy (and to some extent having stronger attitudes towards privacy) are more compliant with the goal of the strategy. That is there is a tendency for them to visit results deemed least risky (Green lights in the Stoplight *nudge* are visited more so by these individuals). This finding provides support for **H2a** - **H2c**.

Transitioning the focus to **H2c** and **H2d**), inspection shows the transparent Stoplight *nudge* does significantly impact behaviour compared to the Control with respect to the colouration of lights (supporting **H2c**) and that users perceptions of the Gray light are seen as more risky than the Gray light and less risky than the Red light (confirming **H2d**). The significance of these findings being that it appears to cause a positive side effect (quite unexpected) that users are also pushed information that is higher quality with respect to the search task.

In summary, it is an interesting juxtaposition the Stoplight *nudge* is effective in pushing users towards less risky results, but does not significantly reduce privacy impacts related to *3rd party trackers*. There are multiple factors to consider that may explain these findings. First, there are many users that care about privacy and some that don't, therefore one would expect to see significant differences with 3rd party tracking impacts using a larger sample of participants, and therefore our sample is likely too small. Second, the other two interventions were much more drastic, and required users to take extra effort to be less privacy protective (e.g. they could have turned off the privacy settings, as introduced in Section 4.2.2), however no users turned off the settings.). Finally, there is the important factor of the actual number of 3rd party trackers tied to each result, for which the Stoplight may not provide enough granularity. Recalling the Green light was linked to the lower median of 3rd party trackers, the lower median still covers a broad range of values (approximately between 0 - 8 dependent on the search task and query results), values for which were not visible to the participants. Contrast this with the Re-ranking system in which the first result was always linked to the lowest number of *3rd party trackers*. Using this comparison, one can understand how the number of *3rd party trackers* would be less for Re-ranking, while the Stoplights do not produce such an effect. Taken together, this suggests that the Stoplight is not enough, and inclusion of the

absolute number of *3rd party trackers* be visually displayed as possible direction for future studies.

### 6.4.2 Findings in the Context of Research Questions

It is important to consider the findings in the context of the broader research questions ([G-RQ-1](#) and [G-RQ-2](#)). Table 6.19 summarises the key findings from the current online study. Given the overall aim of these strategies was privacy prevention, it is clear that all 3 strategies are effective at reducing impacts from *3rd party trackers*. However, as highlighted in the findings, the Stoplight strategy is only partially effective and therefore the least effective. Related to [G-RQ-2](#), this suggests that all 3 strategies are effective, just that strategies [S1](#) and [S2](#) are more so. However, there are very mixed findings with respect to the viability of strategies [S1](#) and [S2](#) (especially [S1](#)) which puts their viability into question. More discussion is provided specific to each of the high-level questions in the subsections that follow.



Evaluation		Filtering <a href="#">S1</a>	Ranking <a href="#">S2</a>	Stoplight <a href="#">S3</a>
Harm	Privacy	Reduced***	Reduced***	Reduced-
	Task Decision	On Par	On Par	Improved*
Behaviour	General Impacts	On Par	Worse-	On Par
	Information Quality	Worse*	Worse*	Improved-
IR System	MRR	Mixed	Mixed	On Par
	Precision	Worse	Worse	Better
	Probabilistic	Worse***	Worse*	Improved***
Preferences	Strategy	Least***	Between	Most***
Stoplights	Compliance	n/a	n/a	Yes***
	Perceptions	n/a	n/a	Colour Risks***
	Privacy Actions/Attitudes	n/a	n/a	Yes*

**Table 6.19:** A summary of key findings useful for evaluating the effectiveness and viability ( [G-RQ-1](#) and [G-RQ-2](#) ) of each strategy.  $- = p < .05$ ,  $* = p < .05$  (with Bonferroni correction),  $** = p < .001$ ,  $*** = p < .0001$  indicate the strongest findings within each factor.

#### 6.4.2.1 Effectiveness of Strategies [G-RQ-2](#)

From the position of proponents of *Nudging*, the strategies were all designed successfully. Re-ranking and Filtering were most effective at reducing privacy impacts, and the ‘Educational’ Stoplight *nudge* also provides some significant findings suggesting it will be effective at least for a part of the population. Furthermore, these systems were all true *nudges* in the sense the default was set to privacy protection and provided a low cost opt-out mechanism (the privacy settings switch introduced in Section 4.2.2). Proponents of *nudging* suggest that the opt-out mechanism will only be used by someone with a rational case for doing so and therefore should rarely ever be used [220].

There is little more to say on the topic of effectiveness of the *nudge* strategies, they certainly work, and in the case of *nudge* [S1](#) and [S2](#). Nonetheless, it does not address the problems surrounding the ethics of the approach, which is perhaps

hinted at through user responses on the preferred option (strategy [S3](#)) which is certainly the most ethical of the 3 approaches.

### 6.4.2.2 Viability of Strategies [G-RQ-1](#)

User preferences aside, which clearly demonstrate the Stoplight approach is most viable, the findings for [H1c](#) or [H1d](#) (summarised by the Behaviour [Information Quality] and IR System groups in [Table 6.19](#)) greatly put into question the viability of the Re-ranking and Filtering *nudge* strategies. Most notably this problem appears with the Filtering approach. However, the Stoplight approach appears to improve exposure to high quality information. It is worth considering plausible explanations for these findings and ultimately why we think the Filtering approach should not be used for harm prevention related to *3rd party trackers*.

The Filtering approach steers users towards results with unknown privacy risk (see [Table 6.16](#)), which is potentially a result of the data used for privacy linking. Recalling the [WhoTracks.me Test Set](#) contains the most popular 10,000 domains and was used for many of the results, and suggests that more popular websites have higher quality information. However, the same analysis (see [Table 6.16](#)) demonstrates the Re-ranking approach steers users towards results with a known (and lowest) privacy impact and away from results with both known and unknown privacy risks. Additionally, one can see that the Stoplight *nudge* also steers users towards results with a known and lowest privacy impact. Put together with the methods to rank results before being displayed to the user (see [Section 4.2.2](#), suggests the highest quality information are the results in the first 10 results of the original ranking (from the commercial API) in combination with results where privacy risks are already known (often the most popular websites). Given that 90 - 95% of all clicks (dependent on the search system) occurred within the first 10 results, it

is therefore not unreasonable that the Stoplight strategy (which keeps the original query ranking) steers users toward results with higher quality information.

Based upon these findings, we conclude that the Filtering strategy is a non-viable approach for harm prevention related to privacy, that caution must be used with the Re-ranking strategy, and that the Stoplight approach is definitely viable.

## 6.5 Summary

An online study was completed to compare three systems implementing *nudge* strategies (S1- S3). This study is an in depth extension of the earlier offline study (Chapter 5) and provides a much richer picture of the overall viability (G-RQ-1) and effectiveness (G-RQ-2) of these approaches to simultaneously reduce privacy impacts during Web search and have limited or no negative impacts to other aspects of Web search. In the online study, a live (real-time) search API was used. Connecting to the API provided a more naturalistic setting and richer experience with respect to choices of results. There was a trade-off of some control for the current study compared to the previous offline study, the methods used allowed for analysis that previous studies could not perform. The following key points were discovered through testing of hypotheses with the methods in an online environment.

- The Filtering strategy S1 is highly effective with respect to privacy protection compared to a Control system. Furthermore, this approach is least preferred (strongly so) by the participant sample when compared to the other two approaches. These findings allow us to conclude that this approach is non-viable as it too greatly increases exposure to misinformation and is not a desirable choice for most.
- The Re-ranking *nudge* based system (strategy S2) is similarly highly effective

tive with respect to reductions in privacy impacts compared to the Control. Like the Filtering strategy, there are also some negative impacts with respect to system performance as it relates to quality of information in the results visited by users. However, these negative impacts are much less noticeable compared to the Filtering result. Furthermore, the Re-ranking strategy does not negatively impact search outcome. We conclude this strategy is effective viable, but is also a strategy that should be used with caution due to our concerns raised around quality of information.

- Strategy [S3](#) is a transparent *nudge* that employs warning lights linked to the privacy harm the system aims to prevent. There is much weaker evidence to support the effectiveness of this strategy compared to the other two strategies. On the contrary, this system by far is the most viable, as it performs on par with (and in some cases even better than) the Control system with respect to important factors such as information quality of results assessed. It is also an approach that is more ethical than the other two strategies. In conclusion, we assess this approach as weakly effective and highly viable, and suggest that caution should be used if the aim is to *nudge* an entire population towards reduced privacy impacts.

With respect to the framework, several novel approaches were used in the evaluation ([FC-Evaluation](#)). For instance, in our study, we adapted precision @ k, a commonly used IR evaluation metric to assess systems based on relevance of documents, for evaluation based on annotations indicative of misinformation. We also demonstrated how one can still evaluate ‘educative’ strategies when the *search outcome* harm metrics (in this case privacy), where in this study flags hidden in the html of all systems (linked to Stoplight colour) allowed for comparison of interactive behaviour indicative of adherence across all strategies. These evaluation methods

and metrics will hopefully be a basis for development of new methods and metrics to evaluate Web search systems designed to reduce harm.

These findings conclude our analyses and discussion specific to the three *nudge* strategies (S1- S3) introduced in the general methods. The remainder of this thesis shifts focus towards the fourth strategy (strategy S4 tests *boosting*), which is based upon an entirely different philosophy to promote harm prevention and risk reduction. Whereas *nudging* promotes harm prevention through often times non-transparent methods (e.g. Filtering and Re-ranking) or ‘Educational’ transparent methods (e.g. Stoplights), *boosting* is designed to be entirely transparent and educational approaches. The removal of the quotes around educational *boosts* (vs. ‘Educational’ *nudges*) is not by accident, as *boosts* are designed to teach people a skill to reduce harm, whereas ‘Educational’ *nudges* do not teach a skill (putting into question their educational value).

In the context of Web search, and search systems, the current real challenge is identifying what skills might be taught as a means to reduce harms such as privacy. This provides the motivation for the empirical Chapter that follows and the subsequent Chapters aimed to compare *nudging* and *boosting* (G-RQ-3).



# Chapter 7

## Useful Cues for Harm Prevention

### 7.1 Overview

The four other empirical studies in this thesis give focus on the development of search systems ([FC-System](#)) that incorporate cognitive harm strategies ([FC-Cognitive](#)) and further introduce evaluation methods ([FC-Evaluation](#)) to assess impacts and quality of the Web search system. However, this study focuses on the methodology and development of informational cues, introduced in [3.2.3.1](#) if [FC-System](#), with the aim to use them in new system and cognitive based harm interventions. Specific to this thesis, identification of informational cues that have the potential for use across *nudging* and *boosting*, is an additional aim as it would permit a quantitative analysis related to high level research question ([G-RQ-3](#)).

Recall that *boosting* strategies are designed to empower individuals with skills necessary to reduce risk to themselves. Furthermore the skills should be applicable to an existing environment and therefore the environment should not need to be changed. That is, the individual learns a skill to better cope with risks in an existing

## 7. USEFUL CUES FOR HARM PREVENTION

---

environment and ultimately prevent and reduce harm. *Boosting* is therefore a stark contrast to *nudging*, which manipulates the environment to produce the desired outcome.

Returning to the research question aiming to compare these approaches ( **G-RQ-3** ) there are two key challenges open related to answering this question in the domain of Web search. The first challenge is to identify factors in the existing environment that can be translated to a skill that can be provided to individuals. The other challenge being the development of a *boost* that effectively ( **G-RQ-2** ) communicates the skill in a viable manner ( **G-RQ-1** ).

The current Chapter presents the study that addresses the first challenge. The findings from this study lay the foundation to address the second challenge (addressed in Chapter 8), which identifies the most effective and viable *boosting* strategy. With both challenges addressed, the research question ( **G-RQ-3** ) can then be addressed.

For consistency across all studies, this study also looks at harms related to privacy and misinformation. Additionally, the SERP and search system is in the spotlight, as the environment may enable users with a competency to protect privacy and / or reduce encounters with information. As such, the primary goal of the current study is determining what (if any) environmental cues within a SERP might be useful in helping users to reduce privacy impacts and / or exposure to misinformation.

In the sections that follow we provide background of the SERP and search environment, motivation for the informational cues considered and the formulation of hypotheses to be tested. After which methods are introduced, along with results of the experiments, and a subsequent discussion on the findings.



## 7.2 Background, Motivation and Hypotheses

It is a desirable goal to identify features that are ‘*easy-to-use*’ for the user as these are expected to be the most useful for comparison of *nudging* and *boosting* strategies (important for **G-RQ-2**). For the purposes of the current study, ‘*easy-to-use*’ features are defined as items that already exist in popular search engine environments (e.g. Google) and have the potential to be evaluated quickly by a user.

### 7.2.1 Motivation and Choice of Features

Learning-to-rank algorithms make use of web page surface features, which provide some insight for ‘easy-to-use’ features. Referring to learning-to-rank features provided by Liu in [139], the main feature families are based upon the web page URL, title, body, document as a whole and anchor text. These families are then represented in different manners such as language models and term counts. These feature families provide useful guidance for identification of ‘easy-to-use’ features.

Within modern SERPs (see examples in Figure 7.1), the main families are search result URL, title, snippet and rank. Referring to Figure 7.1, the top result (Arduino - HelloWorld) is at rank 1 and the title is coloured blue, URL is green and snippet text is black. Examples of ‘*easy-to-use*’ a user could evaluate with minimal effort include:

**Readability of the URL** URLs may contain unusual words auto-generated by spam websites.

**Language of the title** Does the title match the language of the query?

## 7. USEFUL CUES FOR HARM PREVENTION

---

[Arduino - HelloWorld](#)  
<https://www.arduino.cc/en/Tutorial/HelloWorld> -  
"Hello World!" The LiquidCrystal library allows you to control LCD displays that are compatible with the Hitachi HD44780 driver. There are many of them out there, and you can usually tell them by the 16-pin interface. This example sketch prints "Hello World!"

[The Hello World Collection](#)  
[helloworldcollection.de](http://helloworldcollection.de) -  
010 ! hello world in assembler for the hp-85 020 nam hello 030 def runtim 040 def tokens 050 def parse 060 def errmsg 070 def init 100 parse byt 0,0 110 runtim byt 0,0,377,377 120 tokens byt 377 130 errmsg byt 377 140 ! 150 init ldm r26,=msg 160 admd r26,=bintab 170 ldm r36,=12d,0 180 jsb =outstr 190 rtn 200 msg asc "hello world!"

(a) Bing

[C++ "Hello, World!" Program](#)  
[programiz.com/cpp-programming/examples/print-sentence](http://programiz.com/cpp-programming/examples/print-sentence)  
A simple C++ program to display "Hello, World!" on the screen. Since, it's a very simple program, it is often used to illustrate the syntax of a programming language.

[Total immersion, Serious fun! with Hello-World!](#)  
[www.hello-world.com](http://www.hello-world.com)  
Main index for [hello-world](#): links to login and all of the languages

(b) Qwant

**Figure 7.1:** Top results for the query "Hello World" in two commercial search engines (queries were sent from a computer in Germany in September 2019). These examples demonstrate how current SERPs assist users to identify websites that use https protocol (useful for data and privacy protection). Bing displays "https" in the address, unless encryption is not available in which case "http" is hidden. Qwant displays "www." for non-https websites. URL, title, and snippets are available for both examples, and both contain a ranked set of search results. The readability of the snippet for the Bing result at rank 2 is questionable.

**Readability of the snippet** Lower readability of a search result as an indication that the web page contains incorrect information.

**Depth (rank) of result** Encouraging users to consider a broader rank of results (e.g not just results at rank 1 and 2) may expose them to more correct information. Some evidence shows that user interventions for a deeper rank are a helpful method to expose users to more correct (higher quality) information [164]

While rank would be *'easy-to-use'*, encouraging users towards a deeper rank won't necessarily prove fruitful in ensuring privacy protection nor correct information as the economic costs in search [12, 15], such as time, will prevent many users from going beyond the first few results. Another, and quite problematic reason, is that ranking of results are in constant flux due to the automated nature of ranking

## 7.2 Background, Motivation and Hypotheses

---

algorithms, e.g. [139], thus it would be risky (and perhaps dangerous) to suggest to users to make use of such a feature. Thus, we do not consider ranking in our current study.

Due to the subjective nature of evaluating the language of the search result snippets and titles, these 2 families of features are likely to be less ‘*easy-to-use*’ and therefore not considered in the current study.

Thus, the focus of the research is with features in the URL family. However, users are not guaranteed to have a full URL visible for each result as many search engines truncate the visible URL in the search result. Unless the user were to spend extra time mousing over the hyperlink for each result, the user cannot fully assess the URL (e.g. total number of slashes, length of URL, readability of URL). Given time is an economic cost of the search process [12, 15, 177], it is unlikely users would mouse over the hyperlink. Nonetheless, there are three potentially ‘*easy-to-use*’ features common to modern SERPs in the displayed URL: a *cue for HTTPS* (see Figure 7.1), the *top level domain (TLD)* of the website and the *website domain*. Referring back to the first result in Figure 7.1, the result is an *HTTPS* web page, the *TLD* is ‘.cc’ and the *website domain* is ‘arduino.cc’.

**Website domain** The website domain may be an important cue for correctness of information (and potentially for privacy too), however there are millions of domains, making it difficult to find a useful rule of thumb (for this reason web domains are not tested in our research). The ever-growing number of online locations available in a user’s quest for information is one motivation for web domain bias [101] (i.e. sticking with websites the user knows).

## 7. USEFUL CUES FOR HARM PREVENTION

---

**Cue for HTTPS** While it is known HTTPS websites are more secure than sites with HTTP due to encryption capabilities<sup>1</sup>, very little research exists for other uses of this feature. While "HTTPS" presence in the URL is a useful predictor of phishing scams (e.g. [7]), to our knowledge this is a yet to be investigated with respect to other privacy impacts or misinformation. HTTP alone should not be ruled out, as websites using HTTP may contain web pages not requiring individuals to login, with content that is correct and relevant to the search task.

**Top level domain (TLD)** Using TLD for correctness of information arguably makes sense when considering some of the associations for websites in medical search. The Mayo Clinic, Wikipedia and Cochrane Review websites are all '.org'. The National Institute of Health (NIH) and Centers for Disease Control are both '.gov', where many high quality research findings are published. However, while it may seem obvious, no analysis has taken place to confirm that '.org' and '.gov' might be used as a simple heuristic for finding correct information.

TLDs are also potentially useful as an indicator for privacy impacts such as 3rd party tracking. As demonstrating example, consider the search results returned when running the query "cinnamon helps blood sugar" in Google coupled with data from a 3rd party tracking tool to determine the number of trackers<sup>2</sup>. In this example, compare the first 2 commercial medical websites webmd.com (rank 1 with 19 trackers) and medicalnewstoday.com (rank 2 with 21 trackers) with the first 2 non-profit medical websites mayoclinic.org (rank 4 with 2 trackers) and ncbi.nlm.nih.gov (rank 7 with 2 trackers).

Little literature exists investigating the seemingly simple TLD feature for pos-

---

<sup>1</sup>For instance, individuals should not login to a website with HTTP only as their data is susceptible to interception by an external party due to lack of encryption.

<sup>2</sup>The [Ghostery](#) browser plug-in was used for this example.

sible links to privacy impacts and correctness of information. One study included analysis of TLD links to privacy statements [87], finding that many ‘.org’ and ‘.com’ websites do not contain privacy statements. Another study considered TLD in models to predict web page spam [217], a study in the space of web page misinformation classification, for which TLD was a useful baseline feature. A third, and perhaps most relevant study, looked at binary occurrence of web page elements (such as javascript) frequently placed by 3rd parties [134], finding that ‘.com’ websites have the highest Boolean occurrence of such components. Combined together these studies are the most notable investigations into TLDs for privacy protection and reduced exposure to misinformation. To our knowledge, no one has considered TLD within the SERP itself as a possible means for better privacy protection and finding correct information.

### 7.2.2 Research Question and Hypotheses

Referring back to the examples in Figure 7.1, the TLD is available in both examples, as well as Google. HTTPS is a visible feature in Bing and Google, which are by far the most popular global search engines outside of China. Thus, the primary research question for this study is as follows.

- **FEATURE-RQ** “For search tasks, are HTTPS and TLD surface features indicative of A) personal privacy impacts B) quality of information and C) a predictor for both A and B?” .

**FEATURE-RQ** is used to formulate four hypotheses ( **H1** - **H4** ).

Typically .com websites are linked to commercial organizations and .org and .gov websites are linked to non-commercial organizations [212]<sup>3</sup>. The .net TLD is often

---

<sup>3</sup>The .org TLD was originally intended for non-profits only, however this is no longer a requirement.

## 7. USEFUL CUES FOR HARM PREVENTION

---

used and frequently associated with commercial companies as well [95]. 3rd party trackers are used for targeted advertising (a source of revenue) [149, 257], and many commercial websites (e.g. Google) have a strong profit motive to collect a searcher's personal data [267]. Based upon this:

**H1** Websites of commercial organizations (those with .com and .net) are more likely than the websites of non-commercial organizations (.org and .gov) to make use of 3rd party tracking, given the potential for increased revenue source.

As it is known that HTTPS is more secure than HTTP:

**H2** Greater privacy impacts (as measured by existence of 3rd party tracking) will be found for HTTP than HTTPS.

Previous research indicates high quality scientific articles are more strongly associated with correct information [164]. Additionally, anecdotal observations of annotations on the test sets used in our previous studies (**Waterloo Test Set**, **Offline Nudging Test Set** and **Online Nudging Test Set**) for the medical search tasks suggest that many scientific articles are associated with .org and .gov TLDs.

**H3** Thus, .org and .gov TLDs will be the TLDs most strongly associated with correct information for a given search task.

Given that .org and .gov TLDs are likely non-commercial TLDs:

**H4** This feature is a predictor of both privacy and correctness of information simultaneously.

---

The .gov TLD is enforced and organizations must also be part of a government.

## 7.3 Method

To test our hypotheses, we analyse web pages visited during interactions in a SERP for a set of search tasks as well as a broader set of websites popular across the Internet. We considered usage of a synthetic set of queries and results for our analysis, however we opted for analysing web pages visited during interactions in the earlier online study (Chapter 6) as the data was in our view was more naturalistic.

### 7.3.1 Evaluation Test Sets

The evaluation test sets used for the current study are the **Online Nudging Test Set** and **WhoTracks.me Test Set**, which are datasets output from the online lab based study (see Chapter 6 and general methods Section 4.3.4). The misinformation annotations for the current study are at a less granular level than those outlined in the general methods (Section 4.3.2). In the current study, annotations were for only two types of information *Correct* or *Incorrect*, as related to the search task for the given web page.

The definition of *Correct* maps directly to the definition in Section 4.3.2, where the definition of *Incorrect* is any of the other 3 possible annotation classifications in Section 4.3.2. This two class approach of annotations was used to be in line with annotations in **Waterloo Test Set** and allowed for a more straightforward analysis with respect to the quality of the annotations. The annotators classified each web page / search task pair (3 of the 523 unique web pages were linked to multiple tasks, for a total of 526 web page / search task pairs). In line with the general annotation methods, each annotator independently assessed all 526 pairs, then both annotators discussed any disagreements to converge on a final annotation of the web page / search task pair.

### 7.3.2 Evaluation Metrics

Four measures were used as dependent variables (DV). One metric is used to represent correctness of the search result (and associated web page), and three metrics are used for privacy.

#### 7.3.2.1 Web Page Misinformation

**Search Results - Correct** is defined as the total number of *correct* search results. A search result was *correct* if it contained information that agreed with the correct answer for the search task. A search result was marked *incorrect* if it had information that was conflicting or wrong. As our experiment was an online study with access to the live web, all websites were annotated after the experiment (see Section ?? for annotation methodology).

#### 7.3.2.2 Website Privacy

To be in alignment with privacy measures from the previous studies, the number of *3rd party trackers* linked to search results are used as an indicator for privacy risk. Additionally, two categorical variables are derived based upon the quartile values of 3rd party privacy trackers in results visited by users, providing a total of three privacy measures used in the analysis.

**Number of Trackers** The total number of trackers associated with a search result.

**Highest privacy** Based on the number of privacy trackers for all websites visited in the experiment. We define websites with the *Highest privacy* as those with the number of *3rd party trackers* in the lowest quartile of trackers for all results visited within the experiment.



**Good privacy** Calculated in a similar manner as *Highest privacy*, search results with *Good privacy* are defined as those with the number of 3rd party trackers in the lower median of trackers. This classification is applied to the same dataset used for the *Highest privacy* category as well, thus a website in the lowest quartile can be classified as *Highest privacy* and *Good privacy*, however a website with number of trackers between the median and lowest quartile would only be classified as *Good privacy*.

### 7.3.3 Statistical Tests

The independent variables (IV) in Table 7.1 were used for all analyses. Independent t-tests were used to test the effects of IVs on the number of 3rd party trackers, with Cohen’s  $d$  used as a measure of effect size. Logistic regression was used to test the effects of IVs as a predictor of categorical dependent variables (DV) outlined in Table 7.2.

SERP Surface Feature (IV)	Definition
Non-Profit	TLD is .org or .gov
Commercial	TLD is .com or .net
Other	All other TLDs (e.g. .uk, .de)
HTTPS	HTTPS in Web page URL

**Table 7.1:** Independent Variables (IV) - We test four different input variables, three based on the Top Level Domain (TLD) of the website domain and one on the availability of HTTPS encryption. We make the assumption that all .org and .gov websites are linked to not-for-profit organizations and that all .com and .net websites are linked to organizations having profit motives. Any remaining TLDs are placed in a catch-all *Other* category. For a full breakdown of websites and domains encountered for the TLDs, see Table 7.3.

