

INVESTIGATING AUTOMATION RELIANCE: THREE EXPERIMENTS

LARA SURACI

PhD Supervisors: Robin Cubitt & Chris Starmer

*Thesis submitted to The University of Nottingham
for the Degree of Doctor of Philosophy*

SCHOOL OF ECONOMICS
UNIVERSITY OF NOTTINGHAM

30 December 2023



**University of
Nottingham**
UK | CHINA | MALAYSIA

Acknowledgements

In the four years, two months, and two or so weeks that it took me to plan, write and submit this thesis, I never had any doubt that this part would be by far the easiest one to write. I am so incredibly grateful to all the wonderful people who accompanied me on this journey, and without whom — clichéd as it may sound — this thesis would not exist.

First and foremost, to my supervisors: Chris Starmer and Robin Cubitt— for your expertise, your guidance, and your patience throughout the last few years, with both professional and personal matters. I genuinely could not have asked for a more helpful and insightful duo of mentors for on this journey. You two have provided me with an invaluable perspective on experimental research that combines creativity and precision, and that will stay with me wherever I go next.

To my co-author and friend: Irenaeus Wolff— for believing in me when I did not. Your outside-of-the-box thinking, your candour, and your ever-changing email greetings have challenged me, helped me grow as a researcher, and never failed to make me smile.

It goes without saying that I would not have been able to conduct this research without funding from both the Network for Integrated Behavioural Science (NIBS) and the Centre for Decision Research and Experimental Economics (CeDEx). Money makes the world go ‘round but you cannot put a price tag on the scientific exchange that NIBS and CeDEx have facilitated at conferences, seminars and other events, nor on the ideas sparked and the connections formed.

To the friends I made along the way: Aisha Iqbal, Anna Hochleitner, Candice Ford, Fahad Kazmi, Jose Guinot Saporta, Malte Baader, Monika Pompeo, Marcelo Woo, Richard Mills, Sarah and Kiki Bowen, Yuliet Verbel Bustamante, and Zain Kazmi— for the dinner parties, impromptu dance sessions, coffee dates, play dates, movie nights, long walks and longer conversations.

To the friends from home that, for some inexplicable reason, decided to stick around wherever I go: Alina Görtz, Andi Hecker, Lena Kölle, Katja Meyer, Luisa Kerkhoff, Pia Ribbrock, and Ramona Rosalewski— for giving me a reason to always come back home.

To the friends I found in the unlikeliest of places: Amanda Botström, Angela Wang, Emily Walter, Jen Leggett, Laura Elise Bilhimer, Melissa Lim, Molly McPhee, Raquel Garza, and Robin Luttekes— for making me realise that my academic achievements or, indeed, a lack thereof, have not, do not, and will never define me.

To my partner: Tom Beresford— for coming into my life when I least expected it, for fuelling me with coffee, snacks, and pep talks, and of course for going along with my crazy cat adoption plans. I, too, enjoy your company “more often than not”.

To my second family: Roland Hecker and Ulrike Hecker— for making me feel right at home in your midst from the moment we met.

To my aunts and my nonna: Angela Suraci, Beate Görtz, Domenica Musolino, and Karin Robert— for teaching me, in all of your uniquely different ways, how to navigate the world as a child and as a woman of two cultures.

To my uncle, godfather, and birthday twin: Bernhard Görtz— for countless trips to the airport and for telling the best worst dad jokes in the world.

To my cousin Alexander Görtz— for solving all of my computer troubles since Windows 98 and for your invaluable help with some of the trickier programming aspects of this thesis.

To my mother: Petra Suraci— for teaching me independence, resilience, and strength. We both know I haven’t inherited my ambitious streak from you, but I will be forever grateful for the way you nourished it by answering all my annoying questions as a kid and yet never pushing me to be anything other than your daughter.

And last, but never least, to my father: Domenico Suraci— for all that you sacrificed to see me follow my dreams, even when some of them would end up taking me far away from home. I know this thesis would have taken pride of place in your office, just like all the silly little poems and drawings I scribbled on waiting pads at the restaurant did, a lifetime ago.

CONTENTS

General Introduction	1
I Delegating to a Computer or Delegating to a Human: On Task Sensitivity, Programming Information, and How They Interact	10
1 Introduction	10
2 Experimental Design	13
2.1 Treatment Groups	13
2.2 Delegate's Task	14
2.3 Procedures	18
2.4 Hypotheses	20
3 Results	22
3.1 Delegation Decision	22
3.2 Delegate's Performance	26
3.3 Performance Expectations	28
3.4 Determinants of Automation Reliance	32
4 Discussion	36
II Non-Realised Values and Agency: Positive Framing Can Increase Acceptance of Offers of Automated Assistance No Matter Who Makes the Offer	41
1 Introduction	41
2 Experimental Design	45
2.1 Treatment Groups	45
2.2 The Dot Task	47
2.3 Additional Measurements	49
2.4 Hypotheses	51
3 Results	53
3.1 Automation Reliance Decision	53
3.2 Performance and Performance Expectations	59
3.3 Emotional Reactions and Kindness Perceptions	61
3.4 Determinants of Underreliance	65
4 Discussion and Conclusion	67
III Preferences for Normative over Preferential Algorithms With(out) Feedback: Experimental Evidence from a Repeated Lottery Selection Task	73
1 Introduction	73
2 Experimental Design	76

CONTENTS

2.1	Main Experiment	77
2.1.1	Treatment Groups	78
2.1.2	Lottery Menu Construction	80
2.2	RDU Parameter Estimation	86
2.3	Research Hypotheses	89
3	Results	91
3.1	Main Experiment	92
3.2	Lottery Type Selection	105
3.3	RDU Parameter Estimation	110
4	Discussion and Conclusion	113
	 Appendices	 118
A	Appendix for Chapter 1	118
A.1	Additional Analysis	118
A.1.1	Task Evaluation Survey	118
A.1.2	Main Experiment	126
A.2	Instructions	131
A.2.1	Task Evaluation Survey	131
A.2.2	Main Experiment	136
A.2.3	Task Descriptions	143
B	Appendix for Chapter 2	146
B.1	Additional Analysis	146
B.2	Instructions	164
C	Appendix for Chapter 3	180
C.1	Additional Analysis	180
C.2	Instructions	185
C.2.1	Part 1: RDU Estimation	185
C.2.2	Part 2: Main Experiment	188

LIST OF FIGURES

Figure 1.1	Treatment Groups	14
Figure 1.2	Cohort Structure	18
Figure 1.3	Proportions of Human Delegate Choices	23
Figure 1.4	Mean WTA Values	23
Figure 1.5	Performance: Computer Delegate	27
Figure 1.6	Performance: Human Delegate	27
Figure 1.7	Performance: Human Non-Delegate	27
Figure 1.8	Expected Performance: Computer Delegate (By Delegate's Task)	29
Figure 1.9	Expected Performance: Human Delegate (By Delegate's Task)	29
Figure 1.10	Expected Performance: Computer Delegate (By Delegate Choice)	29
Figure 1.11	Expected Performance: Human Delegate (By Delegate Choice)	30
Figure 1.12	Expected Difference: Computer vs Human Delegate	32
Figure 2.1	Treatment Groups	46
Figure 2.2	Dot Task based on Hollard et al. (2016)	48
Figure 2.3	Proportions of Objective Underreliance on Automation	53
Figure 2.4	Mean WTP for Maintaining Automation Reliance Decision	56
Figure 2.5	Proportions of Subjective Underreliance on Automation	57
Figure 2.6	Mean WTP for Maintaining Automation Reliance Decision	58
Figure 2.7	Performance	59
Figure 2.8	Performance Expectations	60
Figure 2.9	Performance Expectation Differences	61
Figure 2.10	Average Emotions Score (ES) and Kindness Score (KS)	63
Figure 3.1	Sample Lottery Menu	77
Figure 3.2	Treatment Groups	79
Figure 3.3	Lottery Menu Construction (Pattern 1)	81
Figure 3.4	Lottery Menu Construction (Pattern 2)	81
Figure 3.5	Lottery Menu Scrambling Procedure	84
Figure 3.6	EV Comparison of P-Bets and \$-Bets in Lottery Menus	85
Figure 3.7	RDU Parameter Estimation Sample Table	88
Figure 3.8	RDU Parameter Estimation Tables based on Murad et al. (2016)	89
Figure 3.9	Proportions of Automation Reliance Across Rounds	92
Figure 3.10	Number of Choices Favouring Automation Reliance	93
Figure 3.11	Automation Reliance Over Time	98
Figure 3.12	Reasons for Decision in Hypothetical Continuation	104
Figure 3.13	Lottery Type Selection Over Time	107
Figure 3.14	Menu-by-Menu Lottery Type Selection	108
Figure 3.15	Lottery Type Selection (Realised vs. Non-Realised)	109
Figure 3.16	RDU Parameter Estimate Distribution: α	111
Figure 3.17	RDU Parameter Estimate Distribution: β	111
Figure 3.18	RDU Parameter Estimate Distribution: γ	112
Figure 3.19	Probability Weighting Function	112
Figure A.1	Gender Decomposition of Proportions of Human Delegate Choices	127
Figure A.2	Expected Performance: Human Non-Delegate (By Delegate's Task)	129

LIST OF FIGURES

Figure A.3	Expected Performance: Human Non-Delegate (By Delegate Choice)	129
Figure A.4	Expected Difference: Computer Delegate vs Human Non-Delegate	129
Figure A.5	Expected Difference: Human Delegate vs Human Non-Delegate	130
Figure B.1	Performance in BASELINE	147
Figure B.2	Performance in BO-CC	147
Figure B.3	Performance in GO-CC	147
Figure B.4	Performance in BO-EC	148
Figure B.5	Performance in GO-EC	148
Figure B.6	Expected Performance in BASELINE	148
Figure B.7	Expected Performance in BO-CC	149
Figure B.8	Expected Performance in GO-CC	149
Figure B.9	Expected Performance in BO-EC	149
Figure B.10	Expected Performance in GO-EC	150
Figure B.11	Performance Expectation Differences in BASELINE	150
Figure B.12	Performance Expectation Differences in BO-CC	151
Figure B.13	Performance Expectation Differences in GO-CC	151
Figure B.14	Performance Expectation Differences in BO-EC	152
Figure B.15	Performance Expectation Differences in GO-EC	152
Figure B.16	Happy Score	153
Figure B.17	Content Score	153
Figure B.18	Grateful Score	154
Figure B.19	Calm Score	154
Figure B.20	Angry Score	155
Figure B.21	Annoyed Score	155
Figure B.22	Disappointed Score	156
Figure B.23	Sad Score	156
Figure B.24	Kind Score	157
Figure B.25	Fair Score	157
Figure B.26	Friendly Score	158
Figure B.27	Generous Score	158
Figure B.28	Mean Score	159
Figure B.29	Unfair Score	159
Figure B.30	Hostile Score	160
Figure B.31	Uncharitable Score	160
Figure C.1	Number of Algorithmic \$-Bet Selections: $EV(\$-Bet) > EV(P-Bet)$	184
Figure C.2	Number of Algorithmic \$-Bet Selections: $EV(\$-Bet) < EV(P-Bet)$	184

LIST OF TABLES

Table 1.1	Treatment Comparisons of Proportions of Human Delegate Choices	24
Table 1.2	Mean WTA for Swapping Initial Delegate Choice	24
Table 1.3	Comparisons of Performance Expectations	30
Table 1.4	OLS Regressions: Human Delegate Choice	34
Table 1.5	Treatment Comparisons of Performance Expectations	35
Table 2.1	Treatment Comparisons of Objective Underreliance (OUR)	55
Table 2.2	Treatment Comparisons of Subjective Underreliance (SUR)	57
Table 2.3	Mean Emotions Scores and Treatment Comparisons	64
Table 2.4	Mean Kindness Scores and Treatment Comparisons	64
Table 2.5	OLS Regression: Underreliance on Automation	66
Table 3.1	Treatment Comparisons of Automation Reliance: Round 1	94
Table 3.2	OLS Regressions: First Round Automation Reliance	95
Table 3.3	Treatment Comparisons of Automation Reliance: Rounds 2 – 8	96
Table 3.4	Average Round-to-Round Transition Frequencies	99
Table 3.5	Correlations: Automation Reliance, Feedback & Satisfaction	100
Table 3.6	OLS Regressions: Final Round Automation Reliance	102
Table 3.7	Treatment Comparisons of Hypothetical Continuation	103
Table A.1	Demographic Sample Composition (Task Evaluation Survey)	118
Table A.2	Spearman’s Correlations	119
Table A.3	Memory Task Evaluation	120
Table A.4	Prediction Task Evaluation	121
Table A.5	Coordination Task Evaluation	122
Table A.6	CAPTCHA Task Evaluation	123
Table A.7	Anagram Task Evaluation	124
Table A.8	Wilcoxon Signed-Rank p -values for COMPUTER	125
Table A.9	Wilcoxon Signed-Rank p -values for HUMAN	125
Table A.10	Wilcoxon Signed-Rank p -values for COMPUTER > HUMAN	125
Table A.11	Wilcoxon Signed-Rank p -values for PROGRAMMING	125
Table A.12	Demographic Sample Composition (Main Experiment)	126
Table A.13	Demographic Treatment Composition	126
Table A.14	Comparisons of Performance	127
Table A.15	Comparisons of Performance Expectations	128
Table A.16	Differences in Performance Expectations (By Delegate Choice)	128
Table A.17	Treatment Comparisons of Performance Expectations	130
Table B.1	Demographic Sample Composition	146
Table B.2	Demographic Treatment Composition	146
Table B.3	Median Emotion Scores	161
Table B.4	Median Kindness Scores	161
Table B.5	Intercorrelations: Positive Emotional Reactions	161
Table B.6	Intercorrelations: Negative Emotional Reactions	161
Table B.7	Intercorrelations: Positive Kindness Perceptions	162
Table B.8	Intercorrelations: Negative Kindness Perceptions	162
Table B.9	Internal Consistency Reliabilities: Emotions & Kindness Perceptions	162

Table B.10	Correlations: Underreliance & Emotional Reactions	162
Table B.11	Correlations: Underreliance & Kindness Perceptions	162
Table B.12	Risk Preferences (Dohmen et al., 2011)	163
Table B.13	Ambiguity Preferences (Cavatorta & Schröder, 2019)	163
Table C.1	Demographic Sample Composition	180
Table C.2	Demographic Treatment Composition	180
Table C.3	Menu-by-Menu Treatment Comparisons of Automation Reliance	181
Table C.4	OLS Regressions: Hypothetical Automation Reliance	182
Table C.5	Reasons for Decision in Hypothetical Continuation	183

General Introduction

Following recent and ongoing advances in automation and Artificial Intelligence (AI), risk-benefit considerations of automation reliance have emerged as a critical focus of economic research. Automation¹ is fast becoming a ubiquitous feature of many decision environments and, albeit not without caveats, research shows that increasing automation reliance could significantly improve social and economic outcomes in a variety of fields.

Much attention has been directed at the impact of increased automation reliance on the stock (see e.g. Angerer et al., 2023; Goldstein et al., 2014; Kirilenko & Lo, 2013) and labour markets (see e.g. Agrawal et al., 2019a; Lankisch et al., 2019; Webb, 2019). Treleaven et al. (2013) report that, as early as 2011, algorithmic trading accounted for over two thirds of the US stock trading volume, while between one in ten and one in two American jobs are predicted to be at risk of automation replacement in upcoming decades (Nedelkoska & Quintini, 2018; Frey & Osborne, 2017). Without disregarding the associated risks, several studies highlight the potential benefits automation reliance holds for resource investment decisions (Agrawal et al., 2019), high-frequency trading (Liaudinskas, 2022; Cohen, 2022), and the forecasting of political and economic events (Chong et al., 2017; Miyoshi & Matsubara, 2018).

Automation reliance has, however, become an integral part of many consumer decision environments as well: not only do we interact with AI-based tools such as predictive chatbots, personalised ads, or sophisticated navigation systems on a near daily basis, but social companion robots, self-driving cars and other once deemed futuristic concepts are now on the verge of becoming available to the general public.

Crucially, automated agents are often less affected by common behavioural biases than their human counterparts. Recent meta-studies further suggest that, as a result, human-computer interactions are frequently characterised by higher levels of rationality, strategic reasoning, and efficiency (March, 2021), and render humans less prone to emotional and social responses than comparable human-human interactions (Chugunova & Sele, 2022). Conditional on appropriate use and careful regulation, reliance on algorithmic decision-aids in employee hiring processes can increase cost-effectiveness and mitigate discrimination (see e.g. Savage & Bales, 2016; Lepri et al., 2018; Raghavan et al., 2020). In the context of autonomous vehicles, higher levels of automation reliance promise improvements in traffic efficiency and the simultaneous reduction of pollution and accidents (see e.g. Van Arem et al., 2006; Spieser et al., 2014; Gao et al., 2014).

¹Automation as an umbrella term for algorithms, computers, and robots may obscure individual nuances of the different types of automated agents. When referring to other studies, we adopt their terminology; otherwise, however, we use these terms interchangeably as differences between them are not our focus.

Algorithms have furthermore been shown to significantly outperform human forecasters in a variety of settings (see e.g. Sele & Chugunova, 2022; Logg et al., 2019; Grove et al., 2000; Goodwin & Fildes, 1999). In light of these findings, public and private organisations alike could benefit from an in-depth investigation into the determinants and deterrents of automation reliance to help them identify tools and strategies to encourage the appropriate use of automation.

Thus far, empirical research into automation reliance is scarce: in fact, March (2021) concludes that computer players have been used for comparison rather than studied as relevant players in their own right. In addition to this, following the seminal works of Meehl (1954) and Dawes (1979), the replacement of human agents with computer agents has primarily been discussed as an ‘antidote’ to cognitive biases, but the questions of *whether* and *when* humans are prepared to rely on automation remain underexplored.

Previous experimental studies do, however, indicate a preference for human agency or an outright aversion to automation (for an overview, see Burton et al., 2020; Jussupow et al., 2020). In Dzindolet et al. (2002), subjects exhibit a strong self-reliance bias when choosing between performing a visual detection task by themselves or with an automated aid: neither favourable expectations of the performance of the automated aid nor the explicit knowledge of its superiority lead to automation reliance. Gogoll & Uhl (2018) similarly observe an “unrationalisable aversion to automation” for which neither differences in perceived utility nor differences in trust can account: on average, subjects correctly estimate the error rates of computer and human delegates to be equal and exhibit no significant differences in trust between the two, but still prefer the latter to perform a calculation task that governs third-party payoffs. Substantiating these results, Schniter et al. (2020) also rule out differences in trust as an explanation for the tendency to choose human over computer delegates: in their experiment, subjects entrust similar amounts to human and computer trustees in a standard trust game; there are, however, significant differences in social emotions like guilt, gratitude and anger when interacting with the two types of trustees (for further explorations of trust in automation see also Dzindolet et al., 2003; Lee & See, 2004; Parasuraman et al., 2014; Hoff & Bashir, 2015; Lee & Rich, 2021).

Suboptimal automation reliance furthermore appears to extend to applied contexts. In spite of the aforementioned benefits of algorithmic decision-aids in employee hiring processes, we see a persistent resistance to standardised methods of employee selection due to an overvaluation of human intuition and experience (Highhouse, 2008; Diab et al., 2011). Similarly, Bonnefon et al. (2016) find high levels of approval of utilitarian automated vehicles — i.e. vehicles that are programmed to sacrifice their passengers for the greater good — but a preference against *using* such vehicles. Human forecasters in Abeliuk et al. (2020) also rarely adjust their own predictions vis-à-vis superior, computer-generated predictions unless they conform to their priors. A field experiment

in an Inspectional Services department tasked with selecting restaurants for inspection concludes that automation may “provide substantial gains over human prediction [but does] not translate into improved decisions” (Glaeser et al., 2021).

All of this is not to say that there are no examples of appropriate usage or even bias towards automation (see e.g. Dietvorst et al., 2018; Logg et al., 2019; Candrian & Scherer, 2022). It is, however, important to note that this is often contingent on “keeping a human in the loop” (Sele & Chugunova, 2022): humans are sometimes willing to rely on the support of automated agents, but are rarely prepared to cede their decision authority (Chugunova & Sele, 2022). Elkins et al. (2013) suggest that computer-generated recommendations are often perceived as a threat to human experts and thus ignored. In addition to this, in line with Abeliuk et al. (2020) and Glaeser et al. (2021), there is evidence that allowing for human intervention can significantly diminish the returns from automation reliance (Angelova et al., 2023).

A representative survey in the Netherlands finds that humans are, to some degree, aware of the advantages of automation reliance: while respondents express concerns about the societal consequences of algorithmic decision-making in general, evaluations of its potential fairness and usefulness are often equal or even more favourable than evaluations of human decision-making (Araujo et al., 2020). Various studies do, however, recognise that perceptions of and attitudes towards automation reliance are highly contingent on context; including performance-based factors like transparency and failure rates (Cameron et al., 2015), psychological factors like the agreeableness of an automated agent (Kim et al., 2021), as well as situational factors like increased uncertainty about economic outcomes in times of crisis (Ben-David & Sade, 2021). Nevertheless, the underlying causes of suboptimal automation reliance are, at best, partially understood.

Liehner et al. (2021) aptly summarise the resulting challenge of reaping the benefits of past, present, and future advances in automation as follows, “For automation to be beneficial and successfully integrated into society it must first be used and accepted by humans. We can develop as much as we want, what good does it do if nobody uses it?”

Consequently, the aim of this thesis is to shed light on three sets of factors that can influence automation reliance: Chapter 1 explores task sensitivity and the role of information about the programming of an automated agent, Chapter 2 focuses on the influence of non-realised values and the ascription of agency, and Chapter 3 investigates preferences for normative versus preferential algorithms with and without outcome feedback. While these factors represent only a small share of the possible determinants of automation reliance, we show, in three experiments, that each of them is an important puzzle piece on the road to understanding the phenomenon of suboptimal automation reliance.