## 7. USEFUL CUES FOR HARM PREVENTION

---

User Search Goal (DV)	Definition
Highest Privacy	# of 3rd Party Trackers in Lowest Quartile
Good Privacy	# of 3rd Party Trackers in Lower Median
Correct Information	Web page Information Correct for Search Task
Correct and Highest Privacy	Web page Correct and Trackers in Lowest Quartile
Correct and Good Privacy	Web page Correct and Trackers in Lower Median

**Table 7.2:** Dependent Variables (DV) which are categorical in nature and are assumed to be desirable user goals in a search task. That is, for some (if not all) users, it is desirable to maintain the highest privacy possible and / or find correct information when performing a search task online. These variables are used in the analysis covered in Section 7.4.3.

## 7.4 Results

We performed analyses to determine the effects of the surface features (Table 7.1) on the total number of 3rd party trackers, and the effects on the categorical dependent variables in Table 7.2. Section 7.4.2 provides results relevant to hypotheses **H1** and **H2**, Section 7.4.3 provides results relevant to all hypotheses **H1** - **H4**. As some of the analysis is underpinned by annotations of the web pages, we begin with an evaluation of the annotations produced for correctness of information within the web pages.

### 7.4.1 Web Page Annotations

Of the entire annotation set of 526 web page / search task pairs, 46 annotations overlapped with previously annotations in the **Waterloo Test Set** and the **Offline Nudging Test Set** (introduced in general methods Section 4.3). Cohen’s  $\kappa$  is used

as a reliability measures of annotations along with the interpretations provided by [151]. Annotator 1 had moderate reliability of  $\kappa = 0.78$ , ( $p < .001$ ), with previous annotations in the **Waterloo Test Set** and the **Offline Nudging Test Set**. Annotator 2 had moderate reliability of  $\kappa = 0.67$ , ( $p < .001$ ), with the same previous annotations. Annotator 1 had weak reliability of  $\kappa = 0.50$ , ( $p < .001$ ) with annotator 2. After discussing disagreements and producing final consensus, both annotators had perfect reliability of  $\kappa = 1.00$ , ( $p < .001$ ) with the 46 annotations in the **Waterloo Test Set** and the **Offline Nudging Test Set**, demonstrating the effectiveness of our annotation approach.

Of the 483 remaining annotations (web page / search task pairs) not included in the **Waterloo Test Set** and the **Offline Nudging Test Set**, annotator 1 had moderate reliability of  $\kappa = 0.67$ , ( $p < .001$ ) with annotator 2; for which both annotators jointly resolved any conflicts after completing their independent assessments. In total, 526 web page / search task pairs were annotated<sup>4</sup>.

## 7.4.2 Effects on Number of Trackers

Welch’s two-sided independent t-test was used to determine the effects on the total number of *3rd party trackers* encountered for each surface feature (**H1** and **H2**). We also calculate Cohen’s  $d$  as a measure of effect size. For the sake of comparison, analysis at the domain level of web pages visited was also performed (523 distinct web pages were visited in the experiment, of which there were 265 distinct web domains). Also included in the analyses are t-tests for the 10,000 most popular web sites provided by **WhoTracks.me** (used to create the **WhoTracks.me Test Set**).

---

<sup>4</sup>Annotations are available via links provided in <https://www.hrbdt.ac.uk/wp-content/uploads/2015/12/README.pdf>.

### 7.4.2.1 Non-Profit TLDs

*Non-profit* TLDs (i.e. web pages and domains ending in .org or .gov) are found to be a very strong predictor of the number of 3rd party trackers encountered compared with remaining TLDs. This finding is true for the web pages collected in our experiment as well as domain level and 10,000 domains [WhoTracks.me Test Set](#).

For the web pages visited in our experiment, we find the total number of 3rd party trackers linked to web pages with *Non-profit* TLDs ( $n = 146, M = 3.09, SD = 2.68$ ) are significantly different compared to web pages without *Non-profit* TLDs ( $n = 377, M = 7.16, SD = 5.21$ ), with a large positive effect  $t(481.68) = -11.68, p < .0001, d = 0.98$ . Using only the web domains for the web pages visited in our experiment, domains with *Non-profit* TLDs ( $n = 49, M = 4.49, SD = 3.28$ ) are also linked to a lower number of 3rd party trackers compared to domains without *Non-profit* TLDs ( $n = 216, M = 6.63, SD = 5.13$ ), and again find significant differences with a positive effect  $t(108.36) = -3.67, p = .0004, d = 0.5$ . Similarly, the analysis of the 10,000 [WhoTracks.me Test Set](#) domains with *Non-profit* TLDs ( $n = 409, M = 5.43, SD = 4.47$ ) compared to domains without *Non-profit* TLDs ( $n = 9591, M = 8.73, SD = 5.68$ ); a significant positive effect is found  $t(466.02) = -14.43, p < .0001, d = 0.64$ .

### 7.4.2.2 Commercial TLDs

*Commercial* TLDs, those ending with .com or .net, are also found to be predictive indicators of privacy impacts. Contrary to *Non-profit* TLDs, *Commercial* TLDs produce a negative effect on privacy. That is, visiting a website ending in .com or .net results in a higher number of 3rd party trackers being encountered over visits to TLDs not defined as *Commercial*.

For web pages visited during the experiment, web pages with TLDs that are *Commercial* TLDs ( $n = 255, M = 7.72, SD = 5.18$ ) contain significantly more trackers compared to remaining web page TLDs ( $n = 268, M = 4.40, SD = 4.21$ ); with a negative effect  $t(489.54) = 8.01, p < .0001, d = -0.70$ . Domains with *Commercial* TLDs ( $n = 149, M = 4.49, SD = 3.28$ ) also have more trackers compared to domains without *Commercial* TLDs ( $n = 116, M = 5.37, SD = 4.54$ ), and found to have a significant negative effect  $t(258.05) = 2.59, p = .0102, d = -0.32$ . The top 10,000 domains provided in the **WhoTracks.me Test Set** have a small negative effect for number of trackers connected to domains with *Commercial* TLDs ( $n = 5299, M = 8.92, SD = 5.90$ ) versus domains that are not *Commercial* TLDs ( $n = 4701, M = 8.22, SD = 5.37$ ) with  $t(9992.01) = 6.21, p < .0001, d = -0.12$ .

#### 7.4.2.3 Other TLDs

Analysis using *Other* TLDs (see definition in Table 7.1) was performed in the same manner as *Non-Profit* and *Commercial* TLDs. Of web pages visited in the experiment, 122 of 523 were classed as *Other* and for the domains in the experiment, 67 of 265 were of this class as well (see full breakdown of all TLDs in Table 7.3). For the **WhoTracks.me Test Set** domains, 4,292 of the 10,000 were defined as *Other*; with over 200 additional unique TLDs in the **WhoTracks.me Test Set** it was not possible to include these in Table 7.3. In all cases, no significant differences were found.

#### 7.4.2.4 HTTPS

Analysis of the HTTPS surface feature indicates web pages with HTTPS have a significantly higher number of trackers ( $n = 493, M = 6.15, SD = 5.00$ ) compared to web pages without HTTPS ( $n = 30, M = 3.83, SD = 4.36$ ), with a negative effect

## 7. USEFUL CUES FOR HARM PREVENTION

---

TLD Type	TLD	Web Pages	Domains
Non-Profit	gov	68	11
	org	78	38
Commercial	com	233	139
	net	22	10
Other	au	8	6
	ca	3	3
	edu	1	1
	ie	5	4
	info	3	2
	int	1	1
	scot	2	2
	uk	99	48

**Table 7.3:** Breakdown (by TLD Type and TLD) of total unique web pages (523) and total unique domains (265) visited by participants in our experiment. There is a particularly high number of .uk TLDs, likely due to localization in the Bing API settings.

$t(33.8) = 2.8, p = .0083, d = -0.49$ . Similar findings occur at the domain level, for domains with HTTPS ( $n = 238, M = 6.55, SD = 4.92$ ) compared to domains without HTTPS ( $n = 27, M = 3.48, SD = 3.91$ ) resulting in a significant negative effect  $t(36.05) = 3.76, p = .0006, d = -0.69$ . Finally, analysis of HTTPS presence with the 10,000 websites in the [WhoTracks.me Test Set](#) could not be performed as the information was not available. It is notable the number of web pages with and without HTTPS is highly imbalanced, due to the low numbers of websites that still use HTTP only and the manner in which HTTP and HTTPS are handled in the SERP (see Figure 7.1).

### 7.4.3 Effects on User Search Goals

Logistic regression was used for the analyses of the five categorical user search goals ([H1](#) - [H4](#)) outlined in Table 7.2, with full details included in Table 7.4. Contingency table values for the different classes are provided in Table 7.5.

IV (Surface Feature)	DV (Search Goal)	Coef.	S.E.	Wald (Z)	PR (>  Z )	OR
Non-Profit	Highest Privacy	2.2328	0.2254	9.90	< .0001	+9.33
Commercial	Highest Privacy	-1.3982	0.2226	-6.28	< .0001	-0.25
Other	Highest Privacy	-0.9971	0.2828	-3.53	.0004	-0.37
Non-Profit	Good Privacy	1.9144	0.2351	8.14	< .0001	+6.78
Commercial	Good Privacy	-1.6917	0.1903	-8.89	< .0001	-0.18
Other	Good Privacy	0.3192	0.2068	1.54	.1227	1.38
Non-Profit	Correct Information	0.7195	0.2029	3.55	.0004	+2.05
Commercial	Correct Information	0.0256	0.1751	0.15	.8838	1.03
Other	Correct Information	-0.8216	0.2104	-3.91	< .0001	-0.44
Non-Profit	Correct and Highest Privacy	2.0935	0.2808	7.45	< .0001	+8.11
Commercial	Correct and Highest Privacy	-1.5228	0.3127	-4.87	< .0001	-0.22
Other	Correct and Highest Privacy	-0.9913	0.3904	-2.54	.0111	-0.37
Non-Profit	Correct and Good Privacy	1.6617	0.2146	7.74	< .0001	+5.27
Commercial	Correct and Good Privacy	-1.2374	0.2160	-5.73	< .0001	-0.29
Other	Correct and Good Privacy	-0.4001	0.2468	-1.62	.1050	0.67
HTTPS	Highest Privacy	-1.1122	0.3798	-2.93	.0034	-0.33
HTTPS	Good Privacy	-1.4472	0.4652	-3.11	.0019	-0.24
HTTPS	Correct Information	0.757	0.3895	1.94	.0519	+2.13
HTTPS	Correct and Highest Privacy	0.3576	0.6224	0.57	.5656	1.43
HTTPS	Correct and Good Privacy	-0.3420	0.4005	-0.85	.3932	0.71

**Table 7.4:** Logistic Regression - TLD Type (IV) on User Search Goal (DV). We consider the three TLD types (Commercial, Non-Profit and Other) and HTTPS as defined in Table 7.1 as input variables. Five user search goals (outlined in Table 7.2) act as the dependent variable for each model. Odds ratios are provided as a signal for effects of the IV on the DV. A significantly positive odds ratio ( $OR$ ) is indicated by  $+$ , where as  $-$  indicates a significantly negative odds ratio to achieve the desired goal. As an example to interpret the results, consider the *Highest Privacy* goal, where the odds ratio is found to be significantly positive for *Non-Profit* TLDs and significantly negative for both *Commercial* and *Other* TLDs; therefore, if *Highest Privacy* is the goal in your search task, you should visit *Non-Profit* websites and should avoid *Commercial* and *Other* websites.

## 7. USEFUL CUES FOR HARM PREVENTION

IV (Surface Feature)	DV (Search Goal)	IV(+) DV(+)	IV(+) DV(-)	IV(-) DV(+)	IV(-) DV(-)	Total
Non-Profit	Highest Privacy	86	60	51	326	523
Commercial	Highest Privacy	34	221	103	165	523
Other	Highest Privacy	17	105	120	281	523
Non-Profit	Good Privacy	118	28	145	232	523
Commercial	Good Privacy	76	179	187	81	523
Other	Good Privacy	69	53	194	207	523
Non-Profit	Correct Information	98	49	187	192	526
Commercial	Correct Information	139	116	146	125	526
Other	Correct Information	48	76	237	165	526
Non-Profit	Correct and Highest Privacy	49	98	22	357	526
Commercial	Correct and Highest Privacy	14	241	57	214	526
Other	Correct and Highest Privacy	8	116	63	339	526
Non-Profit	Correct and Good Privacy	76	71	64	315	526
Commercial	Correct and Good Privacy	38	217	102	169	526
Other	Correct and Good Privacy	26	98	114	288	526
HTTPS	Highest Privacy	122	371	15	15	523
HTTPS	Good Privacy	239	254	24	6	523
HTTPS	Correct Information	274	222	11	19	526
HTTPS	Correct and Highest Privacy	68	428	3	27	526
HTTPS	Correct and Good Privacy	130	366	10	20	526

**Table 7.5:** Contingency Tables - Values here provide a breakdown of each IV and DV. The total number of web pages is 523 for the first 2 search goals (highest and good privacy). Due to 3 web pages being associated with multiple search tasks, 526 total web pages are associated with the 3 remaining search goals (Correct Information, Correct and Highest Privacy & Correct and Good Privacy). The (+) is used to indicate true and the (-) is used to indicate false. For example, for 86 of 523 web pages, it is true that the IV is *Non-Profit* and DV is *Highest Privacy*.



For all five user search goals, it is found that *Non-Profit* TLDs are the most likely search results to provide highest privacy and correct information. Thus, for users with the goal of maintaining *Highest Privacy*, *Good Privacy* and / or the goal of finding *correct* information, results suggest they should always stick with *Non-Profit* websites.

We find no statistical evidence suggesting the information associated with *Commercial* TLDs is more or less correct than information with non *Commercial* TLDs. Therefore we cannot conclude a user is worse off visiting .com and .net websites versus other websites. However, the results strongly indicate *Commercial* websites should be avoided by users concerned about privacy during their search tasks.

Search results linked to *Other* TLDs visited by users in our study are significantly more likely to contain *incorrect* information compared to TLDs not defined as *Other*, suggesting users (with the goal of finding *correct* information) should stick to search results with *Non-Profit* and *Commercial* TLDs. The same findings are also true for users with the goal of maintaining *Highest Privacy* in their search task; with non-significant findings for *Good Privacy*.

Also included in Table 7.4 is the analysis for websites using HTTPS. Although encryption of websites with HTTPS is more secure than HTTP, counter-intuitively search results with HTTP visited by users are more likely to offer privacy protection (based on privacy definitions in Section 7.3.2.2 and Table 7.2). We also perform exploratory analysis on links to correctness of information. Interestingly, though significance is not found with  $\alpha = .05$ , results suggest that correct information is more strongly associated with HTTPS than HTTP websites. Also noted is the skewed sample of search results visited with HTTP ( $n = 30$ ), compared to those with HTTPS ( $n = 493$ ). Regardless of the findings around the HTTP feature, we

suggest that search results with HTTP (especially those requiring login credentials) should be avoided, as they come with privacy risks due to their lack of encryption.

### 7.5 Discussion

The findings from the experiments demonstrate how two ‘*easy-to-use*’ elements common in search environments are useful in differentiating information that is high quality (*correct*) from low quality (*incorrect*) as well as information that is more or less impactful to one’s privacy.

While seemingly obvious that TLDs, such as .org and .gov, are indicative of search results with higher quality information and / or web pages which are more or less privacy impactful, to our knowledge, no one has performed analysis of this type. Our empirical evidence provided in Section 7.4 is therefore a benchmark for future research.

Additionally, TLDs treated as *Non-Profit* TLDs are found to be a very promising feature (as indicated by the results in Section 7.4) to assist users in finding information that is more likely to be correct and furthermore likely to protect personal privacy. Findings for TLDs treated as *Commercial* are congruent with our hypotheses about privacy impacts, that is for users visiting search results with these TLDs, their privacy is more likely to be reduced. It is a promising finding with respect to information quality for *Commercial* TLDs, that visiting search results from commercial websites with these TLDs does not significantly increase your likelihood of encountering incorrect information.

The analyses, in particular those considering the categorical measures of websites offering best odds of highest privacy and correct information suggest a possible link between third party tracking data and the quality of information on the website.

That is, the number of third party trackers is inversely linked to the correctness of information. A plausible explanation of such a link being the TLDs used in our analyses (.org and .gov) are associated with organisations that are both focused on publishing information with the least harmful outcomes (e.g. government funded health campaigns) and simultaneously not driven by profit of users of their website (e.g. *3rd party trackers* are used for display advertising<sup>5</sup>).

In summary, our findings strongly suggest that users should stick to search results with .org and .gov in the URL, for search tasks in the area of medical search (as that was the focus of our study). We suggest that TLD is an ‘easy-to-use’ feature that is not only simple to evaluate, but also can be evaluated quickly, two important factors when taking into consideration the economic factors of search [12, 15], and thus are a pathway for future work. For instance, user studies that consider approaches such as *Nudging* or *Boosting* (as outlined in Section 7.1) might consider TLD as a useful feature to compare both approaches. In the case of *Nudging*, the search system might be altered in a manner that increases the likelihood of encountering *Non-profit* TLDs. In the case of *Boosting*, a search system could highlight and explain to a user how they can alter their search behaviour (based on TLDs) to increase the likelihood of finding correct information, ultimately teaching the user how to better navigate search environments. The findings may also be useful as additional features for algorithms, such as learning-to-rank and those used for recommender systems.

### 7.5.1 Limitations

Our experimental findings demonstrate a pathway for increased likelihood of finding correct information but are not a guarantee, as demonstrated in recent find-

---

<sup>5</sup>For example compare the trackers and advertising on mayoclinic.org vs. healthline.com

## 7. USEFUL CUES FOR HARM PREVENTION

---

ings regarding user trust of website TLDs [35]. Furthermore, the results are only for 10 search tasks within the search domain of medicine. There are many other search domains beyond medicine (e.g. search for financial advice) with potential for very harmful search outcomes that should be investigated. We also only consider search tasks that are fact-based in nature, and have a clear answer. There are many different search tasks, task types and user search intents that come into play in the online environment. Search tasks such as those published by Wildemuth et al. [248] provide a set of other tasks to consider in future work.

We have made the assumption that *3rd party trackers* are a good proxy for privacy impacts for websites visited by users during the search process. There are other potentially nefarious privacy impacts not considered in the experiment (e.g. 1st party cookies, web beacons, browser finger printing, location data), for which 3rd party tracking is assumed to be strongly linked, however no analysis has been performed to confirm this. However, the process to test such a link may prove difficult, as it is not yet established what elements of a website are more or less impactful to one's own privacy. For future work, using a privacy statement corpus (e.g. [250]) is a suggested starting point to confirm or reject such a link.

With respect to the tools used for 3rd party trackers in our experiment, we utilized [WhoTracks.me](#) [117] 3rd party data for our analysis. This choice was made due to the willingness of the authors to provide data to academic researchers. We recognize that there are other tools for 3rd party tracking, such as [Privacy Badger](#), that might also be used. Analysis comparing the two might be one future area for exploration.

We also assume that *3rd party trackers* for each search result and web page are independent of one another, which is not always the case as demonstrated by

[117, 257]. There are multiple organizations (e.g. Google and Facebook) which have cross-platform trackers that are far-reaching, and thus not necessarily independent. A future study might consider these links, and we note that [WhoTracks.me](#) also provides the tracking company (e.g. Google Analytics) information which could be leveraged for users wanting to prevent companies from tracking their search behavior across platforms.

The analysis utilized the [Online Nudging Test Set](#), therefore our experimental setup did not consider search results not visited by users. A future study might annotate all of the search results not visited by users in the experiment.

Finally, we did not consider further breakdowns of the TLDs in Table ???. For instance, the '.uk' TLD could be broken down further into sub-domains and linked to equivalent definitions in Table 7.1, as websites hosted in the United Kingdom sometimes end in '.co.uk' or '.org.uk' which are analogous to '.com' and '.org', respectively.

### 7.5.2 Conclusions

We expose ourselves to risks that we may not be aware of throughout the online search process. Ranking algorithms and personal experience may reduce these risks, however many risks still exist due to the current landscape of the search environment. Searching for information is perhaps analogous to the process of searching for food. Unlike spoiled food, search results and websites presently do not give off smells, or turn brown, when they contain wrong information or take personal (potentially private) information.

Our findings on the analysis of 523 unique web pages visited in a user study indicate that sticking with .gov and .org TLDs is the best option when users care

about privacy and correct information. Commercial websites with .com and .net TLDs are important indicators when a user is concerned about correct information, but not their privacy. We suggest that our findings regarding the TLD features as predictors of correct and more privacy protective search results, are not only a benchmark for future researchers but more importantly will be useful in assisting individuals interested in strategies that help them find correct information more efficiently while simultaneously maintaining their privacy. These features might also be leveraged by designers of IR systems.

### 7.6 Summary

The overall goal of the study completed was to answer **FEATURE-RQ** and ultimately identify features promising for the research question related to the comparison of *boosting* and *nudging* (**G-RQ-3**).

Assuming our findings around the simple features are consistent predictors of reduced privacy impacts and higher quality information, we believe the TLD predictors coupled with appropriate methods, will foster opportunities for individuals to better navigate a somewhat risky and uncertain IR environment. The findings related to these features tie in directly to the behavioural and cognitive methods proposed in **FC-Cognitive**, along with system design considerations (**FC-System**) which include the potential for improvements to learning-to-rank algorithms.

The methods introduced here demonstrate how one might identify new informational cues based on log and metadata, which is yet another aspect of the system component (**FC-System**) of the framework. Based on the methods used and findings in the previous sections, we are confident that TLD is a quite promising feature to use for both harm prevention in the space of both privacy and misinformation.

The current findings provide direction in the next chapter and empirical study, which aims to identify the most effective and viable approach to present a *boost* strategy.

## 7. USEFUL CUES FOR HARM PREVENTION

---



# Chapter 8

## Identifying Effective Boosts

### 8.1 Overview

The following sections outline the methods used and findings from a pilot study to evaluate (FC-Evaluation) different variants of a harm prevention strategy known as *boosting*, an intervention from the cognitive sciences (within FC-Cognitive) designed to empower individuals with skills for improved decision-making. In the current study several variants of a fact box *boost* (strategy S4) are designed and evaluated for learning a competency for reduced privacy impacts during Web search. The target competency, if adhered to, should reduce data sharing with 3rd party companies during the process of Web search.

The primary aim of this study is the identification of effective and viable *boosting* approaches (G-RQ-1 and G-RQ-2) for use in the subsequent empirical study (see Chapter 9) which compares *boosting* and *nudging* (G-RQ-3). In this study, an abstract simulation of a search system (FC-System) is used as part of the evaluation to determine which strategies are most likely to perform best in an interactive search

environment.

### 8.2 Motivation and Hypotheses

Both *boosting* and *nudging* interventions target the decision-making process with the goal of reduced risk and harm. Nonetheless, it is worth a recap of some of the key differences between a *boost* and a *nudge*, because they are quite different. We assume (based on the introduction to behavioural and cognitive interventions in Chapter 2) the distinction between *boosts* and ‘classic’ *nudges* is quite clear, therefore we focus on the transparent ‘educational’ *nudge* for this recap, as this flavour of intervention may cause confusion as it sometimes look similar to a *boost* on the surface.


Recall that ‘educational’ *nudges* (e.g. Stoplight strategy in earlier studies) are suggested as an alternative that offers transparency (and thus more ethically sound) over non-transparent ‘classic’ *nudges* (e.g. Filtering and Re-ranking systems in earlier studies). Additionally, though an ‘educational’ *nudge* may produce less risky decision making, it is not designed to enable users with a competency for better decision making, and therefore an ‘educative’ *nudge* (like a ‘classic’ *nudge*) is unlikely to be effective once the removed. This second point is referred to as ‘reversibility’.

A *boost*, on the other hand, is designed to empower users with a competency to evaluate an environment for risks and / or uncertainty and therefore make more informed decisions after the *boost* is removed (decision making is not reversed). Furthermore, a *boost* is always transparent, whereas aside from the ‘educational’ variety, *nudge* interventions are predominantly non-transparent about their goal.

However, the cost of a *boost* is that it requires effort on the individual, which often requires some motivation, a trait which applies to ‘educative’ *nudges* as well.

**Example comparing ‘educational’ *nudges*** Consider two variants of an ‘educational’ *nudge* for harm prevention in Web search: one variant is the Stoplight *nudge* (see Figure 4.2) evaluated in our earlier studies and the other variant is information nutrition label (see Figure 8.1) as proposed by Fuhr et al. [74]. Both provide informational cues to assist a searcher with in evaluating potential risk with respect to the information they encounter. However, both do not provide any further information, they suggest danger but they do not indicate what outcomes are possible based on the actions the searcher can take (e.g. visit result 1 or result 3). While it is possible that an individual with time may learn what types of information to avoid with these interventions, learning would be purely a side effect of their design (as opposed to a *boost* where for decision making is a primary goal). Also, neither approach gives any indication of what to do when using a system where such interventions are unavailable (a potential ethical concern).