Chapter 1 builds on evidence that the nature of a task plays a role for the willingness to rely on an automated agent (for an overview, see Mahmud et al., 2022) and explores how the availability of programming information interacts with this. Uncertainty about what a computer is programmed to do — often referred to as its ‘black-box’ character — has been suggested as a possible deterrent for automation reliance (Rai, 2020). However, in a delegation experiment in which subjects had to choose between having a human or a computer delegate perform a task for them, we find that the effect of task sensitivity on delegation preferences can be *mitigated* by the availability of programming information: in particular, programming information decreased automation reliance in a task that was ex-ante perceived to be better suited to computers, but had the reverse effect in a task that was ex-ante perceived to be better suited to humans.

Chapter 2 is motivated by frequent observations of humans treating computers as social actors rather than machines (see e.g. Nass & Moon, 2000; Sandoval et al., 2016; Kim et al., 2021) and investigates how this can affect automation reliance. We designed an experiment in which subjects had to rely either on their own performance or that of a computer, whose success probability was always identical in value but differed in its relation to other, non-realised values. Our results suggest that framing an algorithm as superior to others which might have been available may decrease underreliance on automation. Furthermore, in line with the unconscious application of social behaviours and expectations to humans and computers alike, we find no significant effect of whether agency over the choice of the success probability was ascribed to the computer itself or to the experimenter as its programmer.

In light of the human sensitivity to algorithmic errors (see e.g. Dietvorst et al., 2015; Hoff & Bashir, 2015; Madhavan & Wiegmann, 2007), Chapter 3 complements existing research into automation reliance in decision environments in which such algorithmic errors are objective and independent of individual preferences with an investigation of automation reliance in the preferential choice domain. Specifically, we explore the perception and reception of normative versus preferential algorithms in a repeated lottery choice task with and without outcome feedback. Our results show significantly lower levels of reliance on the preferential algorithm than on its normative counterpart. In addition to this, satisfactory feedback reinforced previous automation reliance or non-reliance, whereas unsatisfactory feedback led to changes in automation reliance decisions; an effect that was significantly stronger for the preferential than for the normative algorithm.

References

- Abeliuk, A., Benjamin, D. M., Morstatter, F., & Galstyan, A. (2020). Quantifying Machine Influence Over Human Forecasters. *Scientific Reports*, 10(1), 1–14.
- Agrawal, A., Gans, J. S., & Goldfarb, A. (2019a). Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction. *Journal of Economic Perspectives*, 33(2), 31–50.
- Agrawal, A., Gans, J. S., & Goldfarb, A. (2019b). Exploring the Impact of Artificial Intelligence: Prediction versus Judgment. *Information Economics and Policy*, 47, 1–6.
- Angelova, V., Dobbie, W. S., & Yang, C. (2023). Algorithmic Recommendations and Human Discretion. Working Paper. *NBER Working Paper Series 31747*, available at: <https://www.nber.org/papers/w31747>.
- Angerer, M., Neugebauer, T., & Shachat, J. (2023). Arbitrage Bots in Experimental Asset Markets. *Journal of Economic Behavior & Organization*, 206, 262–278.
- Araujo, T., Helberger, N., Kruikemeier, S., & De Vreese, C. H. (2020). In AI We Trust? Perceptions About Automated Decision-Making by Artificial Intelligence. *AI & Society*, 35, 611–623.
- Ben-David, D. & Sade, O. (2021). Robo-Advisor Adoption, Willingness to Pay, and Trust — Before and at the Outbreak of the Covid-19 Pandemic. Available at SSRN: <https://ssrn.com/abstract=3361710> or <http://dx.doi.org/10.2139/ssrn.3361710>.
- Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The Social Dilemma of Autonomous Vehicles. *Science*, 352(6293), 1573–1576.
- Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A Systematic Review of Algorithm Aversion in Augmented Decision Making. *Journal of Behavioral Decision Making*, 33(2), 220–239.
- Cameron, D., Aitken, J., Collins, E., Boorman, L., Chua, A., Fernando, S., McAree, O., Martinez Hernandez, U., & Law, J. (2015). Framing Factors: The Importance of Context and the Individual in Understanding Trust in Human-Robot Interaction. Available at White Rose Research Online: <https://eprints.whiterose.ac.uk/91238>.
- Candrian, C. & Scherer, A. (2022). Rise of the Machines: Delegating Decisions to Autonomous AI. *Computers in Human Behavior*, 134, 107308.
- Chong, E., Han, C., & Park, F. C. (2017). Deep Learning Networks for Stock Market Analysis and Prediction: Methodology, Data Representations, and Case Studies. *Expert Systems with Applications*, 83, 187–205.

- Chugunova, M. & Sele, D. (2022). We and It: An Interdisciplinary Review of the Experimental Evidence on How Humans Interact with Machines. *Journal of Behavioral and Experimental Economics*, 99, 101897.
- Cohen, G. (2022). Artificial Intelligence in Trading the Financial Markets. *International Journal of Economics & Business Administration (IJEBA)*, 10(1), 101–110.
- Dawes, R. M. (1979). The Robust Beauty of Improper Linear Models in Decision Making. *American Psychologist*, 34(7), 571–582.
- Diab, D. L., Pui, S.-Y., Yankelevich, M., & Highhouse, S. (2011). Lay Perceptions of Selection Decision Aids in US and Non-US Samples. *International Journal of Selection and Assessment*, 19(2), 209–216.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General*, 144(1), 114.
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The Role of Trust in Automation Reliance. *International Journal of Human-Computer Studies*, 58(6), 697–718.
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors*, 44(1), 79–94.
- Elkins, A. C., Dunbar, N. E., Adame, B., & Nunamaker, J. F. (2013). Are Users Threatened by Credibility Assessment Systems? *Journal of Management Information Systems*, 29(4), 249–262.
- Frey, C. B. & Osborne, M. A. (2017). The Future of Employment: How Susceptible Are Jobs to Computerisation? *Technological Forecasting and Social Change*, 114, 254–280.
- Gao, P., Hensley, R., & Zielke, A. (2014). A Road Map to the Future for the Auto Industry. *McKinsey Quarterly*, Oct, (pp. 1–11).
- Glaeser, E. L., Hillis, A., Kim, H., Kominers, S. D., & Luca, M. (2021). Decision Authority and the Returns to Algorithms. *Harvard Business School Working Paper*.
- Gogoll, J. & Uhl, M. (2018). Rage Against the Machine: Automation in the Moral Domain. *Journal of Behavioral and Experimental Economics*, 74, 97–103.
- Goldstein, M. A., Kumar, P., & Graves, F. C. (2014). Computerized and High-Frequency Trading. *Financial Review*, 49(2), 177–202.
- Goodwin, P. & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making*, 12(1), 37–53.

- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus Mechanical Prediction: A Meta-Analysis. *Psychological Assessment*, 12(1), 19.
- Highhouse, S. (2008). Stubborn Reliance on Intuition and Subjectivity in Employee Selection. *Industrial and Organizational Psychology*, 1(3), 333–342.
- Hoff, K. A. & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors that Influence Trust. *Human Factors*, 57(3), 407–434.
- Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why Are We Averse towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion. *Research Papers (168)*, available at: https://aisel.aisnet.org/ecis2020_rp/168.
- Kim, J., Sah, Y. J., & Song, H. (2021). Agreeableness of a Virtual Agent: Effects of Reciprocity and Need for Help. In *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)* (pp. 1–6).: IEEE.
- Kirilenko, A. A. & Lo, A. W. (2013). Moore’s Law versus Murphy’s Law: Algorithmic Trading and Its Discontents. *Journal of Economic Perspectives*, 27(2), 51–72.
- Lankisch, C., Prettner, K., & Prskawetz, A. (2019). How Can Robots Affect Wage Inequality? *Economic Modelling*, 81, 161–169.
- Lee, J. D. & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1), 50–80.
- Lee, M. K. & Rich, K. (2021). Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness Around Healthcare AI and Cultural Mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–14).
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-Making Processes: The Premise, the Proposed solutions, and the Open Challenges. *Philosophy & Technology*, 31, 611–627.
- Liaudinskas, K. (2022). Human vs. Machine: Disposition Effect Among Algorithmic and Human Day Traders. In *Proceedings of the EUROFIDAI-ESSEC Paris December Finance Meeting 2022*, available at SSRN: <https://ssrn.com/abstract=4260869>.
- Liehner, G. L., Brauner, P., Schaar, A. K., & Ziefle, M. (2021). Delegation of Moral Tasks to Automated Agents — The Impact of Risk and Context on Trusting a Machine to Perform a Task. *IEEE Transactions on Technology and Society*, 3(1), 46–57.
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm Appreciation: People Prefer Algorithmic to Human Judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103.

- Madhavan, P. & Wiegmann, D. A. (2007). Similarities and Differences Between Human-Human and Human-Automation Trust: An Integrative Review. *Theoretical Issues in Ergonomics Science*, 8(4), 277–301.
- Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022). What Influences Algorithmic Decision-Making? A Systematic Literature Review on Algorithm Aversion. *Technological Forecasting and Social Change*, 175, 121390.
- March, C. (2021). Strategic Interactions Between Humans and Artificial Intelligence: Lessons from Experiments with Computer Players. *Journal of Economic Psychology*, 87, 102426.
- Meehl, P. E. (1954). *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis: University of Minnesota.
- Miyoshi, T. & Matsubara, S. (2018). Dynamically Forming a Group of Human Forecasters and Machine Forecasters for Forecasting Economic Indicators. In *IJCAI* (pp. 461–467).
- Nass, C. & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1), 81–103.
- Nedelkoska, L. & Quintini, G. (2018). Automation, Skills Use and Training. *Technical Report OECD iLibrary*, available at: <https://doi.org/10.1787/2e2f4eea-en>.
- Parasuraman, R., de Visser, E., Wiese, E., & Madhavan, P. (2014). Human Trust in Other Humans, Automation, Robots, and Cognitive Agents: Neural Correlates and Design Implications. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 58 (pp. 340–344).
- Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 469–481).
- Rai, A. (2020). Explainable AI: From Black Box to Glass Box. *Journal of the Academy of Marketing Science*, 48, 137–141.
- Sandoval, E. B., Brandstetter, J., Obaid, M., & Bartneck, C. (2016). Reciprocity in Human-Robot Interaction: A Quantitative Approach Through the Prisoner’s Dilemma and the Ultimatum Game. *International Journal of Social Robotics*, 8(2), 303–317.
- Savage, D. D. & Bales, R. (2016). Video Games in Job Interviews: Using Algorithms to Minimize Discrimination and Unconscious Bias. *ABAJ Labour & Employment Law*, 32, 211.

- Schniter, E., Shields, T. W., & Sznycer, D. (2020). Trust in Humans and Robots: Economically Similar But Emotionally Different. *Journal of Economic Psychology*, 78, 102253.
- Sele, D. & Chugunova, M. (2022). Putting a Human in the Loop: Increasing Uptake, But Decreasing Accuracy of Automated Decision-Making. *Max Planck Institute for Innovation & Competition Research Paper No. 22-20*, available at SSRN: <https://ssrn.com/abstract=4285645>.
- Spieser, K., Treleaven, K., Zhang, R., Frazzoli, E., Morton, D., & Pavone, M. (2014). Toward a Systematic Approach to the Design and Evaluation of Automated Mobility-on-Demand Systems: A Case Study in Singapore. In *Road Vehicle Automation* (pp. 229–245).
- Treleaven, P., Galas, M., & Lalchand, V. (2013). Algorithmic Trading Review. *Communications of the ACM*, 56(11), 76–85.
- Van Arem, B., Van Driel, C. J., & Visser, R. (2006). The Impact of Cooperative Adaptive Cruise Control on Traffic-Flow Characteristics. *IEEE Transactions on Intelligent Transportation Systems*, 7(4), 429–436.
- Webb, M. (2019). The Impact of Artificial Intelligence on the Labor Market. *Available at SSRN: https://ssrn.com/abstract=3482150*.

CHAPTER 1

Delegating to a Computer or Delegating to a Human: On Task Sensitivity, Programming Information, and How They Interact[†]

1 Introduction

Past research on the delegation of decisions often reports suboptimal levels of delegation that point towards the existence of non-pecuniary reasons (not) to delegate (see e.g. Owens et al., 2014; Danz et al., 2015; Bobadilla-Suarez et al., 2017). Empirical findings suggest preferences for retaining authority (Fehr et al., 2013), non-interference (Neri & Rommeswinkel, 2017) and an intrinsic valuation of decision rights (Bartling et al., 2014) as possible reasons for the reluctance to choose a human delegate over self-reliance — however, far less is known about delegation preferences when, instead, the alternative to the human delegate is not self-reliance but a computer delegate.

Suboptimal delegation behaviour does, however, appear to carry over into the realm of automation reliance: in fact, several studies suggest a preference for human over computer agency, but nevertheless report that performance expectations, as with human delegates, cannot fully explain the reluctance to adopt computer delegates (see e.g. Dzindolet et al., 2002; Gogoll & Uhl, 2018; Ben-David & Sade, 2021).

Relationships between a human delegator and a human delegate are characterised by the *social risk* born of the inherent uncertainty of handing over one’s decision rights to another person, which has, for instance, been explored in the context of betrayal aversion (Bohnet et al., 2008). Similarly, the much discussed ‘black-box’ character of computers (Christin, 2017; Eastwood et al., 2012) creates a source of uncertainty that might act as a deterrent for automation reliance (Rai, 2020). In fact, extrapolation from personal experience may, in many cases, allow us to form reasonable expectations

[†]This chapter is based on joint work with Robin Cubitt and Chris Starmer. I am grateful to NIBS (Network for Integrated Behavioural Science) for funding this research. I also thank CeDEx members for comments and feedback. The study received ethical approval from the Nottingham School of Economics Research Ethics Committee on 24/03/2021.

about how a human would approach a certain task, but precise information on a computer’s programming is rarely available in real life. Whether or not providing such information could mitigate algorithm aversion¹ and, where appropriate, encourage increased automation reliance therefore warrants further investigation.

To our knowledge, however, most empirical studies into automation reliance do not examine the availability of information on the computer delegate’s programming as a relevant factor in and of itself. If information on the computer delegate’s programming is provided at all, it is often reduced to a specific success probability, which abstracts from the intricacies of the programming that control the computer delegate’s performance. Not only does this not reflect the type of information one would typically receive in the real world, but it also does not account for the possibility that information on a computer delegate’s programming may mediate how the uncertainty about its behaviour compares qualitatively to that of a human delegate.

In a recent review of task-specific algorithm acceptance, Kaufmann et al. (2023) find that 21.30% of the tasks in their sample provide some level of information on the algorithm’s success probability, whereas only 9.00% explain the underlying mechanism. Algorithm appreciation is generally lower for tasks *with* information on the algorithm’s success probability than for those in which no such information is available; this is in line with Burton et al. (2020), who hypothesise that “algorithmically illiterate users” may falsely interpret explicit algorithmic error rates as high relative to unknown human error rates. However, within the small number of tasks that do explain the algorithm’s underlying mechanism, Kaufmann et al. (2023) find evidence of algorithm appreciation.

Previous literature has furthermore established that perceptions about the nature of the delegate’s task also play a key role in the context of automation reliance (for an overview, see Mahmud et al., 2022). Castelo et al. (2019), for instance, investigate the link between the perceived objectivity and subjectivity of a task on automation reliance: unlike tasks that rely on intuition or personal judgement, objective tasks, defined via their quantifiability and measurability, are associated with a higher willingness to rely on automation². Bigman & Gray (2018) find that automation aversion is particularly prevalent in legal, medical and military decisions, which is supported by evidence that the moral domain is widely considered to be human territory (Lee, 2018; Niszczoła & Kaszás, 2020). While the nature of the delegate’s task has been shown to induce preferences for a specific type of delegate, the possible interaction of task sensitivity and information on the computer delegate’s programming has yet to be explored.

¹For a comprehensive review of the literature on algorithm aversion, see Burton et al. (2020) or Jussupow et al. (2020).

²Note, however, that Kaufmann et al. (2023) compare several studies with regard to the influence of task objectivity on automation reliance and, overall, find no significant preference for algorithms in more objective tasks.

We therefore conducted a delegation experiment in which subjects had to choose between having a human or a computer delegate perform a task for them. We varied (1) the nature of the delegate’s task and (2) the availability of information about the computer delegate’s programming to investigate if and how the two factors can affect delegation preferences — separately and in conjunction with one another.

Our results suggest that delegation preferences are indeed sensitive to both factors: absent information on the computer delegate’s programming, the proportion of subjects who chose the human delegate was significantly higher for a task that was perceived as human-friendly, and vice-versa for a task that was perceived as computer-friendly; this difference did, however, not persist in the presence of programming information. More specifically, we hypothesise that, if a computer delegate is, *ex-ante*, perceived to be well-suited to a task, information about its programming can draw attention to its fallibility and decrease willingness to adopt it. We also find an overall preference for the human delegate that aligns with previous research, but stands in contrast to the higher success expectations that our subjects reported for the computer delegate; this leads us to conclude that the delegation preferences we observe were not driven by differences in performance expectations alone.

The paper proceeds as follows: in Section 2, we describe the design of the experiment and outline our research hypotheses; Section 3 contains the results of our data analysis, while Section 4 discusses these results and concludes.

2 Experimental Design

Our delegation experiment consisted of three stages. Our main variable of interest is a subject’s decision in the *delegation task* in Stage 1 of the experiment, in which they were asked to choose between relying on the performance of a human delegate or that of a computer delegate in the *delegate’s task* in Stage 2. In the *prediction task* in Stage 3, we elicited performance expectations for both types of delegate in the delegate’s task. We employed a random incentive mechanism to determine which of the three stages was relevant for each subject’s payoff (for evidence on the validity of this method, see Chapter 6 in Bardsley et al., 2009; Cubitt et al., 1998; Charness et al., 2016).

In total, we recruited six cohorts of subjects, which were differentiated by the timing of their entry to the experiment. Further details regarding reasons for and consequences of the sequential structure of the experiment will be explained in Section 2.3; until then, we focus on the *typical cohorts*: Cohorts 2 – 5, who participated in all three experimental stages. In a typical cohort, subjects were informed that the human delegate whose performance they could choose to rely on in the delegation task would always be a subject from the next cohort and that, if chosen, this subject would learn of their choice prior to performing the delegate’s task. Consequently, each member of a typical cohort was a potential delegator for a member of the subsequent cohort and a potential delegate for a member of the previous one.

2.1 Treatment Groups

All subjects were randomly assigned to a BASELINE condition or one of four treatment groups, which varied across two dimensions: (1) the nature of the delegate’s task and (2) whether or not information on the programming of the computer delegate was provided. We used two tasks: one that we denote as ‘human-friendly’ (hereafter: HF Task) and one that we denote as ‘computer-friendly’ (hereafter: CF Task). Details regarding the design and selection of the two tasks as well as our definitions of ‘human-friendliness’ and ‘computer-friendliness’ are provided in Section 2.2.

We are primarily interested in whether subjects delegated the delegate’s task to a human or to a computer delegate. Fig. 1.1 illustrates the treatment-specific differences in what subjects learned about the delegate’s task prior to making their decision: the BASELINE served as a measure of delegation preferences absent any information about the nature of the delegate’s task or the programming of the computer delegate; HF-NPI and CF-NPI then provided insights into if and how these preferences changed in response to the nature of the delegate’s task, while HF-PI and CF-PI enabled us to examine the effect of providing information on the computer delegate’s programming in addition to the nature of the delegate’s task.

	NO PROGRAMMING INFO	PROGRAMMING INFO
NO TASK INFO	BASELINE	
HF TASK	HF-NPI	HF-PI
CF TASK	CF-NPI	CF-PI

Note: HF denotes the HUMAN-FRIENDLY TASK, CF the COMPUTER-FRIENDLY TASK; NPI indicates that no programming information was provided, whereas PI indicates that programming information was provided.

Figure 1.1: Treatment Groups

2.2 Delegate’s Task

The HF Task and the CF Task were selected based on a prior survey, in which we asked 200 respondents to evaluate five candidate tasks: a *Memory Task*, a *Prediction Task*, an *Anagram Task*, a *Coordination Task*, and a *CAPTCHA Task*.

Each candidate task was evaluated via 5-point Likert scales, where 1 corresponded to ‘strongly disagree’ and 5 to ‘strongly agree’ with respect to four statements. Specifically, respondents were asked to express their (dis)agreement with the statements, “*I would expect a computer to perform this task well*” and “*I would expect a human to perform this task well*” to evaluate absolute levels of computer-friendliness and human-friendliness. We further included the statement, “*I would expect a computer to perform this task better than a human*” as measure of relative computer-friendliness and human-friendliness. Finally, we elicited (dis)agreement with the statement, “*I think a computer could be programmed to perform this task well*” as an additional, albeit less direct, measure of computer-friendliness with a focus on the possibilities and limits of programming³ (see Appendix A.2.1 for detailed survey instructions).

³Our aim was to find two tasks that differ significantly in their *perceived* human-friendliness and computer-friendliness; as we are interested only in ex-ante perceptions, we do not require the congruence of *perceived* and *actual* human-friendliness and computer-friendliness of our tasks.

Significant differences in their respective task evaluations led to the selection of the *Coordination Task* as the HF Task and the selection of the *Memory Task* as the CF Task. Overall, the respondents reported consistently favourable expectations of a computer’s performance, but the *Memory Task* stood out as having the highest mean scores for our measures of absolute and relative computer-friendliness as well as the lowest mean score for our measure of absolute human-friendliness. Human performance expectations, on the other hand, were generally low; however, the *Coordination Task* stood out as having the lowest mean scores for our measures of absolute and relative computer-friendliness.

We included the *Memory Task* in the survey as a computer-friendly candidate task because storing memory is and always has been one of the key purposes of computers; here, our priors were confirmed by the survey results that indicate a high level of perceived computer-friendliness in both absolute and relative terms, as well as a low level of absolute human-friendliness. Conversely, we included the *Coordination Task* as a human-friendly candidate task due to our belief that, firstly, humans possess an inherent understanding of focal points⁴ that computers can emulate only when given access to sufficient data, and that, secondly, humans recognise this⁵. Survey results only partly confirm this: given the comparatively low levels of absolute computer-friendliness and human-friendliness, respondents seemingly perceived the *Coordination Task* as rather difficult for computers *and* humans alike. Crucially, however, we show that evaluations of the two candidate tasks differ significantly with respect to absolute and relative computer-friendliness; thus making them a suitable pair of tasks for our between-subject treatment manipulation (see Appendix A.1.1 for detailed survey results).