INFORMATION NUTRITION LABEL		
Best Before: Jan 1, 2018		
Per 1000 words	Recommended Daily Allowance	
<u>Fact</u>	30%	60 %
<u>Opinion</u>	40%	20 %
<u>Controversy</u>	9.0	--
<u>Emotion</u>	6.7	1.3
<u>Topicality</u>	8.7	5.0
<u>Reading Level</u>	4.0	8.0
<u>Technicality</u>	2.0	--
<u>Authority</u>	4.3	9.0
<u>Viralness</u>	--	1.0
Additional substances: advertising, subscription, invective, images (2), tweets, video clips		
Traces: product placement		



**Figure 8.1: Proposed Information Nutrition Label by Fuhr et al. [74]** - The label includes different dimensions (e.g. opinion) to make the searcher aware of the potential risks associated with too much consumption of information high in these dimensions.

## 8. IDENTIFYING EFFECTIVE BOOSTS

---

This recap and examples were used to highlight shortcomings of *nudging* introduced in background Section 2.6 and provide further motivation to investigate *boosting* as an alternative cognitive strategy.

### 8.2.1 Motivation for Fact Boxes

The case has been made for the evaluation of *boosting* as an alternative intervention to *nudging*, but there remains an open question of what type of *boost* to use.

Based upon existing theoretical and empirical research, *boosting* is made up of an extensive set of interventions that can broadly be classed into three types (risk literacy, uncertainty management and motivational) [93] (see examples at [Boosting](#)<sup>1</sup>). Risk literacy is the area we focus upon, as risk literacy is best used for problems where risk is already known (as established in the previous study in Chapter 7).

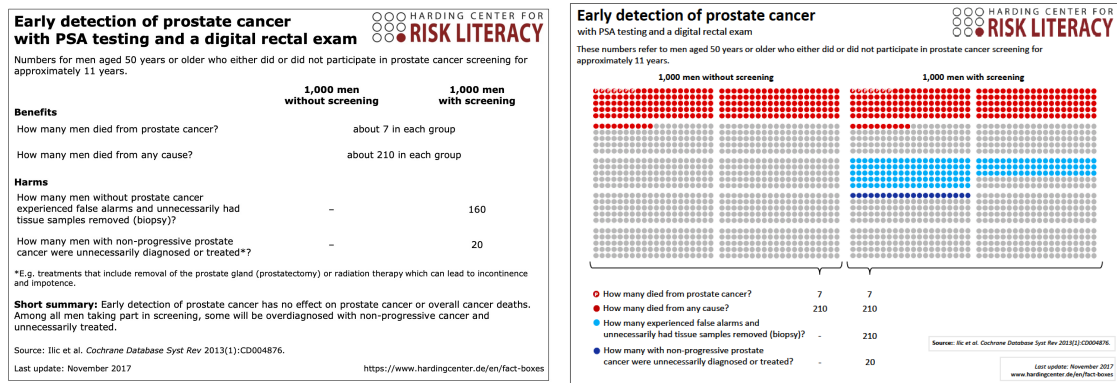
There are a few features seen as necessary for a *boost* intervention in a Web search environment.

- They should be easy to understand (to account for all aptitudes)
- They should be understood quickly (because people don't have much time)
- Displayed in a concise graphical manner (so that they can be embedded in browsers with limited impacts on the overall user experience)

Within the *boosting* literature, there are two approaches which stand out fitting these criteria. Fast and Frugal Trees (FFTs) [82], which are simple decision trees modelled on data linked to outcomes (e.g. heart attack) and independent variables

---

<sup>1</sup>[Boosting](https://scienceofboosting.org/) at <https://scienceofboosting.org/> (LA: 2020-10-01)



**Figure 8.2:** Two Fact Boxes designed by the [Harding Center for Risk Literacy](#), which communicate the same risks graphically but in a different manner. The risks being communicated are those related to screening for Prostate Cancer and the message is the same in both Fact Boxes (outcomes are worse for those screened).

tied to those outcomes (e.g. blood pressure), are one approach. The other approach being Fact Boxes [150], which are graphical mechanisms used to understand risks and possible benefits and harms of different choices (see examples in Figure 8.2).

We rule out FFTs for our studies for two reasons. First, as indicated (see [Boosting](#)), FFTs are more appropriate for environments where uncertainty is a larger concern (not the case in our studies as they are controlled lab settings). Second, comparing the two approaches, it is our view they are less intuitive for someone unfamiliar with decision trees (and therefore less easy to understand and likely require more effort and time).

## 8.2.2 Hypotheses

Based on the method of fact boxes as the *boost* approach chosen for the current study, and the multiple nuances one might consider (e.g. how to best present a fact box in a search environment), the initial aim is to identify the approach(es) which produce the best learning effect. Learning, and related matters such as memory, is an entire area of research much too large to investigate here. However, it must

## 8. IDENTIFYING EFFECTIVE BOOSTS

---

not be overlooked that cognitive effort is required by the individual for a *boost* to be effective [93], and therefore one can conclude that more exposure permits opportunity for cognitive effort.

Combined together, we formulate the following three hypotheses:

**H1** Learning of risks communicated in the fact box will be significantly better for individuals treated with fact boxes compared to those who are not.

**H2** Learning of risks communicated in the fact box will be closer to truth for individuals treated with more exposure than individuals with less exposure.

**H3** Using time as a measure of cognitive effort, it is expected that more time spent with the fact boxes will translate to better learning.

### 8.3 Method

Methods used to evaluate these hypotheses and results of the evaluation are presented below.

#### 8.3.1 Procedure

The main procedures novel to this study are the various designs of the fact box and overall study design used for evaluation.

##### 8.3.1.1 Fact Box Design

Combining findings from the results in Chapter 7 and fact box methods to boost individuals with skills for harm reduction [150], we developed two fact boxes (Figures

8.3 and 8.4) designed to enable users with a skill to reduce the amount of data they share with 3rd party companies.

Returning to the **Harm Prevention Features Test Set** developed in the previous study (Chapter 7), the web pages were grouped by quartiles of the number associated *3rd party trackers*. For instance, in Table 7.2 web pages with ‘Highest Privacy’ are those falling in the lowest quartile of *3rd party trackers* (less than or equal to 2 trackers). Though not included in Table 7.2, the upper quartile of web pages is defined as those having 8 or more *3rd party trackers*. In the fact boxes designed for the current pilot study (Figures 8.3 and 8.4), the lowest quartile of trackers were mapped to **benefits** and the upper quartile mapped to **harms**.

To ensure consistent dimensions for the user, the findings from the previous study were normalized to be out of 100 websites. For example, for .org / .gov TLDs in Table 7.3 a total of 86 .org / .gov websites contained 2 or fewer *3rd party trackers* and 60 containing more than 2 trackers. Setting  $\frac{x}{100} = \frac{86}{60+86}$ , we solve for  $x$  and round it to the nearest whole number and set this value to equal the number of .org / .gov websites out of 100 websites that share data (i.e.  $\frac{59}{100}$  of 100 websites). This same process is repeated for the 3 remaining fact box cells.

The larger fact box (Figure 8.3) is closely aligned with the recommended approaches (see [150]), including a definition of the potential **harms** (as motivation to learn the skill) and citations to the data used for the fact box (for credibility). The larger fact box was the starting point to design the smaller fact box (Figure 8.4). The large fact box was included in our study, as it is in line with fact box design that is shown to be effective (again see [150]). We recognize that such a fact box is unlikely to fit within an operational search engine, which therefore motivated the design of a smaller fact box (see Figure 8.4) that hypothetically can be included in

## 8. IDENTIFYING EFFECTIVE BOOSTS

<b>.org &amp; .gov websites for privacy &amp; health</b>		
<p><b>When using a search engine for medical treatments for health issues, it is shown that visiting websites with top level domains (TLD) ending in .org or .gov increases the likelihood that:</b></p> <ul style="list-style-type: none"> <li>Your personal data will not be shared with 3rd parties (other companies)</li> </ul>		
<p>Numbers are averages based on websites visited by people using a search engine to find information about medical treatments for health issues.</p>		
	<b>100 websites visited with TLD ending in .gov or .org</b>	<b>100 websites visited ending in other TLDs (e.g. .com, .net, .org.uk)</b>
<b>Benefits</b>		
How many websites will share/sell your information to 2 or less 3rd party companies (e.g. Facebook)?	59	14
<b>Harms</b>		
How many websites will share/sell your information to 8 or more 3rd party companies (e.g. Facebook)?	11	33
How can 3rd party data be used to harm you?	3rd party data collected from websites you visit can be used to send you advertisements for medical products potentially harmful to your health.	
<p>Sources: [1] Towards Search Strategies for Better Privacy and Information, October 2019 [2] Privacy Implications of Health Information Seeking on the Web, ACM Communications, 2015</p>		
<p>Date last updated: October, 2019</p>		

**Figure 8.3: Large fact box to *boost* individuals with a skill to reduce privacy impacts.** - A large fact box allowing users to compare the benefits and harms of different TLDs. Much more detail is provided in the large fact box compared to the small fact box (Fig. 8.4).

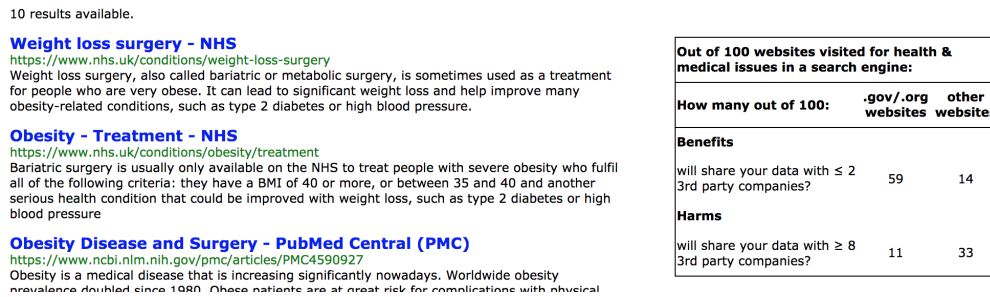
<b>Out of 100 websites visited for health &amp; medical issues in a search engine:</b>		
<b>How many out of 100:</b>	<b>.gov/.org websites</b>	<b>other websites</b>
<b>Benefits</b>		
will share your data with $\leq 2$ 3rd party companies?	59	14
<b>Harms</b>		
will share your data with $\geq 8$ 3rd party companies?	11	33

**Figure 8.4: Small fact box to *boost* individuals with a skill to reduce privacy impacts.** - A small fact box allowing users to compare the benefits and harms of different TLDs.



a modern SERP.

We conclude this introduction to the fact box designs for the current study with a prototype demonstrating how this might look when implemented in a search engine (see Figure 8.5).



**Figure 8.5: Prototype SERP to *boost* users with knowledge and skills to protect personal privacy during Web search.** - The small fact box (Figure 8.4) is placed in the right hand rail of the results.

### 8.3.1.2 Study Design

Figure 8.6 provides an overview of the study design for the four experimental groups (including a control group). All groups completed the same questions and tasks. As a means to provide a baseline comparison against the *boost* interventions under evaluation, the control group was not provided a fact box. The **inline only** group was given the small fact box (Figure 8.4) during task 1 only. The two remaining groups, **large before** and **small before** respectively had the large fact box (Figure 8.3) and small fact box (Figure 8.4) presented during the pre-task instructions. These two groups also had the small fact box available during task 1.

Simulated search tasks (Cochrane Medical Task 2 & 4 from Table 4.1) were used in our pilot study to test the fact boxes. They are simulated because users had no search results available, and only a small selection of URLs to choose from.

For Task 1, the instructions and questions provided in Table 8.1. Aside from

## 8. IDENTIFYING EFFECTIVE BOOSTS

Experimental Groups	Definition	Experiment Progression			
		Pre-Task	Task 1	Task 2	Post Task
<b>Control</b>	No factbox provided at any point in experiment.	Instructions Only	Questions only	Questions Only	Questions Only
<b>Inline Only</b>	Small Factbox provided only during task 1		Small Factbox + Questions		
<b>Small Before</b>	Small Factbox given before task 1 + during task 1				
<b>Large Before</b>	Small Factbox given before task 1 + during task 2				

**Figure 8.6: Overview of the study design for piloting a fact box for boosting skills for better privacy.** - All groups completed 2 mock search tasks. The control group had no fact box during the experiment. The inline group had the small fact box (Fig. 8.4) visible during task 1, as did the remaining two groups. The small and large fact box groups were each introduced to the small fact box 8.4 and large fact box 8.3 respectively in the pre-task instructions. In the second task, no fact box was available across all groups. Post task questions asked participants to estimate the number of websites that shared information with 3rd party companies.

the participants in the control group, the small fact box was visible for all other participants during this task. Participants were given 5 web page URLs to choose from for each question, they were only allowed to choose one URL, and only one URL was the correct answer (e.g. a single web page ending in .org or .gov would be the **least likely** to share information). The questions and answers were randomized. Provided below are the instructions and questions in full for Task 1.

Task 2, which is provided in Table 8.2, asked the same questions (i.e. **least likely** and **most likely** to share information), but used a different medical task (**Do benzodiazepines help alcohol withdrawal?**) and therefore had different URLs provided in as options to choose from. As a reminder, for Task 2, no fact box was available for any of the participants, which was an important element for evaluating the learning effects (**H1** - **H3**).

After completing Task 2, participants were then asked to estimate values (see

task in Table 8.3) corresponding to each cell in the fact box. For the estimation task, no fact box was available for any of the participants (again an important mechanism for evaluating the learning effect). Provided below are the four questions which map the four cells in the fact box. The questions were randomized. Participants could only give answers between 0 and 100.

The entire study was hosted on the Qualtrics online survey platform and questions associated with Task 1, Task 2 and the estimation task, were presented in a randomized order.

## 8. IDENTIFYING EFFECTIVE BOOSTS

---

---

### Instructions

---

Please read the text below carefully and imagine that the situation described is real.

You use your favourite search engine (e.g. Google) to find websites to help you answer:

**Do benzodiazepines help alcohol withdrawal?**

Out of 100 websites visited for health & medical issues in a search engine:		
How many out of 100:	.gov/.org websites	other websites
<b>Benefits</b>		
will share your data with $\leq 2$ 3rd party companies?	59	14
<b>Harms</b>		
will share your data with $\geq 8$ 3rd party companies?	11	33

The following questions include website URLs from the results provided by the search engine you have used. Please answer each question:

---

Which of the following websites is **least likely** to share your information with 3rd party companies?

[https://www.who.int/mental\\_health/mhgap/evidence/alcohol/q2/en/](https://www.who.int/mental_health/mhgap/evidence/alcohol/q2/en/)

<https://www.ukat.co.uk/benzodiazepines/withdrawal-detox/>

<https://www.drugs.com/article/benzodiazepines.html>

[https://www.researchgate.net/publication/42256367\\_Benzodiazepines\\_for\\_alcohol\\_withdrawal](https://www.researchgate.net/publication/42256367_Benzodiazepines_for_alcohol_withdrawal)

<https://americanaddictioncenters.org/withdrawal-timelines-treatments/alcohol-benzos>

---

Which of the following websites is **most likely** to share your information with 3rd party companies?

<https://www.alcoholrehabguide.org/treatment/benzodiazepines/>

<https://www.psychologytoday.com/us/blog/all-about-addiction/201205/treating-alcohol>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4606320/>

<https://www.cochrane.org/CD005063>

<https://pubs.niaaa.nih.gov/publications/arh22-1/38-43.pdf>

---

**Table 8.1: Task 1:** The first task and multiple choice questions asked to test knowledge learned about TLDs and privacy. The fact box was visible for all participants (except Control group) during this question. Correct answers are highlighted.

### Task 1 & 2 (Multiple Choice Questions)

---

**Instructions**

---

Please read the text below carefully and imagine that the situation described is real.

You use your favourite search engine (e.g. Google) to find websites to help you answer:

**Do probiotics help treat eczema?**

The following questions include website URLs from the results provided by the search engine you have used. Please answer each question:

---

---

Which of the following websites is **least likely** to share your information with 3rd party companies?

[http://applications.emro.who.int/imemrf/Iran\\_J\\_Pediatr/Iran\\_J\\_Pediatr\\_2011\\_21\\_2\\_225\\_230](http://applications.emro.who.int/imemrf/Iran_J_Pediatr/Iran_J_Pediatr_2011_21_2_225_230)

<https://www.dailymail.co.uk/health/article-4899194/Seven-steps-rid-eczema.html>

<https://www.drugs.com/npp/probiotics.html>

[https://www.researchgate.net/publication/26336555\\_Probiotics\\_for\\_the\\_treatment\\_of\\_eczema](https://www.researchgate.net/publication/26336555_Probiotics_for_the_treatment_of_eczema)

<https://nationaleczema.org/search-bacterial-balance/>

---

Which of the following websites is **most likely** to share your information with 3rd party companies?

[https://www.cochrane.org/CD006135/SKIN\\_probiotics-treating-eczema](https://www.cochrane.org/CD006135/SKIN_probiotics-treating-eczema)

<https://www.healthline.com/health/skin-disorders/probiotics-for-eczema>

<https://www.worldallergy.org/ask-the-expert/answers/probiotics-in-the-treatment-of-atopic>

<https://www.ncbi.nlm.nih.gov/pubmed/30480774>

<https://clinicaltrials.gov/ct2/show/NCT03863418>

---

**Table 8.2: Task 2:** The second task and multiple choice questions asked to test knowledge learned about TLDs and privacy. The fact box was not visible for any participant during this question. Correct answers are **highlighted**.

---

**Instructions**

---

Please answer the following questions related to websites you might visit while searching for health and medical information.

---

---

**Questions**

Out of 100 **.org** websites, how many websites do you estimate will share your information with **2 or less 3rd** party companies?

Please enter a value between 0 - 100.

Out of 100 **.org** websites, how many websites do you estimate will share your information with **8 or more 3rd** party companies?

Please enter a value between 0 - 100.

Out of 100 **.com** websites, how many websites do you estimate will share your information with **2 or less 3rd** party companies?

Please enter a value between 0 - 100.

Out of 100 **.com** websites, how many websites do you estimate will share your information with **8 or more 3rd** party companies?

Please enter a value between 0 - 100.

---

**Table 8.3: Estimation Task:** Questions asked for the estimation task to test for knowledge gained from the fact box *boost*. Correct answers are found in the fact box, which was not visible to participants for this task.

### 8.3.2 Evaluation Metrics

The following three evaluation metrics were used for analysis on the main tasks questions. All participant responses to the task questions were converted to boolean

## 8. IDENTIFYING EFFECTIVE BOOSTS

---

values.

**All Task Questions Correct**  $Score = 1$  if all 4 multiple choice questions were correct for task 1 and 2, otherwise  $Score = 0$ .

**Task 1 Questions Correct**  $Score = 1$  if both multiple choice questions were correct for task 1, otherwise  $Score = 0$ .

**Task 2 Questions Correct**  $Score = 1$  if both multiple choice questions were correct for task 2, otherwise  $Score = 0$ .

For the post-task estimation questions, the following evaluation metrics were used as dependent variables.

**Total Absolute Deviation of Fact Box Estimate** =  $\sum_{n=1}^N |PV_n - AV_n|$  Where  $PV_n$  = participant estimated value of entry  $n$  in fact box.  $AV_n$  = actual value of entry  $n$  in fact box.  $N$  = the total number of entries in fact box.

### 8.3.3 Statistical Tests

Chi-Square was used to test the overall effect group differences for multiple choice questions. For any significant findings, post-hoc analysis was performed with logistic regression to determine odds-ratios.

One-way Anova was used to determine any significant group differences for the estimation tasks. For any significant findings, post-hoc comparisons between the experimental variant and the control were performed with Welch's two-sample t-test.

### 8.3.4 Participants

A total of 209 participants took part in the experiment and were assigned at random to the 4 experimental groups. Participants were recruited via the [Prolific](#)<sup>2</sup> recruitment platform, which is a Crowdsourcing platform similar to Amazon Mechanical Turk (Prolific is much easier to setup and the company claims to have higher quality samples.). As one goal of the pilot study was to inform design of the United Kingdom based lab study outlined in Chapter 9, the Prolific participants were limited to the same geographic region. The participants reported a mean age ( $M = 39.1, SD = 11.8$ ) of which  $n = 130$  were female,  $n = 78$  were male and  $n = 1$  as other.  $n = 145$  participants reported attainment of an undergraduate degree or higher, and  $n = 191$  were native English speakers. No significant differences were found with distribution of these demographics across the 4 experimental groups.

Participants were required to provide consent for the experiment before taking part in the tasks. Based on a pre-experiment estimation of 6 minutes for the entire experiment all participants were paid £.50 for their time<sup>3</sup>. The median experiment time was 5 minutes and 20 seconds.

## 8.4 Results

### 8.4.1 Task Questions

Using the boolean evaluation metrics as dependent variables and the experimental groups as independent variables, a Chi-Square test was performed to test for overall group differences (see Table 8.4). In all cases, group differences were highly significant.

---

<sup>2</sup>[Prolific](https://www.prolific.co/) at <https://www.prolific.co/> (LA: 2020-03-04)

<sup>3</sup>This payment is equivalent to the Prolific minimum requirement £5 per hour.

## 8. IDENTIFYING EFFECTIVE BOOSTS

We then used logistic regression to perform post-hoc analyses between each group and the control group (see Table 8.5) and calculated the between group odds ratios as a measure of effect. Again, all differences were found to be significant, with the large fact box group control group comparison consistently producing the strongest odds ratios.

		Control	Inline Only	Small Before	Large Before	$\chi^2(df = 3)$	$p$
Sample	N	49	54	52	54	-	-
All Task Questions Correct	Yes	2	20	18	34	39.03	<.0001
	No	47	34	34	20		
Task 1 Questions Correct	Yes	4	20	18	35	35.591	<.0001
	No	45	34	34	19		
Task 2 Questions Correct	Yes	7	22	22	36	29.036	<.0001
	No	42	32	30	18		

**Table 8.4:** Chi Squared Analysis was performed on the 3 boolean metrics for the task questions to compare differences between the 3 experimental groups and the control group.

Dependent Variable	Control vs.	Coef.	S.E.	Wald (Z)	PR (>  Z )	OR
All Task Questions Correct	Inline Only	2.6264	0.7750	3.39	.0007	13.82
	Small Before	2.5210	0.7786	3.24	.0012	12.44
	Large Before	3.6876	0.7750	4.76	<.0001	39.95
Task 1 Questions Correct	Inline Only	1.8897	0.5930	3.19	.0014	6.62
	Small Before	1.7844	0.5977	2.99	.0028	5.96
	Large Before	3.0313	0.5945	5.10	<.0001	20.72
Task 2 Questions Correct	Inline Only	1.4171	0.4933	2.87	.0041	4.12
	Small Before	1.4816	0.4954	2.99	.0028	4.40
	Large Before	2.4849	0.5000	4.97	<.0001	12.00

**Table 8.5:** Post-hoc analysis was performed on all 3 boolean metrics for the task questions. Logistic regression is used to calculate the odds ratio (as a measure of effect) for each experimental group against the control group. In all cases, exposure to the fact box is found to produce significantly positive odds ratio when compared to the control group.

### 8.4.2 Post-Task Estimations

Using the experimental groups as independent variables and the *Total Absolute Deviation of Fact Box Estimate* as the dependent variable, one-way Anova was performed with significant differences found  $F(3, 205) = 11.51, p = < .0001$ . Post-hoc



analyses were performed with Welch’s two sample t-test to compare the experimental fact box groups with the control group (see Table 8.7) and Cohen’s  $d$  was calculated as a measure of effect size (a large negative effect was found in all comparisons). Summary statistics for the estimation metric are provided in Table 8.7.

Group	$N$	$M$	$SD$	T-Test Results (compared to Control)
<b>Control</b>	49	120.2	42.85	-
<b>Inline Only</b>	54	84.31	42.87	$t(100.05) = 4.25, p = < .0001, d = -0.84$
<b>Small Before</b>	52	77.94	46.19	$t(98.98) = 4.77, p = < .0001, d = -0.95$
<b>Large Before</b>	54	70.35	51.95	$t(100.13) = 5.33, p = < .0001, d = -1.05$

**Table 8.6:** Welch’s two sample t-test is used to perform post-hoc analysis for the experimental groups against the control group for the post-task estimation questions which are combined to produce the dependent variable *Total Absolute Deviation of Fact Box Estimate*. Cohen’s  $d$  is calculated as a measure of effect size. A summary of the dependent variable is found in Table 8.7.

Group	Median	$M$	$CI_{lower}$	$CI_{upper}$
<b>Control</b>	117.0	120.2	107.9	132.5
<b>Inline Only</b>	77.5	84.3	72.6	96.0
<b>Small Before</b>	77.5	77.9	65.1	90.8
<b>Large Before</b>	68.0	70.4	56.2	84.5

**Table 8.7:** Summary statistics for the dependent variable *Total Absolute Deviation of Fact Box Estimate*. Median and means ( $M$ ) are provided, along with the lower and upper bounds of the 95% confidence intervals ( $CI_{lower}$  and  $CI_{upper}$  respectively).

## 8.5 Discussion

The results suggest that all three fact box approaches transfer the skill of using .org / .gov TLDs as a means to reduce the risk of harms from 3rd party data sharing.

The findings from the estimation questions indicate that knowledge of 3rd party sharing was significantly improved for all fact box variants when compared to control variant. There was a large effect for all variants, with the effect being strongest for

## 8. IDENTIFYING EFFECTIVE BOOSTS

---

the largest fact box and weakest for the inline only variant.

Given the odds ratios (Table 8.5) and effect sizes (Table 8.6) were strongest with the large fact box variant, this suggests that the details included (e.g. potential harms that could occur) provide motivation to adapt the skill. The effect sizes for both the large and small fact box variants (both providing longer exposure time to the fact box) are stronger than the inline only variant, a finding suggesting that longer exposure time to the fact box is necessary to encourage skill development.