Both the HF Task and the CF Task consisted of five rounds, and all subjects were informed, before making their delegation decision, that: one, each round of a given task had a similar form, and two, human delegates were incentivised to solve the task and computer delegates were programmed to do so, but neither were expected to perform perfectly. Subjects in the BASELINE were given no further information.

Subjects in HF-NPI were presented with the following description of the HF Task:

One of 5 English words needs to be selected. In an experiment reported in 2010, 44 human participants who did not know each other, could not communicate with each other, and had no access to the internet were presented with these five words. The words were arranged in a row from left to right,

⁴There is a long-standing tradition of emphasising the importance of focal points and common knowledge for coordination going back to the seminal works of Schelling (1960) and Lewis (1969). Subsequent research has validated and expanded on the importance of these concepts (see e.g. Mehta et al., 1994a,b; Cubitt & Sugden, 2003).

⁵One could also argue that the *Coordination Task*, which relies on human intuition, fits into the category of ‘subjective’ tasks as defined in Castelo et al. (2019); a category that, in their experiment, is linked to higher levels of algorithm aversion than the category of ‘objective’ tasks.

but the participants were informed that the order of the words would be different for different participants. Each participant was told that they would be rewarded if they selected the same word as another participant who was chosen at random and rewarded in the same way.

The goal of the task is to select the same one of the 5 words as a randomly chosen participant from the 2010 experiment within a time-limit of 10 seconds. Note that success in this task is determined only by whether the goal is met within the time-limit, not by how fast this is accomplished. Note that each of the 5 tasks in the sequence features a different set of words.

Subjects in HF-PI were given the same information, but were furthermore told how the computer delegate, if chosen, would attempt to solve the HF Task:

*The computer accesses **Google search data from 2010 via Google Trends**, a website that analyses the popularity of Google search queries. It compares the number of 2010 Google search queries for the 5 words, and **selects the word with the highest number of Google search queries in that year.***

The HF Task used word sets⁶ and coordination data from Bardsley et al. (2010). In each round of the HF Task, success of the delegate was determined via the random draw of a word from the corresponding word set, using the choice data in Bardsley et al. (2010) to provide the underlying probability distribution, and comparing this word to the delegate’s choice. So, in effect, we randomly paired every subject in our experiment with one of the subjects from Bardsley et al. (2010) and compared whether or not their chosen words matched: in short, the goal of the HF Task was to select the word a human selected when aiming to coordinate with another human.

A human delegate would thus succeed if and only if they selected the same word as the randomly drawn subject from Bardsley et al. (2010), and was most likely to succeed if they selected the modal response in that experiment. The computer delegate, on the other hand, would succeed if and only if the randomly drawn subject chose the word with the highest number of 2010 Google search queries within the set⁷. In other words, the human delegate had two things in common with the randomly drawn subject that the computer delegate did not: (1) their shared humanity, and (2) the goal of coordinating with another human, whereas the computer was drawing on a database of searches that were not made with any coordination in mind; two factors which likely contributed to the (perceived) human-friendliness of the task.

⁶See Appendix A.2.3 for the specific word sets used in the HF Task.

⁷It is worth noting that Google Trends provides relative rather than absolute numbers of search queries, but we ensured that each of our chosen word sets contained one word for which the number of search queries clearly exceeded that of all other words in the set.

Subjects in CF-NPI were presented with the following description of the CF Task:

***A list of 25 English words needs to be memorised.** The order in which the words are presented is randomised. The list will be visible for 5 seconds before it disappears. After the list disappears, 5 words will be shown: 4 of these words were on the list, 1 of them was not.*

***The goal of the task is to select the word that was not on the list within a time-limit of 10 seconds.** Note that success in this task is determined only by whether the goal is met within the time-limit, not by how fast this is accomplished. Note that each of the 5 tasks in the sequence features a **different list of words**.*

Subjects in CF-PI were given the same information, but were furthermore told how the computer delegate, if chosen, would attempt to solve the CF Task:

*The computer records **the first 20 words** on the list. Among the 5 words that are shown after the list disappears, **it identifies the words that it did not record**. If there is only one, it will select this word. If there is more than one, it will **randomly** select one of them.*

The CF Task contained adapted versions of the themed word lists⁸ from Deese (1959). Originally designed to study false recollection bias, we extended these word lists by adding seven words which were connected to the overall theme and three words which were not, in order to render the task adequately complex for our purposes. In each round of the CF Task, we included one of the unconnected words among the five words that the subjects had to choose from after the list disappeared. Given that — as per the design established in Deese (1959) — the correct word was also unconnected to the overall theme of the list, we hoped to moderate the success rate of human delegates by making the two unconnected words stand out as likely solutions.

It follows that the human delegate would succeed if and only if they selected the correct word from among the five words that were shown after the list disappeared, whereas the computer delegate would succeed if (1) none of the four wrong words were among the final five words on the list, or (2) it randomly selected the correct word over any of the wrong words that were among the final five words on the list. Consequently, as with the HF Task, the computer delegate was fallible, but its perfect recall of 20 words and the fact that it was unaffected by false recollection bias likely gave it an advantage over a human delegate; we believe that this contributed to the (perceived) computer-friendliness of the task.

⁸See Appendix A.2.3 for the specific word lists used in the CF Task.

2.3 Procedures

Let us now revisit the sequential structure of the experiment: as shown in Fig. 1.2, Cohorts 2 – 5 are considered the *typical cohorts* because the subjects in these cohorts participated in all three experimental stages; Cohort 1, on the other hand, participated only in Stages 1 and 3, and Cohort 6 participated only in Stage 2.

Recall that delegators learned that their human delegate, if chosen, would be a subject from the next cohort and that, if chosen, this subject would be informed of their choice prior to performing the delegate’s task. In doing so, we preserved what we believe to be a distinctive characteristic of delegation to human beings: a human delegate, unlike a computer delegate, was aware that their performance might affect the delegator’s payoff and might have felt a sense of obligation arise from having been chosen. Rather than recruiting delegators and delegates as two separate groups, the sequential structure allowed us to observe decisions in the delegation task *and* performance in the delegate’s task in all typical cohorts.

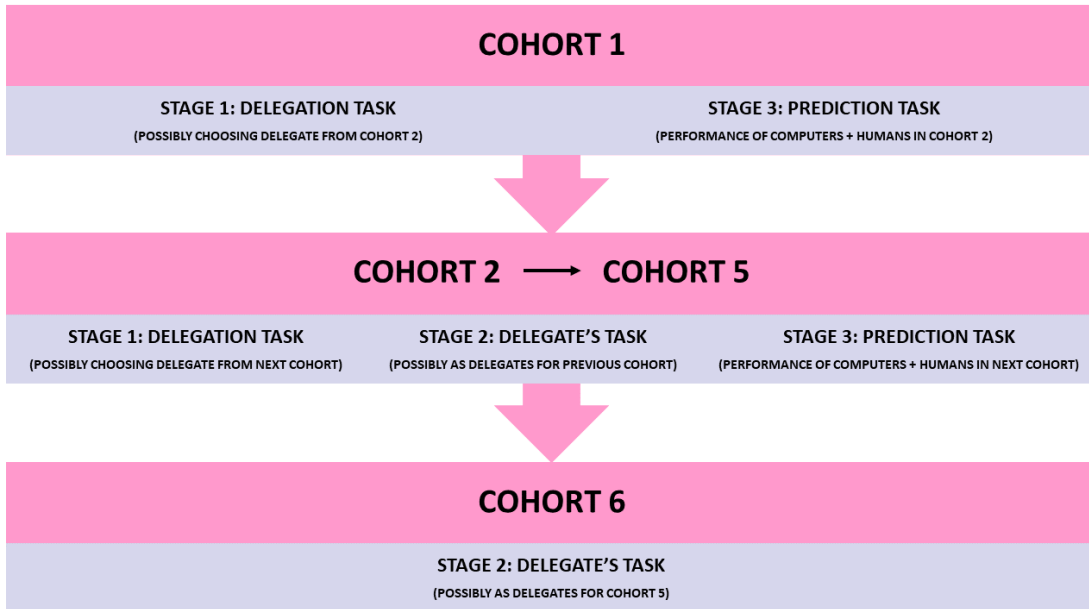


Figure 1.2: Cohort Structure

Given this structure, performance in Stage 2 and belief elicitations in Stage 3 might, however, have been compromised had subjects performed the task that they delegated in Stage 1. Subjects in HF-NPI and HF-PI therefore delegated the HF Task but performed the CF Task, and vice-versa for subjects in CF-NPI and CF-PI⁹. Consequently, subjects did not see the delegate’s task before performing it, and did not perform the task for which they were later asked to report performance expectations.

⁹In the BASELINE, subjects were split into groups B₁ and B₂. Unknowingly, those in B₁ delegated the HF and performed the CF Task, while those in B₂ delegated the CF and performed the HF Task.

It was common knowledge among delegators and human delegates in all treatment groups that *both parties* would earn £0.40 for every round in which a human delegate selected the correct word. We anticipated two possible ways in which other-regarding preferences could act as a motivator for choosing the human over the computer delegate or vice-versa: delegators might choose the human delegate so as not to deprive them of the earning opportunity, or, alternatively, might choose the computer delegate to save the human delegate from the effort of having to perform the delegate’s task. To eliminate these motivators, delegators therefore also learned that, regardless of which type of delegate they chose, the other type of delegate would still perform the task and, if human, receive their own payment for it¹⁰.

In Stage 1, after the delegation task, we also measured the strength of a subject’s delegation preference; this enabled us to later test whether observed preferences for either type of delegate were the result of tie-breaker choices in the face of indifference. To do so, we elicited willingness-to-accept (WTA) values for exchanging a subject’s chosen type of delegate for the other type of delegate using a Price List set-up: on a list of payments ranging from £0.00 to £2.00, the minimum and maximum amounts that a delegate could earn their delegator in the delegate’s task, subjects indicated whether, in exchange for each of these payments, they would swap their chosen type of delegate for the other one. We then randomly drew one of the payments on the list and executed each subject’s decision for this payment (see Appendix A.2.2 for detailed instructions).

In Stage 3, the prediction task, we elicited performance expectations for the delegate’s task to account for differences in performance expectations as a driver of delegation preferences. Subjects were asked to guess in how many of the five rounds each type of delegate would select the correct word; each correct guess earned them £1.00¹¹. Hence, for every human performance, we had to generate a corresponding computer performance, which also enabled us to compare human and computer performances. Moreover, if a subject believed that a human delegate’s awareness of their role can affect performance, their performance expectations for the human delegate would depend on their delegation decision in Stage 1; however, the randomised component of the WTA elicitation prevented subjects from knowing which of the two types of delegate would, ultimately, be acting as their delegate. For this reason, we elicited a total of *three* performance expectations: one for the human delegate if they acted as a delegate (hereafter: human delegate), one for the human delegate if they did not act as a delegate (hereafter: human non-delegate), and one for the computer delegate. We used the two relevant expectations in each subject’s case to determine payoffs in the prediction task.

¹⁰Revealing this also ensured that delegators knew that financial incentives created no conflict of material interests between them and a human delegate. To generate symmetrical assumptions for the computer delegate, subjects were told that “the computer is programmed to solve the task”.