We therefore conclude that both **H1** and **H2** are confirmed. Unfortunately, due to a technical problem with the data collection, we could not test **H3**.

### 8.5.1 Limitations

There are several limitations to the pilot study worth mentioning, which are possible pathways for future work. First, the participant sample was from the United Kingdom to more closely match the expected sample for the planned in lab study, and likely does not represent populations in other places around the globe (e.g. Southeast Asia). Second, there are alternative methods for analysing estimation tasks which were not considered, such as the methods outlined by [37]. There is also the possibility that members of one or more of the groups in our study had biased beliefs about TLDs (e.g. .org and .gov TLDs are more / less safe than other TLDs with respect to privacy), and is a challenge difficult to overcome in the between group design used in the current study. This last point highlights the need for understanding cultural and regional differences, for example running a study that compares beliefs around TLDs and privacy for participants in the United Kingdom versus those in America.

Additionally, fact boxes are one of multiple methods to *boost* an individual with

skills for harm reduction. There are others to consider, such as fast and frugal trees (FFTs), which have not been tested in the current study. Last but not least, we only consider longitudinal learning effects of the intervention over the very short time span of our study, future work should consider much larger time scales (such as days, weeks and months).

### 8.5.2 Conclusions and Recommendations

We have run a study to compare the effectiveness of different fact box approaches to *boost* searchers with a skill to better protect themselves from potential harms from 3rd party data sharing by websites they visit during the search process. *Boosting*, a cognitive intervention to enable people with skills to prevent harms, was the underlying methodology guiding this study. We find that all approaches evaluated in our pilot study are effective at teaching this skill.

Based on our findings, our recommendations are as follows. First, the inline only variant used in our study is highly desirable as it is the least impactful to existing search environments and therefore the most viable (**G-RQ-1**) in a commercial search setting. However, this same variant, though effective (**G-RQ-2**), also produced the weakest effect towards the goal of harm prevention. The inline only approach is nonetheless one that should be examined further. Second, when comparing the two approaches that present a fact box before search tasks, the large fact box is clearly better at teaching the skill when compared to the smaller and less detailed fact box. Given that any intervention providing such fact boxes prior to a search task will require extra time an effort for the searcher, we therefore recommend that the large fact box approach be used as only a limited amount of effort is needed for the searcher over the smaller fact box.

## 8. IDENTIFYING EFFECTIVE BOOSTS

---

We close with some comments with respect to the implementation of fact boxes in practice. We advocate for placement of small fact boxes (e.g. the inline variant) in existing search environments, such as in the right hand rail of the SERP or as Web browser plug-ins. With respect to the more highly detailed large fact boxes, it is unlikely that such an approach could easily be incorporated into an existing search engines (due to space constraints on the page). Nonetheless, the large fact boxes show such a strong effect that other methods should be considered, such as education campaigns in school and providing links to them in the SERPs for users that wish to educate themselves.

### 8.6 Summary

Related to the framework, we have introduced several *boosting* strategies ([FC-Cognitive](#)) and implemented them in a simulated search environment ([FC-System](#)). As *boosting* interventions are designed with transfer of a competency for better decision making, we have evaluated ([FC-Evaluation](#)) the interventions both for immediate effect (results Section 8.4.1) and long-term effect (Sections 8.4.1 and 8.4.2).

The study presented here determined the inline fact box as the most viable ([G-RQ-1](#)) approach with respect to real-world implementation, which appears to have some potential to produce reduced privacy impacts, however this potential appears to be quite weak and therefore expected to be minimally effective([G-RQ-1](#)).

The large fact box was deemed least viable ([G-RQ-1](#)) in a commercial Web search setting, however may have applications in environments where education is a focus (e.g. primary and secondary schools). Nonetheless, this approach is certainly the most effective ([G-RQ-1](#)) with respect to understanding of the risks which are

being communicated and therefore expected to perform the best in an interactive environment.

The findings from this pilot study will be used as a guide for hypotheses in our final empirical chapter which compares the most effective *nudging* and *boosting* strategies a controlled lab environment. Decisions will have to be made as to how best to cope with findings (for the inline only and full fact box approaches) that strongly contrast one another.

## 8. IDENTIFYING EFFECTIVE BOOSTS

---

# Chapter 9

## Boosting vs. Nudging

### 9.1 Overview

Experimental evidence from results presented in the previous studies suggest that both *nudging* (Chapters 5 and 6) and *boosting* (Chapter 8) are effective strategies for steering Web search behaviour in a direction that results in less risky decision making, that in many cases reduced overall harms (with respect to harms from privacy).

Recall that *boosting* provides a competency to cope with the environment (Web search in the current case), versus *nudging* which changes behaviour through ‘choice architecture’ of the environment, a long-term effect should remain for the *boost* after removal of the intervention (see cognitive interventions in background Section 2.6 and recap of interventions in Section 8.2.2).

Based upon these theoretical statements, the main goals of the current study is a) to test the theories across *nudging* and *boosting* (G-RQ-3) and b) to confirm the viability and effectiveness of the interventions (G-RQ-1 and G-RQ-2).

## 9. BOOSTING VS. NUDGING

---

As the high-level research questions ( **G-RQ-1** and **G-RQ-2** ) are common thread through all studies covered the current thesis, the following hypotheses were formulated.

**H1a** Both *nudging* and *boosting* strategies will significantly outperform the Control search environment with respect to the harm being evaluated (privacy impact).

**H1b** Both *nudging* and *boosting* strategies will significantly increase compliance to the intervention (interactions with .org and .gov web pages) when compared to the Control system.

Results from the pilot study already give indications that a long-term effect on harm reductions can be expected based on the recall of risk information communicated to participants. Though the pilot study did not test for an effect in an interactive search environment, results in the pilot study, other studies in this thesis, and the literature (see [88] for example) guide us to formulate the following hypotheses with respect to the *nudging* and *boosting* strategies. Upon removal of the *boost* and *nudge* strategies:

**H2a** The *boost* strategy treatment group will perform significantly better than the Control group.

**H2b** No significant differences will be found between the *nudge* strategy treatment group and the Control group.

**H2c** The *boost* treatment group will perform significantly better than the *nudge* treatment group.

As it relates to the framework in Chapter 3, the methods and results that follow are most heavily focused on the cognitive (**FC-Cognitive**) and evaluation



([FC-Evaluation](#)) components, and demonstrates how future IR researcher can evaluate across theories of decision making.

## 9.2 Method

The methods used to test the hypotheses make use of many already introduced in the general methods. As with the other empirical chapters, we give extra attention to methods unique to this study. Unlike other studies, where procedures were introduced first, here we begin with the evaluation test set.

### 9.2.1 Evaluation Test Set

Only briefly introduced in the general methods, the [Offline Boosting/Nudging Test Set](#) used in the current study is quite similar to the test set in the offline *nudge* study, in that it contains a static set of web pages associated with the same Cochrane medical search tasks. The main motivation for a static test set being that it allows for a highly controlled experiment (like the offline *nudge* studies).

The evaluation test set was developed with data collected during the online *nudge* study (Chapter 6). Using the assumption that commonly submitted queries are representative of search results, the most frequently submitted query for each search task in the online experiment was used to create the current test set. Using the most frequent queries, the first 10 results in Bing (recall Bing was used in the online study) for each query-task pair were included in the dataset. *3rd party tracker* data was then linked to each of the results (as covered in 4.3.1).

To reduce the experiment length, we excluded two tasks from the test set before running in-lab experiments. Criteria for omission was to ensure an equal number of *helpful* and *does not help* search tasks and secondarily based upon imbalances of

## 9. BOOSTING VS. NUDGING

---

results linked to .org and .gov TLDs (the feature indicative of reduced risk of harm for privacy and misinformation, see Chapter 7). Using this criteria, three of the Cochrane tasks (3, 7 and 10) were linked to only 1 result with a TLD of .org or .gov. Task 3 and 10 were both similar and Task 3 (classified as *does not help*) was chosen at random for removal. To maintain balance between *helpful* and *does not help* search tasks, Task 6 (classified as *helpful*) was removed (chosen at random).

Summary metrics of the evaluation dataset are provided in the Table 9.1. Comparing the current dataset to values in the fact boxes (Figures 8.4 and 8.3) and findings on the usefulness of TLD for privacy protection (Chapter 7) we note the following. Using the same extrapolation methods in Section 8.3.1.1 to calculate harms and benefits per 100 websites, the values were calculated for the test set used in the current study as well. This allowed for direct comparisons of the values in the fact boxes presented during the experiment (Figures 8.4 and 8.3) to actual values of the test set used. There were notable differences in the values, however the fact boxes were not changed for this study, as they are seen as more representative of real world data.

Nonetheless, for posterity, it is important to report these differences, noting the following points: First, the mean number ( $M = 4.3$  in Table 9.1) of *3rd party trackers* for .org and .gov TLDs as a ratio to the mean for other TLDs ( $M = 8.2$  in Table 9.1) is approximately half (1.0 : 1.9), and that a ratio (1.0 : 2.3) is found with means in the earlier study (Chapter 7). Second, when comparing the harms and benefits values of .org and .gov TLDs in the current dataset (to those communicated in fact boxes) the benefits are dampened and the harms are increased relative to those expressed in the fact boxes.

		.gov / org	Other TLDs
Overview of Evaluation Dataset	Total Results Available	17	63
	Mean 3rd Party Trackers	4.3	8.2
	Median Trackers	3	6
Shared with ___ 3rd Party Companies	2 or Less (Benefits)	5	6
	8 or More (Harms)	4	25
Sharing Extrapolated (per 100 Websites) for Comparison with Fact Box Presented	Benefits	29	10
	Harms	24	40

**Table 9.1:** Summary of the (Offline Boosting/Nudging Test Set) used for evaluation in the current study. A total of 8 search tasks were used in the study. 10 results were available for each search task. Comparing the last two rows in the table to the fact boxes (Figures 8.4 and 8.3) , one will see that the benefits of .org and .gov are weakened and the harms increased, but nonetheless still offer better protection than other TLDs

## 9.2.2 Procedure

The following subsections introduce the key design elements for the experiment.

### 9.2.2.1 Search Tasks

Search tasks used in the current study are a subset of tasks in Table 4.1 (eight tasks in total + one practice task). Justification for the chosen tasks are included in Section 9.2.1, which outlines the evaluation test set.

### 9.2.2.2 Search Systems (and SERPs)

Many elements (e.g. SERPs and search task decision page) of the systems introduced in the general methods and used in the offline *nudge* study were carried over to the current study, with some tweaks necessary to meet requirements of the study design. Also, much of the code developed (in Python) to run the systems in the current studies was taken from that of earlier studies.

In total, three search systems were developed and used in the current study (Control, *nudge* and *boost*). All systems presented results in a static manner, like

## 9. BOOSTING VS. NUDGING

---

the offline study, therefore participants could not enter queries (as with the offline *nudge* study). As only a few participants in the prior offline and online *nudge* studies made use of privacy protection switch (the mechanism used to opt-out of the *nudge*), the choice was made to remove this switch as removal greatly reduced the technical complexity of implementing the study design.

Those key differences aside from previous studies, the Control system is exactly as described in the general methods Section 4.2.2.1. The result rankings for this Control are the original ranking the **Offline Boosting/Nudging Test Set** detailed in Section 9.2.1.

Given previous findings for the three *nudging* strategies, we implement the Re-ranking system (see Figure 9.1) in the current study, as this *nudge* strategy was found to be highly effective at reducing privacy impacts. Though Re-ranking was determined not as viable as the Stoplight strategy [S3], it was far more effective than the Stoplight strategy and more viable than the Filtering system and combined was the motivation for choice of this *nudge*. The Re-ranking *nudge* system (strategy [S2]) is nearly the same as described in general methods Section 4.2.2.4. However, contrary to the previous ranking approach based upon *3rd party trackers*, result rankings were instead based upon the TLD. In this manner, the original result ranking was maintained, however results with a .org or .gov TLD appear before any other result.

For the *boost* variant, the small fact box evaluated in the pilot study was placed in the right hand rail of the SERP (see Figure 9.2). Rankings used for this variant are the same as the Control system (original rankings). Two sub-variants of the *boost* were included in the study, modelled after the inline only variants and full fact box variants evaluated in the pilot study. The full fact box (the same as Figure 8.3)

was included (presented on a screen before the main experiment), as it was the most effective approach in the pilot and assumed more likely to produce an effect in the participant sample. The inline only variant was also evaluated in the current study (as appears in Figure 9.2).

**MEDICAL QUESTION 1:** Does surgery help obesity?

**HEALTH ISSUE: obesity** - Obesity is a complex disorder involving an excessive amount of body fat. Obesity isn't just a cosmetic concern. It increases your risk of diseases and health problems, such as heart disease, diabetes and high blood pressure. *Source: Mayo Clinic*

**TREATMENT: surgery** - a branch of medicine concerned with diseases and conditions requiring or amenable to operative or manual procedures *Source: Merriam-Webster*

You can click on links below

*helpful:* The medical treatment **helps** if the treatment is effective and has a direct positive influence on the specified illness.

*inconclusive:* The effectiveness of a medical treatment is **inconclusive** if medical professionals are still unsure if the treatment will have a positive, negative or no influence on the specified illness.

*does not help:* The medical treatment **does not help** if the treatment is ineffective and either has no effect or has a direct negative influence on the specified illness.

10 results available.

**Obesity Disease and Surgery - PubMed Central (PMC)**  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4590927>  
 Obesity is a medical disease that is increasing significantly nowadays. Worldwide obesity prevalence doubled since 1980. Obese patients are at great risk for complications with physical and psychological burdens, thus affecting their quality of life.

**Surgical treatment of obesity: a review.**  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2594758>  
 Obesity is a chronic disease due to excess fat storage, a genetic predisposition, and strong environmental contributions. This problem is worldwide, and the incidence is increasing daily. There are medical, physical, social, economic, and psychological comorbid conditions associated with obesity

**Weight loss surgery - NHS**  
<https://www.nhs.uk/conditions/weight-loss-surgery>  
 Weight loss surgery, also called bariatric or metabolic surgery, is sometimes used as a treatment for people who are very obese. It can lead to significant weight loss and help improve many obesity-related conditions, such as type 2 diabetes or high blood pressure.

**Figure 9.1: Re-ranking Search System for *Nudging* Users Towards .org / .gov TLDs.** - The Re-raking *nudge* as appeared for Cochrane Task 9. The results are re-ranked so those with a .org or .gov TLD appear first in the results. The Control system looked exactly the same, however the results used the original ranking.

## 9. BOOSTING VS. NUDGING

**MEDICAL QUESTION 2:** Does surgery help obesity?

**HEALTH ISSUE: obesity** - Obesity is a complex disorder involving an excessive amount of body fat. Obesity isn't just a cosmetic concern. It increases your risk of diseases and health problems, such as heart disease, diabetes and high blood pressure. *Source: Mayo Clinic*

**TREATMENT: surgery** - a branch of medicine concerned with diseases and conditions requiring or amenable to operative or manual procedures *Source: Merriam-Webster*

You can click on links below

10 results available.

**Weight loss surgery - NHS**  
<https://www.nhs.uk/conditions/weight-loss-surgery>  
 Weight loss surgery, also called bariatric or metabolic surgery, is sometimes used as a treatment for people who are very obese. It can lead to significant weight loss and help improve many obesity-related conditions, such as type 2 diabetes or high blood pressure.

**Obesity - Treatment - NHS**  
<https://www.nhs.uk/conditions/obesity/treatment>  
 Bariatric surgery is usually only available on the NHS to treat people with severe obesity who fulfil all of the following criteria: they have a BMI of 40 or more, or between 35 and 40 and another serious health condition that could be improved with weight loss, such as type 2 diabetes or high blood pressure

**Obesity Disease and Surgery - PubMed Central (PMC)**  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4590927>  
 Obesity is a medical disease that is increasing significantly nowadays. Worldwide obesity prevalence doubled since 1980. Obese patients are at great risk for complications with physical and psychological burdens, thus affecting their quality of life.

*helpful:* The medical treatment **helps** if the treatment is effective and has a direct positive influence on the specified illness.  
*inconclusive:* The effectiveness of a medical treatment is **inconclusive** if medical professionals are still unsure if the treatment will have a positive, negative or no influence on the specified illness.  
*does not help:* The medical treatment **does not help** if the treatment is ineffective and either has no effect or has a direct negative influence on the specified illness.

Out of 100 websites visited for health & medical issues in a search engine:		
How many out of 100:	.gov/.org websites	other websites
<b>Benefits</b>		
will share your data with ≤ 2 3rd party companies?	59	14
<b>Harms</b>		
will share your data with ≥ 8 3rd party companies?	11	33

**Figure 9.2: Search System with Fact Box to *Boost* Individuals with the Harms and Benefits of Results (based upon TLDs).** - The *boost* system used in the current study with Cochrane Task 9 displayed. A fact box (the same as Figure 8.4) about harms and benefits related to TLDs appears in the right hand rail. The original result ranking is used (same ordering as used in the Control).

### 9.2.2.3 Study Design

For the most part, study design specifics already introduced the general methods Section 4.2.3 are used in the current study. A between group design was used for in data collection of in lab experiments. Latin Squares and randomization was used for determining search task and system assignment (e.g. Control, *nudge*, *boost*) for each participant. As with the previous lab studies, participants filled out a consent form before the interactive experiment, completed a survey after the interactive experiment and were given a debriefing at the end of the survey. The same labs (and computer setup) used in previous studies were used again here. Once logged into the study, the users would progress through pre-experiment pages (e.g. instructions and practice task), pre-tasks, main tasks and post-tasks (see Figure 9.3).

The experiment was broken into two main phases, phase 1 (tasks 1 - 4)<sup>1</sup> and

<sup>1</sup>These tasks numbers represent the ordering of experiment, which do not map to the Cochrane task numbering.

phase 2 (tasks 5 - 8). The “treatment” phase of the experiment (tasks 1 - 4). The participants were randomly placed into one of the four experiment groups and proceeded through various phases (see Figure 9.3), where participants interacted with either the Control, *boost* or *nudge* system. During the second phase of experiment (tasks 5 - 8) all participants interacted with the Control system only, which was a crucial part of the design for **G-RQ-3** and **H2c**. Finally, to ensure balance of the Cochrane medical tasks, an equal number of *helpful* and *does not help* were assigned to each phase of the experiment and randomly assigned to each phase task.

Experimental Groups	Definition	Experiment Progression			
		Pre-Tasks	Task 1-4	Task 5-8	Post-Tasks Survey
<b>Control</b>	No factbox provided at any point in experiment.	Instructions Only	SERP only	SERP only	Questions From Pilot + Other Questions
<b>Boost - Inline Only</b>	Small Factbox provided during first 4 tasks		SERP + Small Factbox		
<b>Boost - Full Factbox</b>	Large Factbox given before tasks + inline during 1-4	Large Factbox + Instructions			
<b>Nudge</b>	Results re-ranked s.t. org/gov always at top	Instructions Only	SERP only (re-ranked)		

**Figure 9.3: Overview of the Study Design for Comparing *Boost* and *Nudge* Strategies.** - All participants completed 8 search tasks (a randomly selected Cochrane medical task as outlined in section 9.2.1).

**Pre and Post Task Questions** Added for this study, were pre and post-task questions (with Likert scale responses) about 3rd party data sharing. The pre-task questions aim was to gain some understanding into the variations of privacy concerns across the different tasks, as it is assumed some of these tasks may be considered more privacy sensitive than others. The post-task questions were used as a primer (to encourage behaviour for protection from 3rd party data sharing) and to understand their beliefs in their privacy protective behaviour. For the post-task questions, it is assumed that subjects in the boosting experiments groups reporting high privacy protective behaviour and low possibility of 3rd party data sharing would adhere to the boost more so than participants reporting otherwise. We recognize that asking these questions also has the disadvantage of being unrealistic in a real

## 9. BOOSTING VS. NUDGING

---

world environment and furthermore has the potential to bias participants behaviour.

Prior to each search task, participants answered the following questions (with 7 point scale for response)

**How concerned should you be that 3rd party companies may collect information about you during this search task?**

Not all concern (1) - Very concerned (7)

**How much harm is possible due to 3rd party companies collecting information about you during this search task?**

No harm is possible (1) - Great harm is possible (7)

After each search task, participants answered the following questions (with 7 point scale for response)

**To what extent did your search behaviour influence how many 3rd party companies collected information about you during the search task?**

Search behaviour had no influence. (1) - Search behaviour had a lot of influence. (7)

**What is the possibility that 3rd party companies collected information about you during the search task?**

No possibility. (1) - High possibility. (7)

**Post interactive experiment survey** After all 8 search tasks were completed (completion of main experiment), participants were taken to a survey hosted on the



Qualtrics survey platform and completed survey items included in other studies. Additionally, nearly identical questions (multiple choice and estimation questions) to those used in the pilot study (see Section 8.3.1.2) are included in the current study to measure knowledge gained. For the current study, and contrary to the pilot study, fact boxes were not displayed in these questions for any of the participant groups.

### 9.2.3 Evaluation Metrics

The current study makes use of a much smaller set of metrics compared to those used in the previous studies and include only those necessary to test the hypotheses listed.

For evaluating the effects of the strategies on the privacy harm being prevented we use *Average Number of Trackers*, however for the current study we only focus on the harms from privacy and do not focus on the search task decisions. Though data could be analysed in the future, evaluating effects on decisions was deemed much lower priority for this study and analyses was therefore not performed.

Given the hypotheses for harm interventions being evaluated made use of previous findings (Chapter 7) demonstrating that .org and .gov TLDs can greatly reduce the risk of personal data being shared with 3rd parties, three additional evaluation metrics were developed based upon the TLD. The first metric, is simply based on the number of visits to .org and .gov web pages normalized by the total number of web page visits. This measure was motivated by the need to measure compliance to the strategy (as introduced in Section 4.4.2, which was used to develop a compliance measure based upon Stoplight colours in the online *nudge* study). The compliance measure is defined as:

$$Percent_{OrgGov} = \text{Total visits to websites with a .org or .gov TLD} / \text{total}$$

number of website visits

While the `PercentOrgGov` metric is desirable for its simplicity and insights into interactive behaviours, it has several problems. First, this metric does not take into consideration the likelihood of encountering .org / .gov websites. Second, the fact box distinguishes between benefits and harms and it is therefore important to evaluate from this perspective as well. Finally, this metric does not provide a system evaluation in the traditional sense. These problems combined motivated adaptation of IR system metrics (see adaptation of IR metrics in Section 4.4.3.2).

In the case of the current study, adaptation of the normalized discounted cumulative gain (nDCG) [106] is the approach used. Taken from [106], the normalized discounted cumulative gain (nDCG) at rank  $p$  is defined as follows:

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (9.1)$$

Where:

$$DCG_p = \sum_{i=1}^p \frac{rel_i}{\log_2(i+1)} \quad (9.2)$$

Where  $rel_i$  is the graded relevance of result at rank  $i$  and  $rel_i \in \mathbb{R}$  and:

$$IDCG_p = \sum_{i=1}^{Q_p} \frac{rel_i}{\log_2(i+1)} \quad (9.3)$$

Where  $IDCG_p$  is the ideal cumulative gain of query  $Q$  at rank  $p$  and  $Q_p$  is the set of results ordered by relevance score of greatest to lowest value.

With appropriate substitutions for  $rel_i$ , the following two evaluation metrics are additionally used in the analyses.

The compliance measure is defined as:

$nDCB_p$  is defined as  $nDCG_p$  where  $rel_i$  is 1 if result visited is .org or .gov and 0 otherwise.

$nDCH_p$  is defined as  $nDCG_p$  where  $rel_i$  is 0 if result visited is .org or .gov and  $-1$  otherwise.

Finally, measures introduced in the pilot study to test for skills acquired from the strategies are used in the current study. Three binary measures (**All Task Questions Correct**, **Task 1 Questions Correct** and **Task 2 Questions Correct**) and the estimation measure **Total Absolute Deviation of Fact Box Estimate** as defined in the pilot study Section 8.3.2 are used in the current study as well.

#### 9.2.4 Statistical Tests

To account for variations in user behaviour for the interactive experiment, linear mixed effects regression (LMER) was used to test the hypotheses, with the experimental treatment groups (Control, *Nudge*, *Boost* Inline, *Boost* Large) as a single fixed effect and participant and search task as two random effects. All four evaluation metrics ( $Average_{3rdParty}$ ,  $Percent_{OrgGov}$ ,  $nDCB_{10}$  and  $nDCH_{10}$ ) in Section 9.2.3 are used as dependent variables in the analyses.

For a direct comparison between post interactive survey questions in the current study and the same questions in the pilot study (Chapter ??) the same statistical tests are performed.

## 9. BOOSTING VS. NUDGING

---

As was used for analyses in the pilot study, Chi-Squared (logistic regression for post-hoc analyses) is used for tests on the three binary count measures and one-way Anova (Welch’s two-sample t-test for post-hoc analyses) for tests on the estimation measure.

For all analyses,  $\alpha = .05$  is set as the threshold for significance.

### 9.2.5 Participants

Participants were recruited in almost the same manner as previous lab based studies (with  $n = 30$  to be assigned to each group), with the only difference being told the study was about privacy protection in Web search, with a total of  $N = 120$  subjects recruited. However, due to the Covid-19 crisis, and ethical concerns with continuation of a lab study during the crisis, data collection was stopped. In total,  $n = 70$  participants completed the experiment. The participants reported a mean age ( $M = 22.4, SD = 4.4$ ) of which  $n = 50$  were female,  $n = 19$  were male and  $n = 1$  as other.  $n = 26$  participants reported attainment of an undergraduate degree or higher, and  $n = 38$  were native English speakers. No significant differences were found with distribution of these demographics across the 4 experimental groups. As with other lab based studies, each participant was paid £10 for their time.

## 9.3 Results

With respect to the interventions and hypotheses there were a mixture of results found. The results are split between those based on data collected during the interactive Web search phase and the post experiment survey. Across the entire study, no effects were found with the inline only variant of the *boost*, and therefore do not discuss these further (but do include analyses in the tables and figures for

comparison).

### 9.3.1 Main Study Results

The results of the main interactive study are important for discussion related to all hypotheses ( **H1a** - **H2c** ) and high level research questions and in particular **G-RQ-3** .

Tables 9.2 - 9.5 provide results of the analyses using LMER to test hypotheses with the 4 dependent variables (*Average<sub>3rdParty</sub>*, *Percent<sub>OrgGov</sub>*, *nDCB<sub>10</sub>* and *nDCH<sub>10</sub>*). Summary statistics are included in each table for the 4 experiment groups (3 interventions + control). Interventions producing significant differences compared to the control are **bolded** and those demonstrating strong tendencies (defined as  $.05 < p < .10$ ) are *italicized*.

Figures 9.4a - 9.4d are scatterplots (with smoothers) comparing each of the four experimental variants across all 8 search tasks and dependent variables. The visualizations are included to provide additional insights with respect to the interventions during the ‘treatment’ phase (tasks 1 - 4) and the ‘post-treatment’ phase (tasks 5 - 8). Based on the underlying theory of *nudging* and *boosting* and results from previous studies, the idealized visualizations would include the following behaviours:

During the ‘treatment’ phase (task 1 - 4)

- Due to the ‘choice architecture’ of the *nudge*, performance (for all 4 metrics) should be optimal across all 4 tasks. It should also out perform the *boost* and Control.
- Due to the learning effect, as the task number increases, *Boosting* should produce improving performance (i.e. decreasing Average Number of *3rd*

## 9. BOOSTING VS. NUDGING

---

*party trackers* and  $nDCH_p$ , increasing  $Percent_{OrgGov}$  and  $nDCB_p$ . Furthermore, this learning effect should behave asymptotically and at some point outperform the Control.)

- The Control system is a baseline and therefore should remain generally flat.

During the ‘post-treatment’ phase (task 5 - 8)

- Upon removal *nudge*, performance (for all 4 metrics) should perform similarly to the Control as no skill was learned.
- After removal of the *boost*, performance should be maintained as a result of the learning effect. That is all 4 performance metrics should remain at a similar level to the asymptote of curve during the ‘treatment’ phase. Performance should be better than both the *nudge* and the Control)
- Again, the Control system is a baseline and therefore should exhibit similar behaviour to the ‘treatment’ phase

Insights gained from these figures are provided in the subsequent discussion (Section 9.4).

**Strategies on the Harm Being Prevented (Privacy)** Results Table 9.2 and Figure 9.4a indicate a very large effect on reduction of privacy impacts for the *nudge* strategy during the ‘treatment’ phase and ‘post-treatment’ phase of the interactive study when compared to the control. A significant effect is found as well for the *boost* strategy (with large fact box) during the ‘post-treatment’ phase but not during the ‘treatment’ phase.

Group	Summary									
	First 4 Tasks					Last 4 Tasks				
	<i>Median</i>	<i>M</i>	<i>SE</i>	<i>CI<sub>lower</sub></i>	<i>CI<sub>upper</sub></i>	<i>Median</i>	<i>M</i>	<i>SE</i>	<i>CI<sub>lower</sub></i>	<i>CI<sub>upper</sub></i>
Control	6.99	7.93	0.52	6.90	8.96	9.24	9.42	0.50	8.43	10.41
<i>Boost</i> Inline	7.30	8.16	0.54	7.09	9.23	8.35	8.66	0.59	7.49	9.83
<i>Boost</i> Large	6.84	7.57	0.50	6.58	8.56	5.47	6.91	0.50	5.90	7.92
<i>Nudge</i>	6.00	6.64	0.40	5.85	7.44	7.94	7.98	0.48	7.02	8.94

Group	Results										
	First 4 Tasks					Last 4 Tasks					
	<i>vs. Control</i>	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>Pr(&gt;  t )</i>	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>Pr(&gt;  t )</i>
<i>Boost</i> Inline		-0.17	0.53	66.80	-0.33	.7441	-0.13	0.67	66.92	-0.19	.8489
<i>Boost</i> Large		<b>-0.94</b>	<b>0.53</b>	<b>65.41</b>	<b>-1.77</b>	<b>.0812</b>	<b>-1.85</b>	<b>0.67</b>	<b>66.05</b>	<b>-2.76</b>	<b>.0076</b>
<i>Nudge</i>		<b>-1.06</b>	<b>0.52</b>	<b>64.05</b>	<b>-2.04</b>	<b>.0459</b>	<b>-1.65</b>	<b>0.66</b>	<b>65.10</b>	<b>-2.50</b>	<b>.0150</b>

**Table 9.2:** Comparison of *Boost* vs. *Nudge* strategies on harms from privacy. Metric used is Average 3rd Party Trackers. See corresponding scatterplot in Figure 9.4a.

### Compliance with strategy (visit .org and .gov to reduce privacy harms)

Results Table 9.3 and Figure 9.4b indicate a massive effect related to compliance (visiting .org and .gov) with the *nudge* strategy during the ‘treatment’ phase, however this effect disappears entirely during ‘post-treatment’ phase when compared to interactions in the Control system. During the ‘treatment’ phase a significant effect is found as well for the *boost* strategy (large fact box) and a strong tendency for this behaviour is apparent during the ‘post-treatment’ phase.

Group	Summary									
	First 4 Tasks					Last 4 Tasks				
	<i>Median</i>	<i>M</i>	<i>SE</i>	<i>CI<sub>lower</sub></i>	<i>CI<sub>upper</sub></i>	<i>Median</i>	<i>M</i>	<i>SE</i>	<i>CI<sub>lower</sub></i>	<i>CI<sub>upper</sub></i>
Control	0.00	0.13	0.02	0.09	0.18	0.00	0.17	0.02	0.12	0.21
<i>Boost</i> Inline	0.00	0.15	0.02	0.11	0.20	0.00	0.17	0.03	0.11	0.22
<i>Boost</i> Large	0.25	0.27	0.04	0.20	0.34	0.17	0.23	0.03	0.16	0.29
<i>Nudge</i>	0.33	0.34	0.03	0.28	0.40	0.00	0.15	0.02	0.10	0.19

Group	Results										
	First 4 Tasks					Last 4 Tasks					
	<i>vs. Control</i>	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>Pr(&gt;  t )</i>	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>Pr(&gt;  t )</i>
<i>Boost</i> Inline		0.01	0.05	67.50	0.19	.8533	0.01	0.04	68.23	0.34	.7386
<i>Boost</i> Large		<b>0.14</b>	<b>0.05</b>	<b>66.72</b>	<b>2.80</b>	<b>.0067</b>	<i>0.07</i>	<i>0.04</i>	<i>66.89</i>	<i>1.73</i>	<i>.0885</i>
<i>Nudge</i>		<b>0.21</b>	<b>0.05</b>	<b>65.90</b>	<b>4.35</b>	<b>.0000</b>	-0.01	0.04	65.53	-0.37	.7161

**Table 9.3:** Comparison of *Boost* vs. *Nudge* strategies on compliance to the strategy (stick with .org and .gov TLDs). Metric used is the ratio of Visits on .org/.gov Websites. See corresponding scatterplot in Figure 9.4b.

## 9. BOOSTING VS. NUDGING

**System Comparisons with  $nDCB_p$  (Benefits of System)** Turning to system comparisons based upon the adapted normalized discounted cumulative gain metric for system benefits ( $nDCB_p$ ), results in Table 9.4 and Figure 9.4c indicate a strong effect for system benefits of the Re-ranking *nudge* strategy during the ‘treatment’ phase when compared to the Control system, but effects are not present in the ‘post-treatment’ phase. Benefits are not indicated for the *boost* system.

Group	Summary					Summary				
	First 4 Tasks					Last 4 Tasks				
	<i>Median</i>	<i>M</i>	<i>SE</i>	<i>CI<sub>lower</sub></i>	<i>CI<sub>upper</sub></i>	<i>Median</i>	<i>M</i>	<i>SE</i>	<i>CI<sub>lower</sub></i>	<i>CI<sub>upper</sub></i>
Control	0.00	0.14	0.02	0.09	0.18	0.00	0.19	0.03	0.14	0.24
<i>Boost</i> Inline	0.00	0.18	0.03	0.12	0.23	0.00	0.14	0.02	0.09	0.19
<i>Boost</i> Large	0.24	0.23	0.03	0.18	0.28	0.22	0.21	0.03	0.15	0.26
<i>Nudge</i>	1.00	0.79	0.06	0.68	0.90	0.00	0.18	0.03	0.13	0.24

Group	Results					Results				
	First 4 Tasks					Last 4 Tasks				
vs. Control	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>Pr(&gt;  t )</i>	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>Pr(&gt;  t )</i>
<i>Boost</i> Inline	0.02	0.08	66.74	0.28	.7781	-0.03	0.04	67.62	-0.71	.4798
<i>Boost</i> Large	0.10	0.08	66.37	1.26	.2133	0.02	0.04	66.65	0.48	.6325
<i>Nudge</i>	<b>0.65</b>	<b>0.08</b>	<b>65.94</b>	<b>8.63</b>	<b>.0000</b>	0.00	0.04	65.60	-0.09	.9266

**Table 9.4:** Comparison of *Boost* vs. *Nudge* systems with the adapted IR system metric  $nDCB_p$ . See corresponding scatterplot in Figure 9.4.



**System Comparisons with  $nDCH_p$  (Harms of System)** Analyses was performed with a harm based version ( $nDCH_p$ ) of normalized discounted cumulative gain metric. Again, strong effects for the *nudge* strategy [S2] system for the ‘treatment’ phase (but no effects in the ‘post-treatment’ phase) as indicated in results Table 9.5 and Figure 9.4d. The results indicate a strong tendency for reductions in harm with the *boost* fact box strategy [S4] system in the ‘post-treatment’ phase.

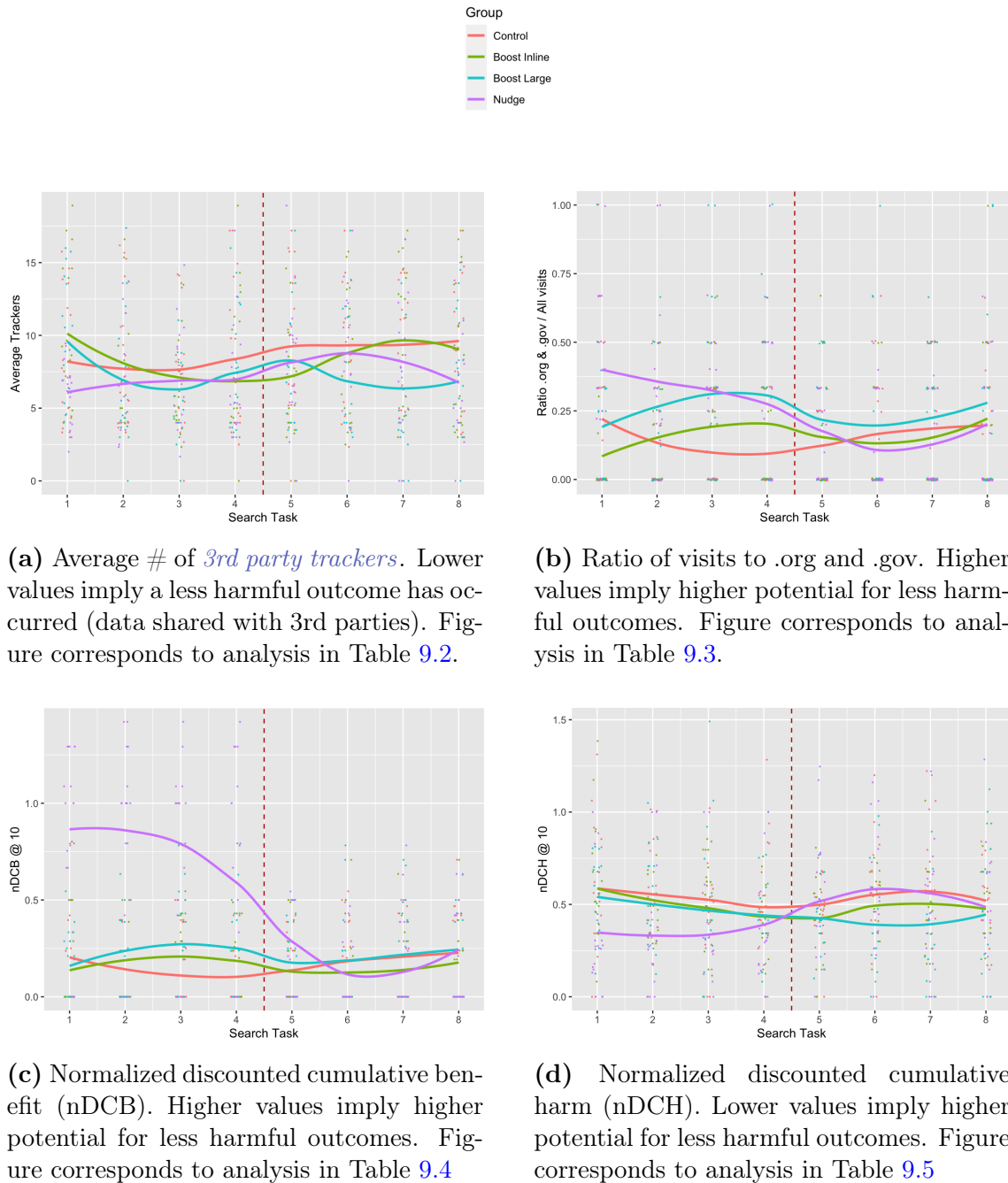
Group	Summary									
	First 4 Tasks					Last 4 Tasks				
	<i>Median</i>	<i>M</i>	<i>SE</i>	<i>CI<sub>lower</sub></i>	<i>CI<sub>upper</sub></i>	<i>Median</i>	<i>M</i>	<i>SE</i>	<i>CI<sub>lower</sub></i>	<i>CI<sub>upper</sub></i>
Control	0.52	0.55	0.03	0.48	0.61	0.56	0.53	0.03	0.47	0.59
<i>Boost</i> Inline	0.48	0.50	0.03	0.44	0.56	0.49	0.48	0.03	0.42	0.53
<i>Boost</i> Large	0.46	0.49	0.04	0.41	0.56	0.40	0.41	0.03	0.35	0.48
<i>Nudge</i>	0.29	0.35	0.03	0.28	0.41	0.51	0.54	0.04	0.47	0.62

Group	Results										
	First 4 Tasks					Last 4 Tasks					
	<i>vs. Control</i>	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>Pr(&gt;  t )</i>	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>Pr(&gt;  t )</i>
<i>Boost</i> Inline		-0.04	0.07	66.83	-0.52	.6039	-0.07	0.07	66.72	-0.96	.3390
<i>Boost</i> Large		-0.06	0.07	66.44	-0.87	.3867	<i>-0.12</i>	<i>0.07</i>	<i>66.29</i>	<i>-1.76</i>	<i>.0833</i>
<i>Nudge</i>		<b>-0.20</b>	<b>0.07</b>	<b>66.00</b>	<b>-2.94</b>	<b>.0045</b>	0.01	0.07	65.78	0.19	.8479

**Table 9.5:** Comparison of *Boost* vs. *Nudge* systems with the adapted IR system metric  $nDCH_p$ . See corresponding scatterplot in Figure 9.5.

## 9. BOOSTING VS. NUDGING



(a) Average # of *3rd party trackers*. Lower values imply a less harmful outcome has occurred (data shared with 3rd parties). Figure corresponds to analysis in Table 9.2.

(b) Ratio of visits to .org and .gov. Higher values imply higher potential for less harmful outcomes. Figure corresponds to analysis in Table 9.3.

(c) Normalized discounted cumulative benefit (nDCB). Higher values imply higher potential for less harmful outcomes. Figure corresponds to analysis in Table 9.4

(d) Normalized discounted cumulative harm (nDCH). Lower values imply higher potential for less harmful outcomes. Figure corresponds to analysis in Table 9.5

**Figure 9.4:** *Boost* vs. *Nudge* scatter plots (with smoother) for dependent variables tested. The first four tasks (treatment phase) and last four tasks are separated by a red dashed line. Participants were assigned at random to the four experimental variants (Control, *Nudge*, *Boost* Inline or *Boost* Large).

### 9.3.2 Evaluation of Knowledge Gained

After the main interactive experiment, all participants responded to the same survey questions asked in the pilot study (Chapter 8). Using the binary count and estimation metrics introduced in Section 9.2.3, analysis was performed as outlined in Section 9.2.4.

Results provided below are relevant for **G-RQ-3** and hypotheses (**H2a** - **H2c**).

**Interventions and Knowledge Gained (Binary Measures)** Though the results are non-significant, there are clear tendencies for the *boost* intervention (with Large fact box) to outperform participants in the Control as well as other interventions (*nudge* and *boost* (with small fact box)).

		Control	Boost Inline	Boost Large	Nudge	$\chi^2(df = 3)$	$p$
Sample	N	18	17	17	18	-	-
All Task Questions Correct	Yes	1	2	5	1	5.8894	0.1171
	No	17	15	12	17		
Task 1 Questions Correct	Yes	2	4	6	2	4.3954	0.2218
	No	16	13	11	16		
Task 2 Questions Correct	Yes	6	3	6	3	2.7143	0.4378
	No	12	14	11	15		

**Table 9.6:** As was performed in the pilot study (Chapter 8), Chi Squared Analysis was performed on the 3 binary metrics for the task questions to compare differences between the 3 experimental groups and the control group.

**Interventions and Knowledge Gained (Estimation Measures)** For the estimation one-way Anova was used for our analysis resulting in non-significant findings  $F(3, 66) = 0.649, p = .586$ . As the results were non-significant, no post-hoc analyses were performed. We do provide summary statistics (see Table 9.7) as it provides useful insights regarding the different interventions. For instance, both *boost* strategies have measures of central tendency indicating performance better than the control and the *nudge*.

Group	<i>Median</i>	<i>M</i>	<i>CI<sub>lower</sub></i>	<i>CI<sub>upper</sub></i>
<b>Control</b>	118.5	123.9	100.5	147.4
<b>Boost Inline</b>	99.0	108.6	75.6	141.6
<b>Boost Large</b>	101.0	117.9	80.5	155.2
<b>Nudge</b>	134.5	134.5	116.6	152.5

**Table 9.7:** Summary statistics for the dependent variable *Total Absolute Deviation of Fact Box Estimate*. Median and means (*M*) are provided, along with the lower and upper bounds of the 95% confidence intervals (*CI<sub>lower</sub>* and *CI<sub>upper</sub>* respectively). The questions asked were the same as those used in the pilot study covered in (Chapter 8) as was the evaluation metric (*Total Absolute Deviation of Fact Box Estimate*).

## 9.4 Discussion

As with the other lab based studies, we break the discussion apart to first highlight findings related to the hypotheses and then focus on the research questions. The key findings from the results of the main interactive experiment are summarised in Table 9.8 and are used as a guide for this discussion.

To simplify the discussion, we address the non-significant findings for the inline only variant of the *boost* here. Though this *boost* strategy is certainly the most viable of the two *boost* approaches evaluated in the experiment, there is no signal detected to suggest it is an effective approach (at least for prevention of harms related to privacy). This approach was the weakest performer of the three *boost* fact boxes evaluated in the pilot study, so the finding is not entirely surprising. With this topic out of the way, the focus is simplified to comparison of the other *boost* (with large fact box presented before the main experiment), the *nudge*, and the Control.

Dependent Variable	First 4 Tasks		Last 4 Tasks	
	Treatment Phase		Post-Treatment Phase	
	Boost	Nudge	Boost	Nudge
$Average_{3rdParty}$	+	*	**	*
$Percent_{OrgGov}$	**	***	+	ns
$nDCB_{10}$	ns	***	ns	ns
$nDCH_{10}$	ns	**	+	ns

**Table 9.8:** A summary of findings across the four performance metrics are provided for comparison of the *boost* (small fact box in SERP with large fact box presented before experiment), *nudge* system with the *Control* system. Interventions were removed during the ‘Post-Treatment’ Phase. \*\*\* =  $p < .001$ , \*\* =  $p < .01$ , \* =  $p < .05$ , + =  $.05 < p < .1$  and ns =  $p > .1$ .

### 9.4.1 Findings Related to Study Specific Hypotheses

#### 9.4.1.1 Treatment Phase (H1a and H1b)

For the ‘treatment’ phase (tasks 1 - 4) the *nudge* approach is a significant intervention when compared to the control across all 4 dependent variables. The  $Percent_{OrgGov}$  for the large fact box intervention is significantly higher when compared to the control and the same intervention produces strong tendencies for  $Average_{3rdParty}$  encounters.

#### 9.4.1.2 Post-Treatment Phase (H2a - H2c)

Ultimately the overall goal of this study (and to a major extent the goal of this thesis) was to compare the claims of the two popular and quite different harm prevention strategies from the behavioural and cognitive sciences (*boosting* and *nudging*) adapted as interventions for harm prevention in Web search.

For the “post-treatment” phase (tasks 5 - 8),  $Average_{3rdParty}$  is significantly less for both the *boost* and *nudge* interventions when compared to the control. During this phase strong tendencies were found for the  $Percent_{OrgGov}$  and  $nDCH_{10}$  dependent variables, however not for the  $nDCB_{10}$  variable. For the *nudge* intervention, none of the findings were significant nor were the tendencies strong.

#### 9.4.1.3 Mixed Phases (H2a - H2c)

**H1b** is central to the overall risk mitigation strategies to prevent harm, in this case being that one should stick to websites with .org and .gov TLDs. The  $Percent_{OrgGov}$ ,  $nDCB_{10}$  and  $nDCH_{10}$  evaluation metrics were motivated by this hypotheses. The results do provide strong evidence the *nudge* intervention is an

effective strategy, however they also provide strong evidence that without this intervention the user behaviour is no different than of users in the Control environment (no intervention), which subsequently confirms [H2b](#).

#### 9.4.1.4 Knowledge Evaluation ([H2a](#) - [H2c](#))

### 9.4.2 Findings in the Context of Research Questions

For this particular study, the focal harm was loss of privacy as a result of 3rd party data sharing and [G-RQ-2](#) was central to this matter. The results provided suggest that both *boosting* and *nudging* strategies exhibit significant harm reduction with respect to this parameter. Interestingly, the results are counter to the expectation that harms would return after removal of *nudge*. However, this finding does not necessarily counter the expectation, as the evaluation dataset was static and a much smaller than planned participant sample was used in the analyses.

Turning to [G-RQ-3](#), which goes beyond Web search and applies to the theoretical claims of *nudging* vs. *boosting*, the findings do suggest that a skill is learned with the *boost* while *nudging* does not produce such a skill. This claim is particularly evident for *nudging* when considering the data in Figures 9.4b - 9.4d, where large drop-offs in performance occur once the intervention is removed. Also notable across all dependent variables (based upon the smoothers in Figure 9.4) a learning effect appears to be present in the ‘treatment’ phase for both *boosting* approaches (however is much more notable for the large fact box group), which is in line with expected behaviour. Furthermore, the data in the last 4 tasks (again based upon smoothers in Figure 9.4 as well as Tables 9.6 and 9.7) suggest that some skill is being learned for *boosting*, whereas there is little (outside of the Average Number of *3rd party trackers*) to suggest this is the case for *nudging*. Both *nudging* and *boosting*

appear to be effective interventions at steering users towards results and decisions that are less risky, however when comparing the results for first 4 tasks (Figures 9.4b - 9.4d) the *nudge* intervention is definitely stronger. This effect is very likely due to the trust bias commonly exhibited in SERPs, where users typically do not go beyond the second result.

Finally, **G-RQ-1** is an important factor that must not be overlooked. As already stated, the in line *boost* is certainly the more viable for commercial settings for privacy harm prevention, however it is certainly not viable for the sample tested. Additional analysis is desirable to consider user differences in a similar manner to the other lab studies (e.g. considerations for privacy actions in daily life), however the small sample did not permit such analyses. The *nudge* approach is certainly the most effective, but previous findings in the online *nudge* study certainly suggest that caution must be used with such an approach. Finally, the large fact box boost may not be viable in a commercial Web search environment, but we certainly argue that it has viability in the context of educational policy. For instance, we argue that it would be quite simple to include such information in the curriculum at a school.

### 9.4.3 Limitations

There are several limitations to our study to consider.

First and foremost, due to the Covid-19 pandemic the study was not run to the full extent that was planned. Just over half ( $n = 70$ ) of the recruited participants ( $N = 120$ ) completed the study. Findings from this study are quite likely underpowered (however we have not performed post-hoc power analyses), thus results should be interpreted cautiously. In line with the underpowered sample, there is a very high possibility that type I errors (related to significant findings) and type II errors



have occurred with respect to the non-significant findings. Also, due to the small sample, we could not test potential differences based on user differences (based upon self-report measures). Running the same study with a larger sample is the obvious pathway to address these matters.

As with previous studies we have run, 3rd party tracking is only one pathway to harm in Web search, there are many others. Future work should consider other pathways, such as the misinformation factors related to TLDs (see findings Chapter 7).