¹¹As both types of delegate always performed the delegate’s task, there was always a unique, correct answer in each case, so all subjects could earn up to £2.00 regardless of their delegate choice.

Finally, for all cohorts and in all treatment groups, the experiment concluded with a post-experiment survey gathering data on individual levels of computer proficiency, attitudes towards automation reliance, and demographic characteristics. As explained above, subjects were paid according to their own and, where applicable, their chosen delegate’s decisions in one, randomly selected stage of the experiment.

2.4 Hypotheses

Hypothesis 1. *In the BASELINE, subjects exhibit a preference for the human delegate.*

Previous research often documents a preference for human over computer agency. While this preference for human agency typically emerges in the presence of *some* context about the nature of the task, we still expected a higher proportion of human delegate choices than computer delegate choices in the absence of information on the nature of the delegate’s task and the programming of the computer delegate. Testing if the corresponding WTA values were significantly higher than zero enables to determine whether, should we observe such a difference in proportions, it is indeed evidence of a preference for human agency rather than the result of human-leaning tie-breaker choices in the face of indifference.

Hypothesis 2. *Relative to the BASELINE, the HF Task shifts delegation preferences towards the human delegate, whereas the CF Task shifts delegation preferences towards the computer delegate.*

Hypothesis 3a. *Information about the programming of the computer delegate affects delegation preferences regardless of the nature of the delegate’s task.*

Hypothesis 3b. *Information about the programming of the computer delegate affects delegation preferences differently depending on the nature of the delegate’s task.*

Similarly, for **Hypotheses 2, 3a** and **3b**, we compare the proportions of human delegate choices between treatments and, should we find any significant differences, verify that these reflect preferences for one type of delegate over the other by examining strength of preference via WTA values.

Note that, for **Hypotheses 3a** and **3b**, we did not specify the *direction* of the hypothesised (difference in) effect: given the nature of the task, the availability of information about the programming of the computer delegate could either reinforce or offset beliefs regarding the suitability of each type of delegate, and thus influence delegation preferences in two directions.

Specifically, the revelation of programming information reduced uncertainty about how the computer delegate approached a task, but it drew attention to its fallibility at the same time; if the former outweighed the latter in either task, we would expect an increase in computer delegate choices and vice-versa if the latter outweighed the former. It is furthermore possible that this effect interacts with the nature of the task, especially since the availability of information on the programming of the computer, in our set-up, entailed slightly different consequences in the two tasks: in the HF Task, the programming information explained how the computer could succeed in the task, whereas in the CF Task, its fallibility was highlighted. Conditional on subjects realising this, programming information could therefore moderate task-sensitivity relative to the BASELINE.

Hypothesis 4. *Subjects who act as delegates outperform those who do not.*

Provided that the subjects in our experiment exhibited other-regarding preferences that place a strictly positive value on the utility of others, delegates who knew that their performance might affect not only their own payoff but also that of another subject had a stronger incentive to perform well¹².

Hypothesis 5a. *Subjects who choose the human delegate expect the human delegate to perform better than the computer delegate.*

Hypothesis 5b. *Subjects who choose the computer delegate expect the human delegate to perform worse than the computer delegate.*

If performance expectations were a key driver of delegation preferences, as one would expect under the assumption of a general preference for more money over less, we should see higher performance expectations for the human delegate among subjects who chose the human delegate than among those who chose the computer delegate, and lower performance expectations for the human delegate among subjects who chose the computer delegate than among those who chose the human delegate.

¹²Remember that delegators and delegates both earned £0.40 for every round in which the delegate selected the correct answer, so inequality aversion should not play a role here.

3 Results

The experiment was programmed using Qualtrics software and conducted on Prolific (Palan & Schitter, 2018) in April 2021. We recruited a total of 360 subjects over six cohorts of 60 members each. Of these 360 subjects, 234 were female and the average age was 33.19 years¹³. All subjects received a participation fee of £1.50 and could earn a bonus payment of up to £4.00 depending on which experimental stage was selected for payoff; the median completion time across all cohorts was 10 minutes and 34 seconds and the average payment, including bonuses, was £2.59¹⁴.

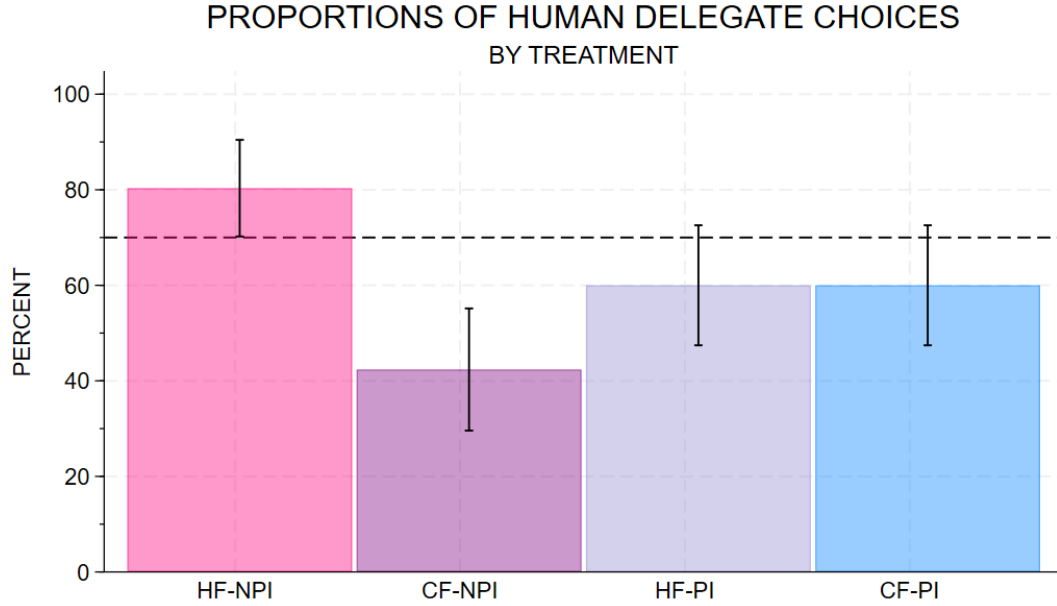
3.1 Delegation Decision

Overall, 300 subjects made a delegation decision in Stage 1: 188 of these subjects (62.67%) chose the human delegate, while the remaining 112 (37.33%) chose the computer delegate; this overall proportion of delegate choices favouring the human delegate is significantly larger than 50% ($p < 0.01$, Two-Sided One Sample z-Test). Note that, hereafter, we focus on these *initial* delegate choices as opposed to the *realised* delegate ‘choices’ after the WTA elicitation; we do, of course, examine WTA values in the interest of determining strength of preference, but post-WTA delegate ‘choices’ are of no further interest to us.

Fig. 1.3 depicts the proportions of human delegate choices in the four treatment groups, with whiskers indicating the 95% confidence intervals and a dashed line marking the BASELINE proportion as a benchmark. Fig. 1.4 shows the corresponding mean WTA values for the two different delegate choices, again with whiskers indicating the 95% confidence intervals and a dashed line marking the BASELINE means as a benchmark. At a first glance, it appears that, relative to the BASELINE, the HF Task increases the proportion of human delegate choices and the CF Task decreases the proportion of human delegate choices in the absence of programming information; these effects are, however, moderated when programming information is introduced. Mean WTA values are also higher for subjects who chose the human delegate than for those who chose the computer delegate in the HF Task, and vice-versa in the CF Task. Table 1.1 reports the corresponding Two-Sided Two-Sample Tests of Proportion for the null hypothesis of between-treatment equality of the proportions of human delegate choices, while Table 1.2 shows in which cases the proportions of human delegate choices and computer delegate choices are associated with a mean WTA value that is significantly higher than 0.

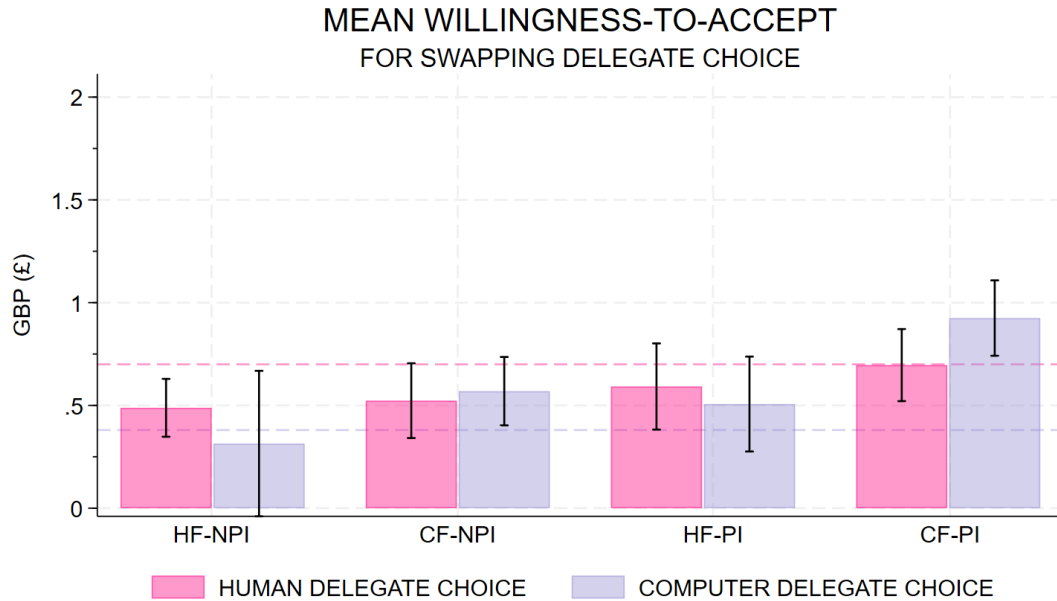
¹³Demographic sample characteristics and a treatment breakdown thereof can be found in Tables A.12 and A.13 in Appendix A.1.2.

¹⁴Both median completion time and average payment were similar if we consider only the typical cohorts, at 11 minutes and 52 seconds and £2.61.



Note: ‘Human Delegate Choice’ denotes a subject’s choice of the human delegate before the WTA elicitation. Whiskers indicate 95% confidence intervals; the dashed line indicates the BASELINE benchmark of 70%.

Figure 1.3: Proportions of Human Delegate Choices



Note: We only use data from the 155 subjects with a single switch point from ‘keep’ to ‘swap’ and the 42 subjects who always selected ‘swap’. Whiskers indicate 95% confidence intervals; dashed lines indicate the BASELINE benchmarks, 0.70 and 0.38.

Figure 1.4: Mean WTA Values

	HDC ₁	HDC ₂	p (H ₀ : HDC ₁ = HDC ₂) <i>Two-Sample Test of Proportion</i>
BASELINE — HF-NPI	0.70	0.80	0.1884
BASELINE — CF-NPI	0.70	0.42	0.0024
BASELINE — HF-PI	0.70	0.60	0.2508
BASELINE — CF-PI	0.70	0.60	0.2508
HF-NPI — CF-NPI	0.80	0.42	0.0000
HF-PI — CF-PI	0.60	0.60	1.0000
HF-NPI — HF-PI	0.80	0.60	0.0145
CF-NPI — CF-PI	0.42	0.60	0.0544

Note: HDC₁ and HDC₂ denote the two proportions of human delegate choices in each row for whose between-treatment equality we test using Two-Sided Two-Sample Tests of Proportion.