There are other interventions that can (and should be) considered. We have evaluated fact boxes as a *boost* (strategy [S4](#)) and a Re-ranking *nudge* (strategy [S2](#)). Studies in the future should compare other *boost* approaches such as Fast and Frugal Trees (FFTs) and *nudge* interventions such as filtering.

Cultural and regional difference certainly exist that have not been tested. For instance, .gov websites are in English and Spanish and would not commonly be used in countries with other languages as the mother tongue.

Last of all, the evaluation dataset is static and therefore not fully representative of a dataset in the wild. User behaviour (and the results visited) would most certainly change if the experiment were connected to an online environment allowing free query entry (in a similar manner to the online *nudge* study). For example, with a large enough participant sample, it is expected that some users provided with the *boost* intervention will enter wild cards (e.g 'alcohol benzodiazepines site:\*.org') in the query knowing that it reduces risk related to loss of privacy in Web search.

### 9.5 Summary

A study investigating two approaches (*boosting* and *nudging*) from the behavioural and cognitive sciences as interventions to reduce potential harms as a result of online Web search was completed.

Similar to the earlier online *nudge* study, there are several novel contributions with respect to the framework worth noting. Related to evaluation ([FC-Evaluation](#)), we adapted  $nDCG$  to compare systems with respect to the harms and benefits of different results in each system. Additionally, we introduced a behavioural interactive measure ( $Percent_{OrgGov}$ ) with respect to the informational cue (TLD), similar to the Stoplight interactive measures developed in the online study, which allowed for comparison of interactive behaviour indicative of adherence across the systems. For the *nudge* system, there are no signals of adherence to the intervention after it's removal, whereas this is seen for the *boost* (demonstrating a new competency). Novel methods were also introduced to compare across different theoretical paradigms for behavioural and cognitive ([FC-Cognitive](#)) based decision making interventions.

The study utilized previously published search tasks and included interventions based upon findings (Chapter 7) showing that results with .org and .gov TLDs are less risky than other TLDs with respect to data being shared with 3rd party companies. The *nudge* intervention demonstrated a strong effect on user behaviour, where users visited significantly more .org and .gov websites and reduced 3rd party data sharing compared to the control. Strong signals (some significant) were also found for the *boost* intervention, however the intervention appears to be weaker than the *nudge* approach. As proponents of *boosting* point out, the intervention is more ethically sound than *nudging* as there is transparency provided, however such interventions are likely most effective for users that are motivated to minimize

personal harm during Web search, which leaves much space for further investigation.



# Chapter 10

## Conclusions

### 10.1 Summary of Thesis

A theoretical framework was proposed as the foundation to address potential harms one might encounter during Web search. The framework suggests that behavioural and cognitive interventions (see [FC-Cognitive](#)) should be considered as part of the evaluation (see [FC-Evaluation](#)) of search systems designed (see [FC-System](#)) with harm prevention as the paramount goal. [FC-System](#) and [FC-Evaluation](#) of the framework (system design and system evaluation) are seen as extensions of approaches already used in the development and evaluation of Web search systems. Policy ([FC-Policy](#)), the remaining component of the framework, is also a critical for efforts to reduce harms to individuals and broader society related to Web search systems. However, as this thesis was focused on the empirical science of harm prevention, no studies made consideration of [FC-Policy](#), and in fact see the other three components as necessary to inform policy.

Taken from [FC-Cognitive](#), this thesis investigated *nudging* and *boosting*, two

## 10. CONCLUSIONS

---

approaches commonly used in many domains (e.g. medicine, finances) to promote decision making for minimized harm to individuals and broader society. Four general strategies ([S1](#)- [S4](#)) and were developed from the *nudging* and *boosting* paradigms, which were first introduced in the general methods and provided again below. These strategies were evaluated across four user studies, of which three were lab-based (two offline and one online).

[S1](#) Filtering *nudge* of to remove results with high privacy risk.

[S2](#) Re-ranking *nudge* to place results with lowest privacy risk at the top and higher privacy threats deeper in rank.

[S3](#) Stoplight *nudge* with coloured lights indicating levels of privacy risk.

[S4](#) Fact box *boost* to teach a skill for reduced privacy risk.

There were two sub-variants of strategy [S4](#) evaluated in the final user study:

[S4 Small](#) One (of two) variants of the fact box *boost* strategy [S4](#). This strategy displayed a small version of the fact box (Figure 8.4) in the right-hand side of the search system.

[S4 Large](#) One (of two) variants of the fact box *boost* strategy [S4](#). Provided a large and more detailed fact box (Figure 8.3) before the experiment began, and was in addition to the small version in SERP (as in strategy [S4 Small](#)).

The first two lab-based studies (Chapter 5 and 6) focused specifically on the *Nudge* based strategies ([S1](#)- [S3](#)). The final lab based study (Chapter 9), which compared *boosting* and *nudging* against a Control Web search environment, relied

upon findings from a non-lab based user study run on a popular Crowdsourcing platform (Chapter 8) and result of a study to identify useful cues for harm prevention in Web search (Chapter 7).

The first of these studies utilized a highly controlled offline environment (Chapter 5) and was the foundation for the development and refinement of general methods (Chapter 4) used through much of the other user studies. One of the key findings from this study being that strategies [S1](#) and [S2](#) are both highly effective at harm prevention related to privacy in Web search. Another main finding being that for a subset of participants, those taking some form of privacy protective action outside of the lab, the Stoplight strategy [S3](#) also appears to be effective at reducing privacy impacts. There were no significant impacts across these strategies with respect to impacts to task behaviour (i.e. time to complete) and *search outcome* (medical decisions). The conclusions of this study provided some initial answers to [G-RQ-1](#) and [G-RQ-2](#).

The second study (Chapter 6) investigated the same *nudge* strategies, in a much more naturalistic online Web search environment connected to the live web. The study was less controlled, but had a fully interactive environment and therefore permitted a much richer set of analyses. Analyses considered the target harm of privacy impacts, and additionally considered exposure to misinformation as an additional harm. This dual axis of harms considered provided in depth insights related to the three *nudge* strategies, which were not visible with privacy impacts alone. A key finding from this study being that *nudge* strategy [S1](#) is highly effective at privacy reductions, but non-viable due to a quite significant increase of exposure to low quality information. This dual consideration gave quite surprising findings related to the Stoplight system (strategy [S3](#)) as it compared to the Control system (across the participant sample), being that privacy impacts were not significantly reduced

## 10. CONCLUSIONS

---

but exposure to high quality information was increased (in some cases significantly). Other findings were fairly consistent with the offline *nudge* study, most notably that only a subset of users, those who take actions for privacy protection outside of the lab, are the individuals most compliant with the Stoplight approach.

The last study (Chapter 9), compared Control search system with systems designed upon *nudge* strategy [S2] and *boost* strategies [S4 Small] and [S4 Large]. The study was in an offline environment, was much like the offline *nudge* study, in that participants could not enter queries. It was much different from both *nudge* studies (which compared the strategies with a *within-group* design), in that a *between-group* design was used where each participant was randomly assigned to one and only one strategy (or Control group) for the entire study. This key difference in design was necessary as *boosting* strategies are designed to teach skills; where learning is a temporally driven process, and therefore motivated the *between-group* approach as the best means to give time for learning. One key finding from this study indicates that the large fact box *boost* (strategy [S4 Large]) was the only effective of the two *boost* strategies evaluated ([S4 Small] and [S4 Large]). Furthermore, the Re-ranking *nudge* was again found to be highly effective at reduced harms related privacy. The most interesting findings related to **G-RQ-3** were several findings in support of the claims related to *boosting* and *nudging*, where for example, participants treated with strategy [S4 Large] were compliant with the skill taught after the *boost* was removed, whereas participants treated with *nudge* [S2] were non-compliant after removal of the *nudge*.

Fundamental to **G-RQ-3**, which entailed the evaluation of strategies [S4 Small] and [S4 Large] with a *nudge* based approach (strategy [S2] in the current thesis), was the identification of cue(s) which could be translated to heuristics for harm prevention. Chapter 7 introduces the methods used to identify the cues used in



subsequent studies, to first pilot different fact boxes (Chapter 8), and second to compare *nudging* and *boosting* approaches (Chapter 9). While the methods developed and findings related to cue identification (Chapter 7) are important to the overall body of work presented here, they (and to some extent the pilot study to compare fact boxes) are given much less attention in the discussion that follows. However, the methods developed in these studies (Chapters 7 and 8) are relevant for future directions of research.

## 10.2 Discussion

Ultimately, the aim of the thesis was to understand the viability and effectiveness of each of these strategies, and furthermore to evaluate theoretical claims about *Nudging* and *Boosting*. This aim helped formulate the overall research questions, which are restated below.

**G-RQ-1** “*Are the behavioural and cognitive strategies viable for prevention of harm during Web search?*”

**G-RQ-2** “*Which of these behavioural and cognitive strategies are most effective at harm prevention?*”

**G-RQ-3** “*To what extent do the claims about *boosting* compared to *nudging* exist within the experimental search environment?*”

Revisiting these questions in turn are the main focus of the discussion, for which one must keep in mind these questions can only be addressed in the context of the experiments and the specific harms they considered (primarily loss of privacy and secondarily exposure to misinformation).

### 10.2.1 Viability of Strategies (G-RQ-1)

The question around viability G-RQ-1 is, in our view, the most important question with respect to each of the strategies investigated. Reason being, no matter how effective a strategy may be at preventing the harm it is designed to address, if this same strategy results in other issues, issues such as increased exposure to misinformation, degraded user experience, or simply users just don't like the strategy, then the strategy is likely non-viable.

In the case of the strategies evaluated, three of the strategies (S2, S3 and S4) are deemed viable approaches in the context of strategies for reduced harms related to privacy.

The Filtering *nudge* strategy is non-viable, for two reasons. First, and perhaps most concerning, is the strategy greatly increased exposure to low-quality information discovered in the online *nudge* study. Though medical decisions were not significantly impacted with this strategy, this result is likely due to a limitation of the study, being the low number of participants. There is already evidence from previous findings (see findings in [179]) with the search tasks used in our study, that results biased towards lower quality information will result in poorer decisions with potentially grave medical outcomes, and there is no reason to expect why poorer decisions will not occur with strategy S1. This finding alone would be sufficient evidence to avoid using it for prevention of loss of privacy for *3rd party trackers*. But there is also the other finding from the online study that participants really do not like this approach when compared to the other two *nudge* strategies, and is a finding that shows this approach has limited (if any) commercial viability.

The two other *nudge* strategies (S2 and S3) are viable as contrast to the two reasons that Filtering is not viable. With respect to misinformation, though

strategy [S2](#) performs slightly worse than the Control, the differences are minimal (as indicated in the online study) and can likely be improved with some tweaks to the Re-ranking approach (e.g. only Re-rank the first 10 results rather than all 50 returned by the API used). Findings from the Stoplight strategy [S3](#) support Re-ranking in this manner, where it was found that information quality increased in the first 10 results compared to the Control system (see online study Discussion for more specifics), most likely due to the association of Green lights to higher quality information. Both of these strategies (counter to Filtering) were more preferred (especially the Stoplight approach), and therefore may be okay to use in commercial settings.

The *boost* approach, considered two sub-strategies ([S4 Small](#) and [S4 Large](#)), which are certainly both viable from an implementation standpoint. The small fact box (strategy [S4 Small](#)) can certainly be placed in the SERP itself (as our study demonstrated) without having other negative impacts, but it is not effective for reductions to privacy impacts for the participants sample used in the evaluation. Turning to the other *boost* approach (strategy [S4 Large](#)), as the discussion in the Chapter 9 indicates, the large fact box is not fit for purpose in a commercial search engine, as it takes up too much space. To reiterate our suggestion from this study, we see this approach most viable in education focused settings, such as education campaigns in schools or through browser based plug-ins for teaching skills.

Several major limitations were also raised with respect to the *boost* vs. *nudge* study that suggest one must use caution related to the viability of the *boost* strategies. First, data collection had to be cut short due to the Covid-19 pandemic, therefore a more limited analysis was performed. Though data was collected for self-report measures, encounters with misinformation and search task decisions, these analyses were not performed due to the low sample size. Additionally, the questions

## 10. CONCLUSIONS

---

related to preferences for each strategy were incorrectly entered and therefore could not be determined.

### 10.2.2 Effectiveness of Strategies (G-RQ-2)

In the studies covered, effectiveness was evaluated with respect to two general objective measures, with the primary measures related to reductions in privacy loss due to *3rd party trackers* and the secondary measures related to compliance to the strategy (e.g. visiting .org and .gov sites) significantly more so than the Control.

There is not much else to say with respect to the primary measures (e.g. mean number of *3rd party trackers*) that was not already said in study specific results and discussions. Non-transparent strategies (S1 and S2) are highly restrictive with this matter compared to the Control system. Where as the transparent Stoplight strategy (S3) demonstrated no global effect, but did show some effects for users that do take privacy actions in their daily lives. Similarly, the transparent *boost* strategy (only strategy S4 Large) showed a small positive impact compared to the Control. These findings, of course could have turned out entirely different in the instance other harms (e.g. hate speech reduction) were made the primary focus, which is a limitation of the studies.

Measures related to compliance were an important factor overlooked in the initial offline *nudge* study, but investigated heavily in the later two lab based studies. In the online *nudge* study, the measures were based upon compliance to the Stoplight colours (and related risks), where as the final study (comparing *nudging* and *boosting*) focused on compliance to the TLD risks in the Fact box. Evaluation of strategy compliance was motivated by the possibility of type II errors with the analyses of the primary privacy measures, due limitations of the sample size and evaluation

test sets. Analysis of compliance was particularly important for the transparent strategies (S3 and S4), as in both cases their primary measures did not indicate harm reductions. However, the less granular compliance measures used in both studies indicate that participants were being significantly more compliant with these strategies when compared to the Control.

### 10.2.3 *Nudging* vs. *Boosting* (G-RQ-3)

Much was already covered (see Chapter 9 and Section 9.4) with respect to investigating the claims related to *nudging* and *boosting*. Related to the measures used in the study, especially the measures specific to compliance with the strategy, the findings are in line with what the theory states. *Nudging* and *boosting* both work well when the intervention is turned on (e.g. results are Re-ranked or a fact box is displayed). After the interventions were removed, the participants that were treated with the *boost* continued to demonstrate compliance to the intervention (i.e. continued to visit .org and .gov) more so than the Control, and the participants treated with strategy S2 returned to performance in line with the Control. Therefore, a finding demonstrating that users (at least some) were enabled with a new decision making competency, which is in line with findings from the earlier pilot study.

However, there were many limitations in this study, which were detailed in the study specific discussion, of which some are worth re-highlighting here. Limitations mainly caused by the already mentioned incomplete sample as a result of the Covid-19 pandemic, a limitation which greatly raises the possibility of both type I and type II errors. Furthermore, the study compared a transparent *boost* with a non-transparent ‘classic’ *nudge*, where a comparison against a transparent ‘educational’ *nudge* is an important comparison one should make in the future. Additionally, strategy S4 Small, suggested to be much more viable in commercial settings than strategy

[S4 Large](#), showed no positive or negative effect in our sample when compared to the Control.

### 10.2.4 Validity of Studies

The general methods chapter (Chapter 4) included motivation for the use of lab based studies, specifically the highly controlled environment they offer. A key challenge of this approach is that all of the experiments had a partially artificial component to them, thus validity of the findings presented must be framed within the constraints of the studies.

One artificial component to highlight is in the offline *nudging* and *boosting* studies. Participants were instructed to imagine they had performed a search for the medical task, as they could not enter queries (a static set of results were provided). This is not how search tasks are performed in the wild and one must certainly be cautious with extrapolation of findings from these studies more broadly.

The online *nudge* study permitted users to enter queries, thus no instructions were necessary for them to imagine they had entered a query. Nonetheless, participants had no choice with respect to the search tasks provided, whereas in a non-lab setting users perform their search tasks based upon a naturalistic information need.

All participants across all studies took part because they chose to do so, which means that the sample is non-representative of Web searchers in general. This is one disadvantage of the ethical requirements for informed consent academic research, as opposed to Web platforms (e.g. Google, Facebook) which can run an A/B test [96] on a random sample of users without telling them. Experiments on Web platforms are ecologically valid because participants are never even aware they took part in an experiment.

### 10.2.5 Other Limitations

As with any study, there are many limitations one must consider, in addition to those mentioned above.

A *within-group* design was used in both *nudge* studies, as it allowed for more participants assigned to each strategy. This reduced the number of search tasks assigned to each strategy to two. This was not so problematic in the offline evaluation the [Offline Nudging Test Set](#), as the data collected still met assumptions for the statistical tests. However, the noisier online [Online Nudging Test Set](#) was less well behaved with respect to assumptions (e.g. homoscedasticity) for parametric statistical tests. As a result, we performed analyses with non-parametric tests in addition to parametric tests. Both type I and type II errors are therefore more likely with the findings reported in the online *nudge* study when compared to findings in the other empirical studies.

Furthermore, eye tracking was not considered for any of the studies, which would be a useful method for measuring participant efforts (and underlying motivating factors) with respect to strategies [S3](#), [S4 Large](#) and [S4 Small](#).

With respect to measuring encounters with misinformation, the annotation methodology to classify the quality of information (e.g. *correct*, *incorrect*) was limited in that annotators were graduate students, not medical professionals (e.g. medical doctors, researchers with the Cochrane review board).

Specific to the privacy measures used, tools that give better coverage of to identify leaks of data to 3rd parties were considered during study design, however all tools considered would have greatly degraded the search experience. For example, the [Selenium automated browser](#) library (tested in the design phase of the experiment)

## 10. CONCLUSIONS

---

indicated that with parallelisation (4 pages loaded in parallel), at best we could expect the first 10 results could be opened in no less than 10 seconds (often times many more seconds). Given the length of time, such approaches were determined to be infeasible for a user study, and lookup tables containing *3rd party tracker* data were used instead.

The background section also introduced several other cognitive strategies (e.g. Technocognition) which were not considered in our current studies. As the choice of comparisons was driven by the digital human rights focus of the overall project supporting this thesis, a comparison was made between the highly ethical *boost* with the less ethical *nudge*.

### 10.3 Avenues for Future Work

There are many directions one could head with our framework in future work. Here we highlight what would come next if given the chance.

#### 10.3.1 *Nudging* vs. *Boosting*

An obvious choice are analyses of data collected with a larger participant sample in the *nudge* versus *boost* study. Furthermore, a study comparing the two approaches could (and should) be run in an online environment similar to the online *nudge* study, as the data analysis would be much richer and allow for greater insights. A comparison of a transparent ‘educative’ *nudge* (e.g. the Stoplight) with a *boost* would also be a natural step forward.



### 10.3.2 Harms and Strategies

In the current thesis, we focused on the harms related to privacy, finding that participants have a broad set of beliefs and practices related to privacy on the Web. It therefore would also be wise to consider other harms. In the introduction, other harms of Web search including misinformation, climate impacts and hateful speech were just some examples to consider in addition to privacy. Furthermore, just like privacy, other harms are likely to be more or less important dependent on the context (e.g. searcher preferences and the domain of the search task) and overall priority based upon societal norms.

There are a number of different strategies one might develop and evaluate. Findings from the study (Chapter 7) identified TLDs as a useful feature for both misinformation and privacy impacts (we designed strategies specific to privacy). In future research, the findings on TLDs suggest a different strategy for misinformation, where individuals use majority voting to determine the correct answer for the search tasks used (i.e. the answer to the Cochrane medical question is determined when two out of three results visited agree). In a similar vein, fact boxes may not be the most appropriate method, Fast and Frugal Trees (a type of *boost*) are one alternative strategy to consider.

There are potentially many different strategies one might discover through investigation of behavioural logs captured in naturalistic settings. Several notable studies have identified strategies in the logs including: identification of links between known cognitive models of risk preference (prospect theory) [77], expert searcher behaviour [22], and behaviour indicating the usage of cues for credibility assessments ([94]). All of the studies mentioned here are a template for discovering new strategies. Furthermore, these studies identified skills that could be transferred to people that do

## 10. CONCLUSIONS

---

not have them (a *boost*), however only the expert searcher investigation actually attempted this [22].

Returning to the extensive set of cues highlighted by Smith and Rieh [208], there is certainly opportunity for novel information extraction techniques (e.g. privacy impact metrics gathered for all Web sites in a search index). Addressing these opportunities opens the door to identification of more strategies, as well as presentation of cues in the search environment itself (e.g. with an information nutrition label *nudge* [74]).

Purely algorithmic strategies, though seemingly the focus of much research at present, should continue to be investigated and improved. Specific to privacy, suggested approaches, such as those to minimize data collection [28], are just one important area of interest.

### 10.3.3 Different Contexts

When one considers the many search contexts, there are again many different directions for research.

In the current thesis, we focused on medical search tasks with known answers. Search task, as introduced in the background, is a very broad topic in and of itself. Exploratory and navigational search are areas yet to be considered. Furthermore, different domains of task outside of medical search also have the potential for harms, for instance domains such as finance and law.

Also, we investigated Web search in the context of a search environment simulating a search engine. Other examples where information is searched for include mobile applications, social media platforms and conversational assistants. These are

all contexts allowing one to evaluate strategies through a different lens.

## 10.4 Closing Remarks

We introduced a framework (Chapter 3) as a pathway to reduce the risk of harms present in modern Web search. Central to the framework are behavioural and cognitive science based decision making tools along with three further components: policy, system design and overall evaluation. In the current set of studies, we made use of existing and study specific evaluation metrics to evaluate and compare the two popular decision making paradigms (*nudging* and *boosting*) in the context of Web search and the proposed framework. The findings are promising, but only a beginning. Further investigations may eventually lead to real world applications aimed at addressing some of the bigger problems at present in Web search.

Baeza-Yates' recent commentary on the interactions between IR systems and searchers [16] as a cause of harms in Web search systems, is direct real-world evidence of the pervasive nature of the current Web search setup. Implementing one (or more) of the strategies evaluated in our studies are a possible pathway to improving the overall nature of Web search and the related risks. Given current algorithms and their ability to learn from log data, hypothetically it would only require a subset of users concerned about harms in Web search to shift the results for everyone else (i.e. a positive externality). The environment could naturally evolve to something more protective for individuals and society as a whole.

## 10. CONCLUSIONS

---

# References

- [1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. 2017. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Comput. Surv.* 50, 3, Article Article 44 (August 2017), 41 pages.
- [2] Alessandro Acquisti, Leslie K John, and George Loewenstein. 2013. What Is Privacy Worth? *The Journal of Legal Studies* 42, 2 (2013), 249–274.
- [3] Aman Agarwal, Kenta Takatsu, Ivan Zaitsev, and Thorsten Joachims. 2019. A General Framework for Counterfactual Learning-to-Rank. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*.
- [4] Eugene Agichtein, Walt Askew, and Yandong Liu. 2008. Combining Lexical, Syntactic, and Semantic Evidence for Textual Entailment Classification.. In *TAC*.
- [5] Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. 2018. Intent-aware Query Obfuscation for Privacy Protection in Personalized Web Search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 285–294.
- [6] Johannes Aigner, Amelie Durcharadt, Thiemo Kersting, Markus Kattenbeck, and David Elweiler. 2017. Manipulating the Perception of Credibility in Refugee Related Social Media Posts. In *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval (CHIIR '17)*. ACM, 297–300.
- [7] Mofleh Al-diabat. 2016. Detection and Prediction of Phishing Websites using Classification Mining Techniques. *International Journal of Computer Applications* 147, 5 (2016).
- [8] A. Alhindi, U. Kruschwitz, C. Fox, and M-D. Albakour. 2015. Profile-Based Summarisation for Web Site Navigation. *ACM Transactions on Information Systems (TOIS)* 33, 1, Article 4 (March 2015), 39 pages.
- [9] Julia Angwin, Charlie Savage, Jeff Larson, Henrik Moltke, Laura Poitras, and James Risen. 2015. AT&T Helped U.S. Spy on Internet on a Vast Scale. *The New York Times* (2015). (Accessed on 06/2019).
- [10] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 4 (2008), 555–596.
- [11] Leif Azzopardi. 2011. The Economics in Interactive Information Retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. 15–24.
- [12] Leif Azzopardi. 2014. Modelling Interaction with Economic Models of Search. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '14)*. ACM, 3–12.

## REFERENCES

---

- [13] Leif Azzopardi, Diane Kelly, and Kathy Brennan. 2013. How Query Cost Affects Search Behavior. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. ACM, 23–32.
- [14] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the Utility of Search Engine Result Pages: An Information Foraging Based Measure. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '18)*. 605–614.
- [15] Leif Azzopardi and Guido Zuccon. 2016. An Analysis of the Cost and Benefit of Search Interactions. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (ICTIR '16)*. 59–68.
- [16] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (2018), 54–61.
- [17] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 2011. *Modern Information Retrieval*. Vol. 463.
- [18] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.
- [19] David Bamman, Chris Dyer, and Noah A Smith. 2014. Distributed Representations of Geographically Situated Language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 828–834.
- [20] Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Propopy: Organizing the news based on their propagandistic content. *Information Processing & Management* 56, 5 (2019), 1849–1864.
- [21] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 54–63.
- [22] Scott Bateman, Jaime Teevan, and Ryen W. White. 2012. The Search Dashboard: How Reflection and Comparison Impact Search Behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, 1785–1794.
- [23] Marcia J Bates. 1989. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review* 13, 5 (1989), 407–424.
- [24] Nicholas J. Belkin. 2008. Some(What) Grand Challenges for Information Retrieval. *SIGIR Forum* 42, 1 (2008), 47–54.
- [25] Susan Benesch. 2017. Civil Society Puts a Hand on the Wheel: Diverse Responses to Harmful Speech. *Harmful Speech Online* (2017), 31.
- [26] Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Counterspeech on Twitter: A Field Study. *The Dangerous Speech Project* (2016).
- [27] Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. 2013. Inferring the demographics of search users: Social data meets search queries. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. ACM, 131–140.
- [28] Asia J. Biega, Peter Potash, Hal Daumé, Fernando Diaz, and Michèle Finck. 2020. Operationalizing the Legal Principle of Data Minimization for Personalization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, 399–408.
- [29] Asia J. Biega, Rishiraj Saha Roy, and Gerhard Weikum. 2017. Privacy through Solidarity: A User-Utility-Preserving Framework to Counter Profiling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. Association for Computing Machinery, 675–684.
- [30] Sarah Bird, Vikas Mishra, Steven Englehardt, Rob Willoughby, David Zeber, Walter Rudametkin, and Martin Lopatka. 2020. Actions speak louder than words: Semi-supervised learning for browser fingerprinting detection. [arXiv:cs.CR/2003.04463](https://arxiv.org/abs/cs.CR/2003.04463)

## REFERENCES

- [31] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
- [32] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 4349–4357.
- [33] Rachel Botsman. 2017. Big Data meets Big Brother as China moves to rate its citizens. *Wired UK* 21 (2017).
- [34] Dale E. Brashers and Timothy P. Hogan. 2013. The appraisal and management of uncertainty: Implications for information-retrieval systems. *Information Processing and Management* 49, 6 (2013), 1241–1249. <https://doi.org/10.1016/j.ipm.2013.06.002>
- [35] Joel Breakstone, Mark Smith, Sam Wineburg, Amie Rapaport, Jill Carle, Marshall Garland, and Anna Saavedra. 2019. Students' civic online reasoning: A national portrait. *Stanford History Education Group & Gibson Consulting* (2019). <https://purl.stanford.edu/gf151tb4868>
- [36] Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (2002), 3–10.
- [37] Norman R. Brown. 2002. Real-world estimation: Estimation modes and seeding effects. *Psychology of Learning and Motivation*, Vol. 41. Academic Press, 321–359.
- [38] Pete Burnap and Matthew L. Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science* 5, 1 (2016), 11. <https://doi.org/10.1140/epjds/s13688-016-0072-6>
- [39] Carole Cadwalladr and Emma Graham-Harrison. 2018. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian* 17 (2018), 22. <https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election>
- [40] Juan Pablo Carrascal, Christopher Riederer, Vijay Erramilli, Mauro Cherubini, and Rodrigo de Oliveira. 2013. Your Browsing Behavior for a Big Mac: Economics of Personal Information Online. In *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*. 189–200.
- [41] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE, 9–16.
- [42] Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. 2015. Misleading Online Content: Recognizing Clickbait as "False News". In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection (WMDD '15)*. Association for Computing Machinery, New York, NY, USA, 15a–19.
- [43] Sophie Chesney, Maria Liakata, Massimo Poesio, and Matthew Purver. 2017. Incongruent Headlines: Yet Another Way to Mislead Your Readers. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*. 56–61.
- [44] Laurie Clarke. 2019. Your online shopping will soon require more than just a bank card. <https://www.wired.co.uk/article/online-shopping-psd2-strong-customer-authentication>. *Wired UK* (2019).
- [45] Cyril W. Cleverdon. 1991. The Significance of the Cranfield Tests on Index Languages. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '91)*. Association for Computing Machinery, New York, NY, USA, 3a–12.
- [46] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. 2020. Threats of a Replication Crisis in Empirical Computer Science. *Commun. ACM* 63, 8 (2020), 70a–79.
- [47] Anne Cocos and Chris Callison-Burch. 2017. The Language of Place: Semantic Value from Geospatial Context. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (EACL 2017)*, 99–104.

## REFERENCES

---

- [48] Kate Conger and Davey Alba. 2020. Twitter Refutes Inaccuracies in Trump’s Tweets About Mail-In Voting - The New York Times. <https://www.nytimes.com/2020/05/26/technology/twitter-trump-mail-in-ballots.html>. (Accessed on 06/14/2020).
- [49] Alissa Cooper. 2008. A Survey of Query Log Privacy-Enhancing Techniques from a Policy Perspective. *ACM Trans. Web 2, 4*, Article 19 (2008), 27 pages.
- [50] W. Bruce Croft. 2019. The Importance of Interaction for Information Retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’19)*. ACM, New York, NY, USA, 1–2.
- [51] W Bruce Croft, Donald Metzler, and Trevor Strohman. 2015. *Search Engines: Information Retrieval in Practice*. Pearson Education.
- [52] Thomas H. Davenport and D.J. Patil. 2012. Data Scientist: The Sexiest Job of the 21st Century. *Harvard business review* 90, 5 (2012), 70–76.
- [53] Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *arXiv preprint arXiv:1703.04009* (2017).
- [54] Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Scientific Reports* 6 (2016).
- [55] Michela Del Vicario, Fabiana Zollo, Guido Caldarelli, Antonio Scala, and Walter Quattrociocchi. 2017. Mapping social dynamics on Facebook: The Brexit debate. *Social Networks* 50 (2017), 6–16.
- [56] Brenda Dervin. 1998. Sense-making theory and practice: an overview of user interests in knowledge seeking and use. *Journal of Knowledge Management* 2, 2 (1998), 36–46.
- [57] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [58] Fernando Diaz. 2016. Worst Practices for Designing Production Information Access Systems. *ACM SIGIR Forum* 50, 1 (2016), 2–11.
- [59] Shiri Dori-Hacohen and James Allan. 2015. Automated Controversy Detection on the Web. In *European Conference on Information Retrieval*. Springer, 423–434.
- [60] Doteveryone. 2018. People, Power and Technology: The 2018 Digital Understanding Report. (2018). [https://doteveryone.org.uk/wp-content/uploads/2019/07/Doteveryone\\_PeoplePowerTechDigitalUnderstanding2018.pdf](https://doteveryone.org.uk/wp-content/uploads/2019/07/Doteveryone_PeoplePowerTechDigitalUnderstanding2018.pdf)
- [61] David Elweiler, Christoph Trattner, and Morgan Harvey. 2017. Exploiting Food Choice Biases for Healthier Recipe Recommendation. In *Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’18)*. ACM, 575–584.
- [62] Steven Englehardt and Arvind Narayanan. 2016. Online Tracking: A 1-million-site Measurement and Analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS ’16)*. ACM, 1388–1401.
- [63] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W. Felten. 2015. Cookies That Give You Away: The Surveillance Implications of Web Tracking. In *Proceedings of the 24th International Conference on World Wide Web (WWW ’15)*. 289–299.
- [64] Rob Ennals, Beth Trushkowsky, and John Mark Agosta. 2010. Highlighting Disputed Claims on the Web. In *Proceedings of the 19th International Conference on World Wide Web (WWW ’10)*. Association for Computing Machinery, New York, NY, USA, 341–350. <https://doi.org/10.1145/1772690.1772726>
- [65] Robert Epstein and Ronald E. Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (2015), E4512–E4521.



## REFERENCES

---

- [66] European Union. April. General Data Protection Regulation (GDPR) - Official Legal Text. <https://gdpr-info.eu/> (Accessed on 04/02/2018).
- [67] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rightsbased Approaches to Principles for AI. *Berkman Klein Center Research Publication* (2020).
- [68] B. J. Fogg. 2003. Prominence-Interpretation Theory: Explaining How People Assess Credibility Online. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems (CHI EA '03)*. 722 – 723.
- [69] Imane Fouad, Nataliaia Bielova, Arnaud Legout, and Natasa Sarafjanovic-Djukic. 2020. Missed by Filter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels. *Proceedings on Privacy Enhancing Technologies* 2020, 2 (2020), 499–518.
- [70] Floyd J Fowler Jr. 2013. *Survey research methods*. Sage publications.
- [71] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human-Computer Interaction* 35, 6 (2019), 456–467.
- [72] Wai Tat Fu and Herre van Oostendorp. 2020. Introduction. In *Understanding and Improving Information Search*. Springer, 1–9.
- [73] Norbert Fuhr. 2017. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *ACM SIGIR Forum* 51, 2 (2017).
- [74] Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Jarvelin, Rosie Jones, Yiqun Liu, Josiane Mothe, Wolfgang Nejdl, Isabella Peters, and Benno Stein. 2017. An Information Nutritional Label for Online Documents. *ACM SIGIR Forum* 51, 3 (2017), 46–66.
- [75] Benjamin CM Fung, Ke Wang, Rui Chen, and Philip S Yu. 2010. Privacy-Preserving Data Publishing: A Survey of Recent Developments. *Comput. Surveys* 42, 4 (2010), 1–53.
- [76] Elisa Gabbert. 2019. The 3 Types of Search Queries & How You Should Target Them. <https://www.wordstream.com/blog/ws/2012/12/10/three-types-of-search-queries>. (Accessed on 04/18/2020).
- [77] Yingqiang Ge, Shuyuan Xu, Shuchang Liu, Zuohui Fu, Fei Sun, and Yongfeng Zhang. 2020. Learning Personalized Risk Preferences for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 409–418.
- [78] Charulata Ghosh and Matthew S. Eastin. 2020. Understanding Users' Relationship with Voice Assistants and How It Affects Privacy Concerns and Information Disclosure Behavior. In *HCI for Cybersecurity, Privacy and Trust*. Springer International Publishing, Cham, 381–392.
- [79] Gerd Gigerenzer. 2008. Why Heuristics Work. *Perspectives on Psychological Science* 3, 1 (2008), 20–29.
- [80] Gerd Gigerenzer. 2015. On the Supposed Evidence for Libertarian Paternalism. *Review of philosophy and psychology* 6, 3 (2015), 361–383.
- [81] Gerd Gigerenzer and Henry Brighton. 2009. Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science* 1, 1 (2009), 107–143.
- [82] Gerd Gigerenzer and Daniel G Goldstein. 1996. Reasoning the Fast and Frugal Way: Models of Bounded Rationality. *Psychological Review* 103, 4 (1996), 650.
- [83] Gerd Gigerenzer, Ralph Hertwig, and Thorsten Pachur (Eds.). 2011. *Heuristics: The Foundations of Adaptive Behavior*. Oxford University Press.

## REFERENCES

---

- [84] Daniel G Goldstein and Gerd Gigerenzer. 2002. Models of Ecological Rationality: The Recognition Heuristic. *Psychological review* 109, 1 (2002), 75.
- [85] Google. [n.d.]. Search Education - Google. <https://www.google.com/insidesearch/searcheducation/>. (Accessed on 06/14/2020).
- [86] Peter C. Götzsche. 2019. *Death of a Whistleblower and Cochrane's Moral Collapse*. People's Press.
- [87] Mark A Graber, Donna M D'Alessandro, and Jill Johnson-West. 2002. Reading level of privacy policies on Internet health Web sites. *Journal of Family Practice* 51, 7 (2002), 642–646.
- [88] Till Grüne-Yanoff and Ralph Hertwig. 2016. Nudge Versus Boost: How Coherent are Policy and Theory? *Minds and Machines* 26, 1 (01 Mar 2016), 149–183.
- [89] Matthias Hagen, Martin Potthast, Michel Büchner, and Benno Stein. 2015. Webis: An ensemble for twitter sentiment detection. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. 582–589.
- [90] Juho Hamari, Jonna Koivisto, and Tuomas Pakkanen. 2014. Do Persuasive Technologies Persuade? - A Review of Empirical Studies. In *Persuasive Technology*. Springer International Publishing, 118–136.
- [91] Marti Hearst. 2009. *Search User Interfaces*. Cambridge University Press.
- [92] Ralph Hertwig. 2017. When to consider boosting: some rules for policy-makers. *Behavioural Public Policy* 1, 2 (2017), 143–161.
- [93] Ralph Hertwig and Till Grüne-Yanoff. 2017. Nudging and Boosting: Steering or Empowering Good Decisions. *Perspectives on Psychological Science* 12, 6 (2017), 973–986.
- [94] Brian Hilligoss and Soo Young Rieh. 2008. Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management* 44, 4 (2008), 1467–1484.
- [95] Chris Hoffman. 2012. *The Difference Between .com, .net, .org and Why We're About To See Many More Top-Level Domains*. <https://www.howtogeek.com/126670/the-difference-between-.com-.net-.org-and-why-were-about-to-see-many-more-top-level-domains/>
- [96] Katja Hofmann, Lihong Li, and Filip Radlinski. 2016. Online Evaluation for Information Retrieval. *Foundations and Trends® in Information Retrieval* 10, 1 (2016), 1–117.
- [97] Eric Horvitz and Deirdre Mulligan. 2015. Data, privacy, and the greater good. *Science* 349, 6245 (2015), 253–255.
- [98] Dirk Hovy. 2015. Demographic Factors Improve Classification Performance.. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 752–762.
- [99] Matthew B. Hoy. 2018. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical Reference Services Quarterly* 37, 1 (2018), 81–88.
- [100] Hugo C Huurdeman and Jaap Kamps. 2020. Designing Multistage Search Systems to Support the Information Seeking Process. In *Understanding and Improving Information Search*. Springer, 113–137.
- [101] Samuel Jeong, Nina Mishra, Eldar Sadikov, and Li Zhang. 2012. Domain Bias in Web Search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. ACM, 413–422.
- [102] Peter Ingwersen. 1996. Cognitive Perspectives of Information Retrieval Interaction: Elements of A Cognitive IR Theory. *Journal of Documentation* 52, 1 (1996), 3–50.
- [103] Peter Ingwersen and Kalervo Järvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context*. Vol. 18. Springer Science & Business Media.

## REFERENCES

---

- [104] Ganesh J, Manish Gupta, and Vasudeva Varma. 2016. Doc2Sent2Vec: A Novel Two-Phase Approach for Learning Document Representation. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, New York, NY, USA, 809–812. <https://doi.org/10.1145/2911451.2914717>
- [105] Bernard J Jansen, Danielle L Booth, and Amanda Spink. 2008. Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management* 44, 3 (2008), 1251–1266.
- [106] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446.
- [107] Stephen P Jenkins, Lorenzo Cappellari, Peter Lynn, Annette Jäckle, and Emanuela Sala. 2006. Patterns of consent: evidence from a general household survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169, 4 (2006), 701–722.
- [108] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data As Implicit Feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*. ACM, 154–161.
- [109] Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender.. In *Proceedings of the 19th Conference on Computational Language Learning (CoNLL)*. 103–112.
- [110] Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. 2007. “I Know What You Did Last Summer”: Query Logs and User Privacy. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (CIKM '07)*. 909–914.
- [111] Dan Jurafsky and James H Martin. 2000. *Speech and Language Processing*. Pearson.
- [112] Dan M Kahan. 2012. Cultural cognition as a conception of the cultural theory of risk. In *Handbook of risk theory*. Springer, 725–759.
- [113] Dan M Kahan. 2013. A Risky Science Communication Environment for Vaccines. *Science* 342, 6154 (2013), 53–54.
- [114] Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Macmillan.
- [115] Daniel Kahneman and Amos Tversky. 1984. Choices, values, and frames. *American psychologist* 39, 4 (1984), 341.
- [116] Yvonne Kammerer and Saskia Brand-Gruwel. 2020. Trainings and Tools to Foster Source Credibility Evaluation During Web Search. In *Understanding and Improving Information Search*. Springer, 213–243.
- [117] Arjaldo Karaj, Sam Macbeth, Rémi Berson, and Josep M Pujol. 2018. WhoTracks. Me: Shedding light on the opaque world of online tracking. *arXiv preprint arXiv:1804.08959* (2018).
- [118] Jaana Kekäläinen and Kalervo Järvelin. 2002. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology* 53, 13 (2002), 1120–1129.
- [119] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.
- [120] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. 2015. Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In *Proceedings of the 2015 International Conference on the Theory of Information Retrieval (ICTIR '15)*. ACM, 101–110.
- [121] Jeonghyun Kim. 2009. Describing and Predicting Information-Seeking Behavior on the Web. *Journal of the American Society for Information Science and Technology* 60, 4 (2009), 679–693.
- [122] Jason Kincaid. 2010. EdgeRank: The secret sauce that makes Facebook’s news feed tick. *TechCrunch, April* (2010).
- [123] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

## REFERENCES

---

- [124] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805. <https://doi.org/10.1073/pnas.1218772110> arXiv:<http://www.pnas.org/content/110/15/5802.full.pdf>
- [125] Anastasia Kozyreva, Stephan Lewandowsky, and Ralph Hertwig. In Press 2020. Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools. *Psychological Science in the Public Interest* (In Press 2020).
- [126] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790. <https://doi.org/10.1073/pnas.1320040111> arXiv:<http://www.pnas.org/content/111/24/8788.full.pdf>
- [127] Udo Kruschwitz. 2005. *Intelligent Document Retrieval: Exploiting Markup Structure*. Vol. 17. Springer Science & Business Media.
- [128] Udo Kruschwitz, Charlie Hull, et al. 2017. Searching the Enterprise. *Foundations and Trends® in Information Retrieval* 11, 1 (2017), 1–142.
- [129] Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2016. Freshman or fresher? quantifying the geographic variation of internet language. (2016), 615–618.
- [130] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 102 (2018), 31 pages.
- [131] Michael D. Lee, Natasha Loughlin, and Ingrid B. Lundberg. 2002. Applying one reason decision-making: the prioritisation of literature searches. *Australian Journal of Psychology* 54, 3 (2002), 137–143.
- [132] Pedro Giovanni Leon, Blase Ur, Yang Wang, Manya Sleeper, Rebecca Balebako, Richard Shay, Lujo Bauer, Mihai Christodorescu, and Lorrie Faith Cranor. 2013. What matters to users?: factors that affect users’ willingness to share information with online advertisers. In *Proceedings of SOUPS*. ACM, 7.
- [133] Stephan Lewandowsky, Ullrich K.H. Ecker, and John Cook. 2017. Beyond Misinformation: Understanding and Coping with the “Post-Truth” Era. *Journal of Applied Research in Memory and Cognition* 6, 4 (2017), 353–369.
- [134] Timothy Libert. 2015. Privacy Implications of Health Information Seeking on the Web. *Commun. ACM* 58, 3 (2015), 68–77.
- [135] Steven L Lima. 1998. Stress and decision-making under the risk of predation: recent developments from behavioral, reproductive, and ecological perspectives. *Advances in the Study of Behaviour* 27, 8 (1998), 215–290.
- [136] Halden Lin, Shobhit Hathi, Aishwarya Nirmal, Matthew Conlen, Fred Hohman, and lilianliang. 2020. The Hidden Cost of Digital Consumption. (October 2020). Issue 02. <https://parametric.press/issue-02/streaming/>
- [137] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.
- [138] Bin Liu, Anmol Sheth, Udi Weinsberg, Jaideep Chandrashekar, and Ramesh Govindan. 2013. AdReveal: improving transparency into online targeted advertising. In *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*. ACM, 12.
- [139] Tie-Yan Liu. 2009. Learning to Rank for Information Retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.
- [140] Philipp Lorenz-Spreen, Stephan Lewandowsky, Cass R. Sunstein, and Ralph Hertwig. 2020. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour* (2020).
- [141] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one* 14, 8 (2019).

## REFERENCES

---

- [142] Mary Madden, Lee Rainie, Kathryn Zickuhr, Maeve Duggan, and Aaron Smith. 2014. Public Perceptions of Privacy and Security in the Post-Snowden Era. *Pew Research Center* 12 (2014).
- [143] Carin Magnhagen. 1991. Predation risk as a cost of reproduction. *Trends in Ecology & Evolution* 6, 6 (1991), 183–186. [https://doi.org/10.1016/0169-5347\(91\)90210-0](https://doi.org/10.1016/0169-5347(91)90210-0)
- [144] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Vol. 1. Cambridge University Press.
- [145] Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (2006), 41–46.
- [146] Gary Marchionini, Gary Geisler, and Ben Brunk. 2000. Agileviews: A Human-Centered Framework for Interfaces to Information Spaces. In *Proceedings of the Annual Conference of the American Society for Information Science*. 271–280.
- [147] Ninya Maubach, Janet Hoek, and Damien Mather. 2014. Interpretive front-of-pack nutrition labels. Comparing competing recommendations. *Appetite* 82 (2014), 67–77.
- [148] David Martin Maxwell. 2019. *Modelling Search and Stopping in Interactive Information Retrieval*. Ph.D. Dissertation. University of Glasgow.
- [149] Jonathan R Mayer and John C Mitchell. 2012. Third-party web tracking: Policy and technology. In *2012 IEEE Symposium on Security and Privacy*. IEEE, 413–427.
- [150] Michelle McDowell, Felix G. Rebitschek, Gerd Gigerenzer, and Odette Wegwarth. 2016. A Simple Tool for Communicating the Benefits and Harms of Health Interventions: A Guide for Creating a Fact Box. *MDM Policy & Practice* 1, 1 (2016).
- [151] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.
- [152] Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. 2012. Adding Semantics to Microblog Posts. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. Association for Computing Machinery, 563a–572.
- [153] Fernando Melo and Bruno Martins. 2017. Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Transactions in GIS* 21, 1 (2017), 3–38.
- [154] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*. ACM, 1–8.
- [155] Bhaskar Mitra and Nick Craswell. 2017. Neural Models for Information Retrieval. arXiv:cs.IR/1705.01509
- [156] Bhaskar Mitra and Nick Craswell. 2018. *An Introduction to Neural Information Retrieval*.
- [157] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.
- [158] Neema Moraveji, Daniel Russell, Jacob Bien, and David Mease. 2011. Measuring Improvement in User Search Performance Resulting from Optimal Search Tips. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. Association for Computing Machinery, 355–364.
- [159] Blake Murdoch, Stuart Carr, and Timothy Caulfield. 2016. Selling falsehoods? A cross-sectional study of Canadian naturopathy, homeopathy, chiropractic and acupuncture clinic website claims relating to allergy and asthma. *BMJ Open* 6, 12 (2016). <https://doi.org/10.1136/bmjopen-2016-014028> arXiv:https://bmjopen.bmj.com/content/6/12/e014028.full.pdf
- [160] G Craig Murray and Jaime Teevan. 2007. Query Log Analysis: Social and Technological Challenges. In *ACM SIGIR Forum*, Vol. 41. ACM New York, NY, USA, 112–120.
- [161] Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. 2016. Computational Sociolinguistics: A Survey. *Computational Linguistics* 42, 3 (2016), 537–593.

## REFERENCES

---

- [162] Jakob Nielsen and Rolf Molich. 1990. Heuristic Evaluation of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90)*. 249–256.
- [163] Patricia A Norberg, Daniel R Horne, and David A Horne. 2007. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of consumer affairs* 41, 1 (2007), 100–126.
- [164] Alamir Novin and Eric Meyers. 2017. Making Sense of Conflicting Science Information: Exploring Bias in the Search Engine Result Page. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. ACM, 175–184.
- [165] Heather L O'Brien and Lori McCay-Peet. 2017. Asking Good Questions: Questionnaire Design and Analysis in Interactive Information Retrieval Research. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. ACM, 27–36.
- [166] Brendan O'Connor, Jacob Eisenstein, Eric P Xing, and Noah A Smith. 2010. Discovering demographic language variation. In *Proceedings of the NIPS Workshop on Machine Learning for Social Computing*.
- [167] Adam Oliver. 2015. Nudging, Shoving, and Budging: Behavioural Economic-Informed Policy. *Public Administration* 93, 3 (2015), 700–714.
- [168] Christopher Olston and Marc Najork. 2010. Web Crawling. *Foundations and Trends in Information Retrieval* 4, 3 (2010), 175–246.
- [169] Amy Orben. 2020. Teenagers, screens and social media: a narrative review of reviews and key studies. *Social Psychiatry and Psychiatric Epidemiology* (2020), 1–8.
- [170] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web*. Technical Report. Stanford InfoLab.
- [171] Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2, 1-2 (2008), 1–135.
- [172] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, Vol. 10. Association for Computational Linguistics, 79–86.
- [173] Eli Pariser. 2011. *The filter bubble: What the Internet is Hiding From You*. Penguin UK.
- [174] Mauricio Perez, Sandra Avila, Daniel Moreira, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. 2017. Video pornography detection through deep learning techniques and motion information. *Neurocomputing* 230 (2017), 279–293.
- [175] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. *arXiv preprint arXiv:1802.05365* (2018).
- [176] Lawrence Phillips, Kyle Shaffer, Dustin Arendt, Nathan Hodas, and Svitlana Volkova. 2017. Intrinsic and Extrinsic Evaluation of Spatiotemporal Text Representations in Twitter Streams. *ACL 2017* (2017), 201–210.
- [177] Peter Pirolli and Stuart Card. 1999. Information Foraging. *Psychological Review* 106, 4 (1999), 643–675.
- [178] Peter Pirolli and Stuart Card. 2005. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. In *Proceedings of International Conference on Intelligence Analysis*, Vol. 5. McLean, VA, USA, 2–4.
- [179] Frances A Pogacar, Amira Ghenai, Mark D Smucker, and Charles LA Clarke. 2017. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '17)*. ACM, 209–216.