Table 1.1: Treatment Comparisons of Proportions of Human Delegate Choices

	MEAN WTA IN £	p (H ₀ : WTA > 0) <i>One-Sided Sign Test</i>
HUMAN DELEGATE CHOICE		
BASELINE	0.70	0.0000
HF-NPI	0.49	0.0000
CF-NPI	0.52	0.0005
HF-PI	0.59	0.0000
CF-PI	0.70	0.0000
COMPUTER DELEGATE CHOICE		
BASELINE	0.38	0.0039
HF-NPI	0.31	0.1250
CF-NPI	0.57	0.0000
HF-PI	0.51	0.0010
CF-PI	0.93	0.0000

Note: We only use data from the 155 subjects with a single switch point from ‘keep’ to ‘swap’ and the 42 subjects who always selected ‘swap’.

Table 1.2: Mean WTA for Swapping Initial Delegate Choice

Firstly, in line with **Hypothesis 1**, 42 of 60 subjects (70.00%) in the BASELINE chose the human delegate, which, like the overall proportion of human delegate choices, is also significantly larger than 50% ($p < 0.01$, Two-Sided One-Sample z-Test). Furthermore, mean WTA values in the BASELINE are significantly higher than 0 irrespective of delegate choice. Of course this may, at least in part, reflect an endowment effect, but the higher mean WTA for those who chose the human delegate in particular suggests that, in line with previous research, this is evidence of a preference for human agency rather than the result of human-leaning tie-breaker choices in the face of indifference.

Result 1. *In the BASELINE, subjects exhibit a clear preference for the human delegate.*

Next, we cannot reject the null hypothesis of an equal proportion of delegate choices favouring the human and the computer delegate in CF-NPI (42.37%), HF-PI (60.00%), and CF-PI (60.00%). In HF-NPI, on the other hand, the proportion of human delegate choices (80.33%) is again significantly larger than 50% ($p < 0.01$, Two-Sided One-Sample z-Test). WTA values in HF-NPI for the human, albeit not for the computer delegate, are also significantly higher than 0; indicating that, like in the BASELINE, this does, indeed, reflect a preference for the human delegate.

Relative to the BASELINE, we observe a higher proportion of human delegate choices in HF-NPI and a lower proportion in CF-NPI; however, only the latter of the two differences is significant. Reflecting the respective natures of the delegate’s task, we furthermore find that the proportion of human delegate choices is significantly higher in HF-NPI than in CF-NPI. Corresponding mean WTA values for both types of delegates or, in case of HF-NPI for the type of delegate that was chosen more frequently, are all significantly higher than 0. In summary, **Hypothesis 2** is largely supported.

Result 2. *In the absence of information about the computer delegate’s programming, the proportion of human delegate choices is significantly higher in the HF Task than in the CF Task.*

Conversely, between HF-PI and CF-PI, there is no difference in the proportion of human delegate choices. Moreover, the proportion of human delegate choices is significantly higher in HF-NPI than in HF-PI, but the reverse is the case for CF-NPI and CF-PI. In other words, the effect of the nature of the delegate’s task on delegate choices is neutralised by opposing effects of information on the computer delegate’s programming. Here as well, the corresponding mean WTA values for both types of delegates are all significantly higher than 0. **Hypotheses 3a** and **3b** are supported.

Result 3. *Information about the computer delegate’s programming decreases the proportion of human delegate choices in the HF Task, but increases the proportion of human delegate choices in the CF Task.*

One explanation for this finding is that, in the absence of programming information, the natures of the two tasks induce different beliefs about the capabilities of the computer. By design, the HF Task and the CF Task are designed to appear human-friendly and computer-friendly, respectively; hence, subjects may struggle to imagine how the computer delegate could solve the HF Task, but believe it to be near infallible in the CF Task. Consequently, information on the computer delegate’s programming may correct these beliefs — by highlighting the computer delegate’s ability to solve the HF Task and its fallibility in the CF Task — and so moderate the previously observed differences in delegation preferences. We will revisit this explanation when analysing performance expectations in the prediction task.

3.2 Delegate’s Performance

Of the 300 subjects to make a delegation decision in Stage 1, *after* the WTA elicitation, 148 were paired with a human delegate and 152 were paired with a computer delegate. It follows that, in Stage 2, 148 subjects acted as a human delegate and 152 subjects acted as a human non-delegate.

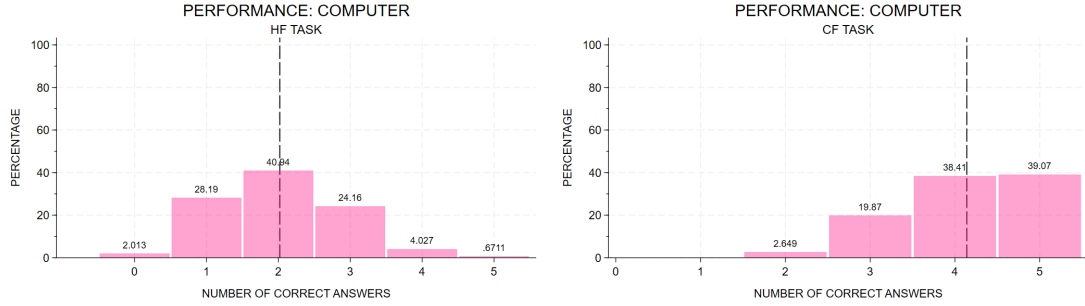
Fig. 1.5 — Fig. 1.7 show the distributions of correct answers of computer delegates, human delegates, and human non-delegates depending on the delegate’s task, with vertical dashed lines marking the mean numbers of correct answers. Computer delegates averaged 2.02 correct answers in the HF Task and 4.14 correct answers in the CF Task; human delegates and human non-delegates, on the other hand, averaged 1.88 and 1.58 correct answers in the HF Task, respectively, and 3.29 and 3.38 correct answers in the CF Task, respectively.

Hence, overall, all types of delegate performed better in the CF Task than in the HF Task, which matches our result from the prior survey that respondents generally perceived the *Coordination Task* as comparatively difficult for computers *and* humans. Computer delegates also consistently outperformed human non-delegates in the HF Task and in the CF Task, and human delegates in the CF Task but not in the HF Task¹⁵ ($p < 0.01$, Mann-Whitney U Test) (see Table A.14 in Appendix A.1.2 for further details).

Partially supporting **Hypothesis 4**, human delegates narrowly outperformed human non-delegates in the HF Task ($p < 0.05$, Mann-Whitney U Test) but not in the CF Task. However, performance expectations, which we analyse in detail in Section 3.3, do not capture this: human delegates and human non-delegates were, on average, expected to give 2.58 and 2.62 correct answers in the HF Task, respectively, and 3.02 and 3.01 in the CF Task, respectively; holding the nature of the task constant, neither of the differences in performance expectations are statistically significant between human delegates and human non-delegates. In other words, delegators, in a sense, underappreciated the possible increase in performance resulting from the human-specific *awareness* of acting as a delegate, and we can rule this out as a significant contributor to the overall preference for the human delegate over the computer delegate.

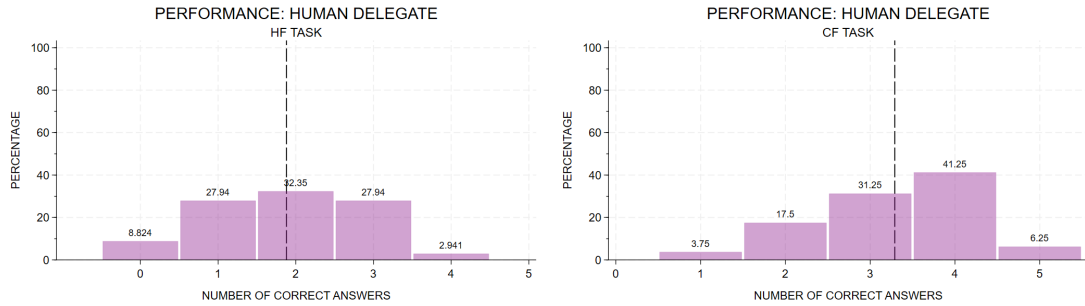
Result 4. *Human delegates outperform human non-delegates in the HF Task but not in the CF Task, but delegators do not anticipate this difference in performance.*

¹⁵Recall that our treatment manipulations rely exclusively on differences in ex-ante perceptions of human-friendliness and computer-friendliness, so we are not concerned about the fact that, ex-post, humans performed worse in the HF Task than in the CF Task nor about the fact that human non-delegates were outperformed by computers in the HF Task.



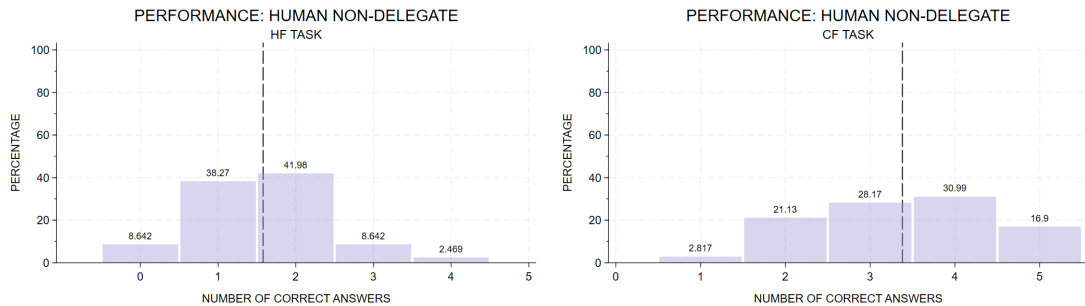
Note: Distribution of the number of correct answers from the 300 computer delegates in the HF Task and the CF Task. Dashed vertical lines indicate the mean numbers of correct answers, 2.02 and 4.14.

Figure 1.5: Performance: Computer Delegate



Note: Distribution of the number of correct answers from the 148 human delegates in the HF Task and the CF Task. Dashed vertical lines indicate the mean numbers of correct answers, 1.88 and 3.29.

Figure 1.6: Performance: Human Delegate



Note: Distribution of the number of correct answers from the 152 human non-delegates in the HF Task and the CF Task. Dashed vertical lines indicate the mean numbers of correct answers, 1.58 and 3.38.

Figure 1.7: Performance: Human Non-Delegate

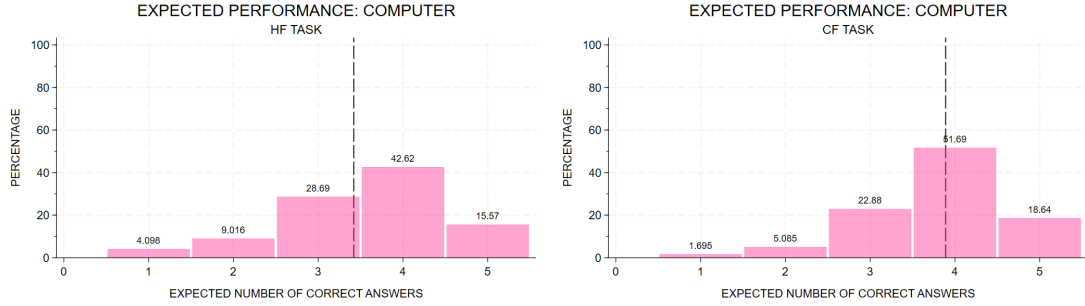
3.3 Performance Expectations

We now turn our attention towards the prediction task in Stage 3. We have previously seen that performance expectations for human delegates and human non-delegates, at least on average, rarely differed, and even if such differences did exist, their relevance for our research objectives is comparatively low. Consequently, we hereafter focus only on performance expectations for human delegates and computer delegates; however, analogous analyses for human non-delegates can be found in Fig. A.2 — Fig. A.5 and Tables A.15 — A.17 in Appendix A.1.2.