## REFERENCES

---

- [180] Martin F Porter. 1980. An algorithm for suffix stripping. *Program* 14, 3 (1980), 130–137.
- [181] Privacy International. 2019. Your mental health for sale. How websites about depression share data with advertisers and leak depression test results. (2019). <https://privacyinternational.org/sites/default/files/2019-09/Your%20mental%20health%20for%20sale%20-%20Privacy%20International.pdf>
- [182] R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>
- [183] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. ACM, 117–126.
- [184] Samuli Reijula and Ralph Hertwig. 2020. Self-Nudging and the Citizen Choice Architect. *Behavioural Public Policy* (2020), 1–31. <https://doi.org/10.1017/bpp.2020.5>
- [185] CarrieLynn D. Reinhard and Brenda Dervin. 2012. Comparing situated sense-making processes in virtual worlds: Application of Dervin’s Sense-Making Methodology to media reception situations. *Convergence* 18, 1 (2012), 27–48.
- [186] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook. In *Recommender systems handbook*. Springer, 1–35.
- [187] Giuseppe Rizzo and Raphaël Troncy. 2012. NERD: a framework for unifying named entity recognition and disambiguation extraction tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EARS '12)*. 73–76.
- [188] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *preprint arXiv:1701.08118* (2017).
- [189] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.
- [190] Daniel Russell and Mario Callegaro. 2019. How to be a Better Web Searcher: Secrets from Google Scientists Researchers who study how we use search engines share common mistakes, misperceptions and advice. *Scientific American* (2019). <https://blogs.scientificamerican.com/observations/how-to-be-a-better-web-searcher-secrets-from-google-scientists/>
- [191] Ian Ruthven. 2008. Interactive Information Retrieval. *Annual Review of Information Science and Technology* 42, 1 (2008), 43–91.
- [192] Tetsuya Sakai. 2016. Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, New York, NY, USA, 5–14.
- [193] Tetsuya Sakai. 2020. On Fuhr’s Guideline for IR Evaluation. In *ACM SIGIR Forum*, Vol. 54. ACM, 14–22.
- [194] Gerard Salton. 1968. Automatic information organization and retrieval. (1968).
- [195] Aaron Sankin. 2020. Want to Find a Misinformed Public? Facebook’s Already Done It. <https://themarkup.org/coronavirus/2020/04/23/want-to-find-a-misinformed-public-facebooks-already-done-it>. (Accessed on 04/30/2020).
- [196] Tefko Saracevic. 1996. Relevance reconsidered. In *Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2)*. ACM New York, 201–218.
- [197] Tefko Saracevic. 1997. The stratified model of information retrieval interaction: Extension and applications. In *Proceedings of the annual meeting-american society for information science*, Vol. 34. 313–327.
- [198] Reijo Savolainen. 2018. Berrypicking and information foraging: Comparison of two theoretical frameworks for studying exploratory search. *Journal of Information Science* 44, 5 (2018), 580–593.

## REFERENCES

---

- [199] Wolfgang Schulz. 2018. Regulating Intermediaries to Protect Privacy Online - the Case of the German NetzDG. *Personality and Data Protection Rights on the Internet* (2018).
- [200] Julia Schwarz and Meredith Morris. 2011. Augmenting Web Pages and Search Results to Support Credibility Assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, 1245–1254.
- [201] Leslie Scism. [n.d.]. New York Insurers Can Evaluate Your Social Media Use - If They Can Prove Why It's Needed. *The Wall Street Journal* (January [n.d.]). <https://www.wsj.com/articles/new-york-insurers-can-evaluate-your-social-media-use-if-they-can-prove-why-its-needed-11548856802> (Accessed on 02/2019).
- [202] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.
- [203] Michael Siegrist, Rebecca Leins-Hess, and Carmen Keller. 2015. Which front-of-pack nutrition label is the most efficient one? The results of an eye-tracker study. *Food Quality and Preference* 39 (2015), 183–190.
- [204] Fabrizio Silvestri. 2010. Mining Query Logs: Turning Search Usage Data into Knowledge. *Foundations and Trends in Information Retrieval* 4, 1-2 (2010), 1–174.
- [205] Herbert A Simon. 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics* 69, 1 (1955), 99–118.
- [206] Paul Slovic. 2016. *The perception of risk*. Routledge.
- [207] Catherine L. Smith and Paul B. Kantor. 2008. User Adaptation: Good Results from Poor Systems. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 147–154.
- [208] Catherine L. Smith and Soo Young Rieh. 2019. Knowledge-Context in Search Systems: Toward Information-Literate Actions. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*. ACM, New York, NY, USA, 55–62.
- [209] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 254–263.
- [210] Karen Spärck Jones. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation* 28, 1 (1972), 11–21.
- [211] Karen Spärck Jones. 2003. Privacy: what's different now? *Interdisciplinary Science Reviews* 28, 4 (2003), 287–292.
- [212] P. A. Stout, J. Villegas, and H. Kim. 2001. Enhancing learning through use of interactive tools on health-related websites. *Health Education Research* 16, 6 (2001), 721–733.
- [213] Fred Stutzman, Robert Capra, and Jamila Thompson. 2011. Factors mediating disclosure in social network sites. *Computers in Human Behavior* 27, 1 (2011), 590–598.
- [214] Jessica Su, Aneesh Sharma, and Sharad Goel. 2016. The Effect of Recommendations on Network Structure. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. 1157–1167.
- [215] Matt Summers. 2020. Facebook isn't free: zero-price companies overcharge consumers with data. *Behavioural Public Policy* (2020), 1–25.
- [216] Cass R Sunstein. 2016. *The Ethics of Influence: Government in the Age of Behavioral Science*. Cambridge University Press.
- [217] Krysta M. Svore, Qiang Wu, Chris J. C. Burges, and Aaswath Raman. 2007. Improving Web Spam Classification Using Rank-time Features. In *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '07)*. ACM, 9–16.



## REFERENCES

---

- [218] Briony Swire-Thompson and David Lazer. 2020. Public Health and Online Misinformation: Challenges and Recommendations. *Annual Review of Public Health* 41, 1 (2020), 433–451.
- [219] Jaime Teevan, Susan T Dumais, and Eric Horvitz. 2005. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*. ACM, 449–456.
- [220] Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin.
- [221] The United Nations. 1948. Universal Declaration of Human Rights.
- [222] The United Nations General Assembly. 1966. International Covenant on Civil and Political Rights. *Treaty Series* 999 (December 1966), 171.
- [223] The United Nations General Assembly. 1966. International Covenant on Economic, Social, and Cultural Rights. *Treaty Series* 999 (December 1966), 171.
- [224] Clive Thompson. 2006. Google’s China Problem (and China’s Google Problem). *The New York Times* (2006). (Accessed on 06/2020).
- [225] Eran Toch, Yang Wang, and Lorrie Faith Cranor. 2012. Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 203–220.
- [226] Vincent Toubiana, Arvind Narayanan, Dan Boneh, Helen Nissenbaum, and Solon Barocas. 2010. Adnostic: Privacy Preserving Targeted Advertising. In *Proceedings Network and Distributed System Symposium*.
- [227] Minh Tran, Xinshu Dong, Zhenkai Liang, and Xuxian Jiang. 2012. Tracking the Trackers: Fast and Scalable Dynamic Analysis of Web Content for Privacy Violations. In *International Conference on Applied Cryptography and Network Security*. Springer, 418–435.
- [228] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 32–41.
- [229] Raphaël Troncy. 2003. Integrating Structure and Semantics into Audio-visual Documents. In *The Semantic Web - ISWC 2003*. Springer Berlin Heidelberg, 566–581.
- [230] Zeynep Tufekci. 2015. Facebook said its algorithms do help form echo chambers, and the tech press missed it. *New Perspectives Quarterly* 32, 3 (2015), 9–12.
- [231] Zeynep Tufekci. 2017. We’re building a dystopia just to make people click on ads. <https://www.ted.com/>
- [232] Andrew H. Turpin and William Hersh. 2001. Why Batch and User Evaluations Do Not Give the Same Results. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*. Association for Computing Machinery, New York, NY, USA, 225–231.
- [233] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131.
- [234] Kazutoshi Umemoto, Takehiro Yamamoto, and Katsumi Tanaka. 2020. Search Support Tools. In *Understanding and Improving Information Search*. Springer, 139–160.
- [235] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. 2012. Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. In *Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS '12)*. ACM, Article 4, 4:1–4:15 pages.
- [236] U.S. Government Publishing Office. 1996. 47 USC 230: Protection for private blocking and screening of offensive material. <https://www.govinfo.gov/content/pkg/USCODE-2011-title47/pdf/USCODE-2011-title47-chap5-subchapII-partI-sec230.pdf>. (Accessed on 06/2020).

## REFERENCES

---

- [237] Erica Van Herpen and Hans CM Van Trijp. 2011. Front-of-pack nutrition labels. Their effect on attention and choices when consumers have varying goals and time constraints. *Appetite* 57, 1 (2011), 148–160.
- [238] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272.
- [239] Vivienne Waller. 2011. Not Just Information: Who Searches for What on the Search Engine Google? *Journal of the American Society for Information Science and Technology* 62, 4 (2011), 761–775.
- [240] Zeerak Waseem. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the 1st Workshop on Natural Language Processing and Computational Social Science*. 138–142.
- [241] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of NAACL-HLT*. 88–93.
- [242] Ingmar Weber and Carlos Castillo. 2010. The demographics of web search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*. ACM, 523–530.
- [243] Ingmar Weber, Venkata Rama Kiran Garimella, and Erik Borra. 2012. Mining Web Query Logs to Analyze Political Issues. In *Proceedings of the 4th Annual ACM Web Science Conference (WebSci '12)*. ACM, 330–334.
- [244] Ryen White. 2013. Beliefs and biases in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. ACM, 3–12.
- [245] Ryen W White. 2016. *Interactions with Search Systems*. Cambridge University Press.
- [246] Ryen W White and Ahmed Hassan. 2014. Content bias in online health search. *ACM Transactions on the Web (TWEB)* 8, 4 (2014), 25.
- [247] Ryen W White and Resa A Roth. 2009. Exploratory Search: Beyond the Query-Response Paradigm. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–98.
- [248] Barbara M. Wildemuth and Luanne Freund. 2009. Search Tasks and Their Role in Studies of Search Behaviors. In *Third Annual Workshop on Human-Computer Interaction and Information Retrieval (HCIR)*.
- [249] Max L Wilson. 2011. Search User Interface Design. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 3, 3 (2011), 1–143.
- [250] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 1330–1340.
- [251] Tom D Wilson. 1981. On User Studies and Information Needs. *Journal of Documentation* 37, 1 (1981), 3–15.
- [252] Tom D Wilson. 1997. Information Behaviour: An Interdisciplinary Perspective. *Information processing & management* 33, 4 (1997), 551–572.
- [253] Tom D Wilson. 1999. Models in Information Behaviour Research. *Journal of Documentation* 55, 3 (1999), 249–270.
- [254] Tom D Wilson. 2000. Human Information Behavior. *Informing Science* 3, 2 (2000), 49–56.
- [255] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. *preprint arXiv:1610.08914* (2017).

## REFERENCES

---

- [256] Kimberly S Young. 2004. Internet Addiction: A New Clinical Phenomenon and Its Consequences. *American Behavioral Scientist* 48, 4 (2004), 402–415.
- [257] Zhonghao Yu, Sam Macbeth, Konark Modi, and Josep M Pujol. 2016. Tracking the trackers. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, 121–132.
- [258] Michael Zimmer. 2008. Privacy on Planet Google: Using the Theory of Contextual Integrity to Clarify the Privacy Threats of Google's Quest for the Perfect Search Engine Google: An Intersection of Business and Technology. *Journal of Business & Technology Law* 3 (2008), 109–126.
- [259] Michael Zimmer. 2010. *Web Search Studies: Multidisciplinary Perspectives on Web Search Engines*. Springer Netherlands, 507–521.
- [260] Steven Zimmerman. 2018. Exploring Potential Pathways to Address Bias and Ethics in IR. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. Association for Computing Machinery, 1445.
- [261] Steven Zimmerman, Chris Fox, and Udo Kruschwitz. 2018. Improving Hate Speech Detection with Deep Learning Ensembles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (Miyazaki, Japan) (LREC 2018)*.
- [262] Steven Zimmerman, Stefan Herzog, David Elsweiler, Jon Chamberlain, and Udo Kruschwitz. 2020. Towards a Framework for Harm Prevention in Web Search. In *Bridging the Gap between Information Science, Information Retrieval and Data Science - BIRDS*.
- [263] Steven Zimmerman and Udo Kruschwitz. 2017. Speaking of the weather: Detection of meteorological influences on sentiment within social media. In *Computer Science and Electronic Engineering (CEECE), 2017*. IEEE, 1–6.
- [264] Steven Zimmerman, Alistair Thorpe, Jon Chamberlain, and Udo Kruschwitz. 2020. Towards Search Strategies for Better Privacy and Information. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval (CHIIR '20)*. Association for Computing Machinery, 124–134.
- [265] Steven Zimmerman, Alistair Thorpe, Chris Fox, and Udo Kruschwitz. 2019. Investigating the Interplay Between Searchers' Privacy Concerns and Their Search Behavior. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. ACM, 953–956.
- [266] Steven Zimmerman, Alistair Thorpe, Chris Fox, and Udo Kruschwitz. 2019. Privacy Nudging in Search: Investigating Potential Impacts. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*. 283–287.
- [267] Shoshana Zuboff. 2015. Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology* 30, 1 (01 Mar 2015), 75–89.
- [268] Frederik J Zuiderveen Borgesius, Damian Trilling, Judith Moeller, Balázs Bodó, Claes H de Vreese, and Natali Helberger. 2016. Should we worry about filter bubbles? *Internet Policy Review* 5, 1 (2016), 1–16.

## REFERENCES

---

# Appendix A

## General Methodology (Appendices)

### A.1 Questionnaires and Scales

#### A.1.1 Pre/post-task

A set of pre-task and post-task (Table A.1) questions were given to participants during each search task in the experiment. These questions were used to formulate metrics outlined in general methods Section 4.4.4.3

**Table A.1:** Two questions were given before and after each search task in the lab studies (both offline and online). Questionnaire item numbers and language used to capture pre-task knowledge of the search task and post-task confidence on their decision are included along with Likert scales used. Items are grouped by pre-task and post-task questions.

Item		Language Used
1	Pre-Task	How knowledgeable are you of this health issue?
2	Pre-Task	What level of knowledge do you have in regards to this treatment?
	Scale	1 = I have no knowledge 7 = I have expert knowledge
1	Post-Task	How confident are you in your decision about the effectiveness of the treatment for the medical condition?
	Scale	1 = I am very unconfident in my decision 7 = I am very confident in my decision
2	Post-Task	To what extent are you certain about your decision?
	Scale	1 = I am very uncertain about my decision 7 = I am very certain about my decision

## A. GENERAL METHODOLOGY (APPENDICES)

---

Experiment progress:

**MEDICAL QUESTION 2: Does surgery help obesity?**  
**HEALTH ISSUE: obesity** – Obesity is a complex disorder involving an excessive amount of body fat. Obesity isn't just a cosmetic concern. It increases your risk of diseases and health problems, such as heart disease, diabetes and high blood pressure. *Source: Mayo Clinic*  
**TREATMENT: surgery** – a branch of medicine concerned with diseases and conditions requiring or amenable to operative or manual procedures *Source: Merriam-Webster*

**Questions (Understanding your prior knowledge of treatment and health issue):**

**How knowledgeable are you of this health issue?**

1 I have no knowledge     2     3     4 Neither/Nor     5     6     7 I have expert knowledge

**What level of knowledge do you have in regards to this treatment?**

1 I have no knowledge     2     3     4 Neither/Nor     5     6     7 I have expert knowledge

**Figure A.1:** Interface used to ask pre-task and post task questions. In the screen shot the pre-task questions are visible.

### A.1.2 Privacy Attitudes

The following tables (A.2 - A.3) contain questions asked to participants during offline and online lab based studies. These questions were used to formulate metrics introduced in general methods Section 4.4.4.4 related to individual attitudes related to privacy.

### A.1.3 Privacy Protective Behaviors / Actions

The following tables (A.4 - A.6) contain questions asked to participants during offline and online lab based studies related to individual actions taken to protect privacy. These questions were used to formulate metrics introduced in general methods Section 4.4.4.4.

## A.1 Questionnaires and Scales

**Table A.2:** Questionnaire item numbers and language used to capture general attitudes towards privacy. Likert scales used varied across the questions (which are grouped appropriately).

Item	Language Used
1	How likely do you think it is that personal information submitted / shared on the internet will be: Shared or sold to others
2	Used by others to harm you
3	Use virtual private networks (VPNs) when viewing information on the internet
Scale	1 = Not at all Likely 7 = Very Likely
4	I believe that in giving personal information to online: The damage that could be caused by a data security breach is
5	The likelihood of a data security breach is
Scale	1 = Very Low 7 = Very High
6	In general, how worried are you about your personal privacy?
Scale	1 = Not Worried At All 7 = Very Worried
7	Different private and public organizations have personal information about us. How concerned are you about whether or not they keep this information confidential?
Scale	1 = Not Concerned At All 7 = Very Concerned

**Table A.3:** Questionnaire item numbers and language used to capture attitudes towards privacy in the domain of health information. These questions were asked for other information domains (e.g. Age, Location, Religion) but those data were not included in the analysis. Likert scales used were consistent for all three questions.

Item	Language Used
1	How concerned are you about the sensitivity of the following <u>personal information</u> that you share?
2	How concerned are you about your privacy when sharing the following <u>personal information</u> ?
3	How concerned are you about personal damage when sharing the following <u>personal information</u> ?
Scale	1 = Not Concerned At All 7 = Very Concerned

**Table A.4:** Questionnaire items, labels, and questionnaire language (6 point scale where 0=I have never heard of this, 1=Never, 2=Sometimes, 3>About half the time, 4=Most of the time, 5=Always). Items 1 - 10 were asked in all lab based studies. Items 11 - 14, denoted by underlined questions were developed after the initial offline *nudge* study (Chapter 5).

Item	Label	Language "When using the internet, how regularly do you do the following?"
1	Anon. Com.	Use anonymous communications networks (e.g. Tor)
2	Encrypted	Use end-to-end encryption tools for messaging (e.g. Signal)
3	3rd Party	Run browser extensions to block 3rd party tracking cookies (e.g. Ghostery, Privacy Badger)
4	VPN	Use virtual private networks (VPNs) when viewing information on the internet
5	Fingerprint	Run software to prevent browser fingerprinting (e.g. uBlock, Privacy Badger)
6	Cookie Del.	Delete your browsing cookies automatically (with software)
7	HTTPS	Use software to ensure HTTPS communications with websites
8	Javascript	Disable javascript in your browser
9	Cookie Dis.	Disable cookies in your browser
10	Anti-virus	Have anti-virus software installed on your devices (e.g. Norton, Sophos)
11	<u>Do Not Track</u>	Use the 'do not track' feature of my browser
12	<u>Incognito</u>	Use the 'incognito' / 'anonymous' mode of my browser
13	<u>Read State. 1</u>	Read the privacy policies of websites I visit
14	<u>Read State. 2</u>	Review privacy statement updates sent to me by applications I use

## A. GENERAL METHODOLOGY (APPENDICES)

---

**Table A.5:** In a post experiment questionnaire given to participants during all lab based studies the following question was asked. "*When using the internet, how frequently do you use the following Web browsers?*" Privacy enhancing browsers were defined as any browser listed on the [PrivacyTools](#) resource Website.).

Item	Label	Privacy Enhancing?
1	Chrome	No
2	Internet Explorer	No
3	Safari	No
4	Mozilla Firefox	Yes
5	Tor	Yes
6	Brave	Yes

**Table A.6:** In a post experiment questionnaire given to participants during all lab based studies the following question was asked. "*When using the internet, how frequently do you use the following search engines?*" Privacy enhancing browsers were defined as any search engine listed on the [PrivacyTools](#) resource Website.).

Item	Label	Privacy Enhancing?
1	Google	No
2	Bing	No
3	Yahoo	No
4	Baidu	No
5	Yandex	No
6	Duck Duck Go	Yes
7	Qwant	Yes

## A.2 Latin and Graeco-Latin Square Design

To ensure a balanced design for the within-group online and offline *nudge* studies, we follow the methodology of Pogacar et al. [179] who created a 10 x 10 Graeco-Latin matrix to provide randomization and balance of search task and system treatments. The first step to creation of this matrix was building three pre-cursor Latin squares: one each for the *helpful* and *does not help* search tasks (Tables A.7 and A.8 respectively) along with one (see Table A.9 containing the three experimental treatments (S1- S3) + Control treatment + Baseline treatment (task without search results). Following the procedures of [179], we overlaid the Latin square of the systems with each of the *helpful* and *does not help* Latin squares, resulting in two 5x5 Graeco-Latin squares for these task-system pairs. Finally, as with [179], we randomized the



## A.2 Latin and Graeco-Latin Square Design

---

columns and rows of the Graeco-Latin overlays twice to produce the 10x10 matrix used in our studies (see Table [A.10](#)).

T2	T5	T6	T8	T9
T5	T6	T8	T9	T2
T6	T8	T9	T2	T5
T8	T9	T2	T5	T6
T9	T2	T5	T6	T8

**Table A.7:** Latin Squares table of all search tasks classified as *helpful*. Full definitions of the tasks are found in methods Table [4.1](#)

T1	T3	T4	T7	T10
T3	T4	T7	T10	T1
T4	T7	T10	T1	T3
T7	T10	T1	T3	T4
T10	T1	T3	T4	T7

**Table A.8:** Latin Squares table of all search tasks classified as *does not help*. Full definitions of the tasks are found in methods Table [4.1](#)

Control	Filter	Baseline	Re-ranking	Stoplight
Baseline	Re-ranking	Stoplight	Control	Filter
Stoplight	Control	Filter	Baseline	Re-ranking
Filter	Baseline	Re-ranking	Stoplight	Control
Re-ranking	Stoplight	Control	Filter	Baseline

**Table A.9:** Latin Squares table of the 5 experimental treatments for the online and offline *nudge* studies. Full descriptions of the variants are found in methods section [4.2.2](#).

## A. GENERAL METHODOLOGY (APPENDICES)

---

T7-Re	T2-Co	T5-Fi	T8-Re	T4-Ba	T6-Ba	T9-St	T3-Fi	T10-St	T1-Co
T3-Ba	T8-St	T2-Fi	T4-Re	T6-Re	T9-Co	T7-St	T5-Ba	T1-Fi	T10-Co
T4-St	T8-Co	T9-Fi	T6-St	T7-Co	T3-Re	T1-Ba	T5-Re	T2-Ba	T10-Fi
T5-St	T6-Co	T10-Ba	T2-Re	T8-Fi	T3-St	T9-Ba	T1-Re	T7-Fi	T4-Co
T9-Re	T4-Fi	T7-Ba	T1-St	T3-Co	T2-St	T5-Co	T8-Ba	T6-Fi	T10-Re
T2-Re	T10-Fi	T9-Fi	T3-Ba	T8-Co	T5-Ba	T1-Re	T6-St	T4-St	T7-Co
T6-Co	T1-Ba	T9-Re	T3-St	T5-St	T7-Fi	T10-Re	T4-Co	T8-Fi	T2-Ba
T2-St	T6-Fi	T5-Co	T9-Ba	T7-Re	T1-St	T8-Re	T10-Ba	T3-Co	T4-Fi
T8-Ba	T10-St	T4-Re	T5-Fi	T9-St	T6-Re	T1-Co	T7-Ba	T3-Fi	T2-Co
T2-Fi	T7-St	T3-Re	T1-Fi	T10-Co	T5-Re	T8-St	T4-Ba	T6-Ba	T9-Co

**Table A.10:** Graeco Latin Square Design - Task and Experimental Variants used in the online and offline *nudge* studies. A 10 x 10 Graeco-Latin squares matrix showing the different variants used in the experiments. Ba = Baseline, Co = Control, Fi = Filtering Re = Re-ranking and St = Stoplight.

### A.3 Supplemental to Online Nudge Study

#### Sending Queries to Commercial API and Linkage to Privacy Impacts

All queries submitted in the SERPs (3 experimental + control) were sent to the [Microsoft Azure Cognitive Services Bing API](#). In the query request, we lower cased all queries for normalization, we set the market to the country (United Kingdom) where our experiment was run, and requested the maximum number of results (which was 50 at time of our experiment). Per API documentation, we received related searches and spelling suggestions in the API response but did not use these in our experiment (a possible avenue for future research). It is noteworthy that the Bing API automatically corrects spelling in the query (e.g. a query for ‘catt videos’ would return results for ‘cat videos’), we accepted the corrections as another form of query normalization.

Once results were retrieved by the Bing API, we used a lookup table to retrieve the number of privacy trackers for each result. Privacy trackers were available for many (but not all) of the results. We calculated descriptive statistics for the subset of results where the number of privacy trackers was known during the experiment

### A.3 Supplemental to Online Nudge Study

---

(e.g. if 22 of 50 results were returned from the Bing API could be linked to privacy trackers, we would calculate descriptive statistics for the 22 results). We recorded the maximum number of privacy trackers for all results in the subset. We also calculated the median and upper quartiles for number of trackers in the subset, which were used as demarcations for low risk, medium risk and high risk to personal privacy (this is the same method used in our offline studies). We also marked each result where no tracker was available (using the same example, there would be 28 websites of 50 that did not have trackers). At this point, all results could be updated appropriately dependent on the SERP variant the subject was given. Descriptions of the different SERPs and how these results were displayed are further outlined below. For procedures to produce the lookup table for *3rd party trackers* and explanation of why *3rd party trackers* are not available for all results, see Section 4.3.1.

It is worth noting that query response times, the time between when the user submitted a query and when results were displayed to them, were well under 1 second. This was the case for all experimental SERP variants.