Fig. 1.8 and Fig. 1.9 illustrate performance expectations for computer delegates and human delegates via the distribution of expected correct answers, depending on the delegate’s task. Computer delegates were, on average, expected to give 3.42 correct answers in the HF Task, 3.89 in the CF Task, and 3.73 in the BASELINE where delegators were not given any information on the nature of the delegate’s task or the programming of the computer delegate. In comparison, human delegates were, on average, expected to give 2.58 correct answers in the HF Task, 3.02 in the CF Task, and 2.92 in the BASELINE. Results in Table 1.3 confirm that subjects reported significantly higher performance expectations for the computer delegate in both tasks. This is in line with actual performance, but at odds with the high proportions of human delegate choices we see in all treatment groups, particularly in the BASELINE and HF-NPI.

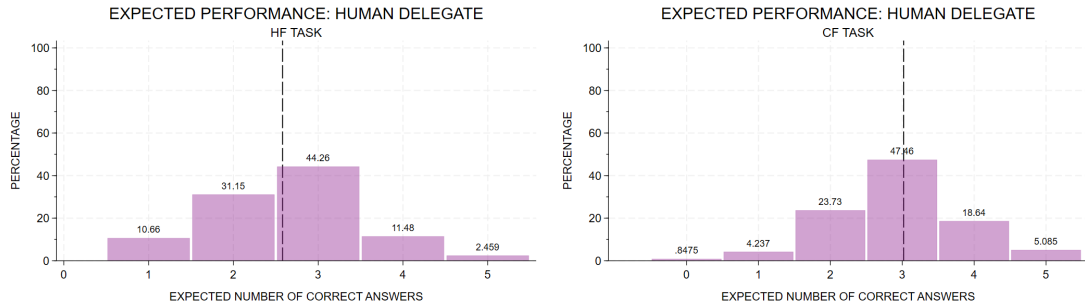
Built on the underlying assumption that subjects prioritise their material interests, **Hypotheses 5a** and **5b** predict that subjects who chose the human delegate expected the human delegate to outperform the computer delegate and vice-versa for subjects that chose the computer delegate.

Fig. 1.10 and Fig. 1.11 illustrate performance expectations for computer delegates and human delegates via the distribution of expected correct answers, depending on delegate choices in Stage 1. Subjects who chose the computer delegate, on average, expected the computer delegate to give 3.82 correct answers and the human delegate to give 2.77 correct answers. In comparison, subjects who chose the human delegate, on average, expected the computer delegate to give 3.58 correct answers and the human delegate to give 2.85 correct answers. Results in Table 1.3 confirm that subjects reported significantly higher performance expectations for computer delegates regardless of delegate choice. Furthermore, subjects who chose the computer delegate reported significantly higher performance expectations for the computer delegate than subjects who chose the human delegate ($p < 0.05$, Mann-Whitney U Test); subjects who chose the human delegate, on the other hand, did not report significantly higher performance expectations for the human delegate than subjects who chose the computer delegate. Our findings thus support **Hypothesis 5b** but contradict **Hypothesis 5a**.



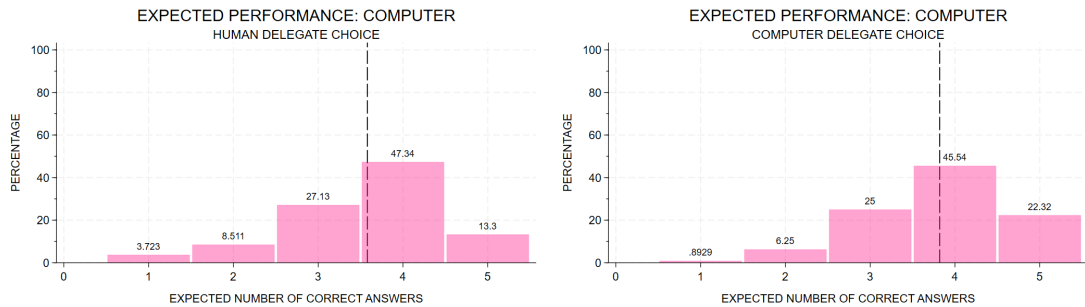
Note: Distribution of the number of expected correct answers for the computer delegate, depending on the delegate's task. Dashed vertical lines indicate the mean numbers of expected correct answers, 3.42 and 3.89.

Figure 1.8: Expected Performance: Computer Delegate (By Delegate's Task)



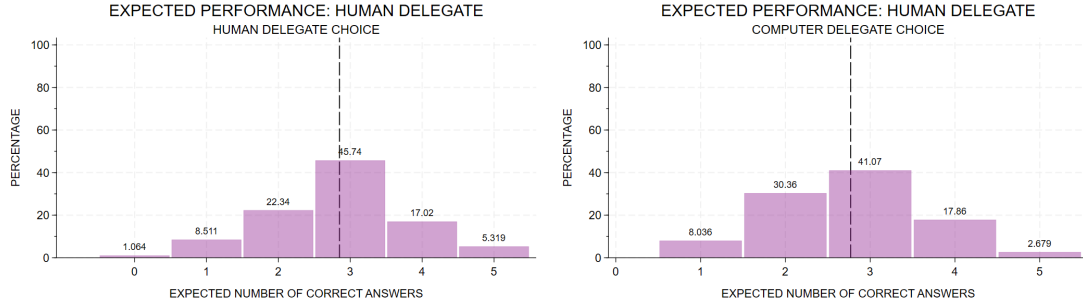
Note: Distribution of the number of expected correct answers for the human delegate, depending on the delegate's task. Dashed vertical lines indicate the mean numbers of expected correct answers, 2.58 and 3.02.

Figure 1.9: Expected Performance: Human Delegate (By Delegate's Task)



Note: Distribution of the number of expected correct answers for the computer delegate, depending on delegate choice. Dashed vertical lines indicate the mean numbers of expected correct answers, 3.58 and 3.82.

Figure 1.10: Expected Performance: Computer Delegate (By Delegate Choice)



Note: Distribution of the number of expected correct answers for the human delegate, depending on delegate choice. Dashed vertical lines indicate the mean numbers of expected correct answers, 2.77 and 2.85.

Figure 1.11: Expected Performance: Human Delegate (By Delegate Choice)

	μ_{COMPUTER}	μ_{HUMAN}	$p \text{ (} H_0: \text{EQUAL EXP)}$ <i>Wilcoxon Signed-Rank Test</i>
NO INFO	3.73	2.92	0.0000
HF TASK	3.42	2.58	0.0000
HF-NPI	3.44	2.62	0.0000
HF-PI	3.40	2.53	0.0000
CF TASK	3.89	3.02	0.0000
CF-NPI	4.02	3.05	0.0000
CF-PI	3.77	2.98	0.0000
COMPUTER DELEGATE CHOICE	3.82	2.77	0.0000
HUMAN DELEGATE CHOICE	3.58	2.85	0.0000

Note: We have 300 performance expectations for each type of delegate; μ_{COMPUTER} and μ_{HUMAN} denote the mean numbers of expected correct answers for the computer delegate and the human delegate, respectively. NO INFO refers to the BASELINE, in which delegators received no information on the nature of the delegate's task.

Table 1.3: Comparisons of Performance Expectations

Result 5a. *Subjects, on average, anticipated computer delegates to outperform human delegates in all treatment groups and regardless of delegate choice.*

We have so far shown that **Hypothesis 5a** is not supported by *absolute* performance expectations; however, we can also investigate *relative* performance expectations to see if the difference in performance expectations between human delegates and computer delegates differs depending on a subject's delegate choice. Delegators who chose the human delegate might expect the computer delegate to outperform the human delegate, but to a lesser degree than delegators who chose the computer delegate.

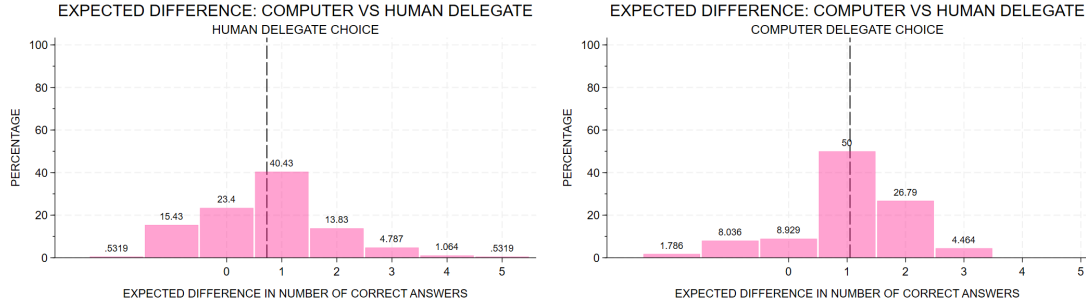
Fig. 1.12 illustrates the distribution of differences in performance expectations depending on delegate choice. We have calculated these by subtracting individual performance expectations for the human delegate from those for the computer delegate: a value of 1 would, for example, indicate that a subject expected the computer delegate to give one more correct answer than the human delegate, a value of -1 would indicate that a subject expected the human delegate to give one more correct answer than the computer delegate and so forth.

Subjects who chose the computer delegate, on average, expected the computer delegate to outperform the human delegate by 1.11; subjects who chose the human delegate, on the other hand, expected the computer delegate to outperform the human delegate by 0.73. The difference in performance expectations is significantly higher than zero ($p < 0.01$, Two-Sided Wilcoxon Single Sample t-test) irrespective of delegate choice, which again indicates consistently higher performance expectations for the computer than for the human delegate and thus further contradicts **Hypothesis 5a**. Comparing these differences, we do, however, also see that the amount by which subjects expect computer delegates to outperform human delegates is significantly smaller among those who chose the human delegate than among those who chose the computer delegate ($p < 0.01$, Mann-Whitney U Test).

Nevertheless, many subjects who chose the human delegate still expected it to be outperformed by the computer delegate. To be precise, of the 188 subjects who chose the human delegate, only 30 (15.96%) expected it to outperform the computer delegate while 114 (60.64%) believed the reverse to be true; the remainder reported equal performance expectations. Of the 112 subjects who chose the computer delegate, on the other hand, 91 (81.25%) expected it to outperform the human delegate while 11 believed the reverse to be true (9.82%); the remainder, again, reported equal performance expectations. Indeed, the proportion of subjects who expected their chosen delegate to outperform the other is significantly higher among those who chose the computer delegate than among those who chose the human delegate ($p < 0.01$, Two-Sided Two-Sample Test of Proportion).

In conclusion, delegators consistently expected computer delegates to outperform human delegates across both tasks, all treatment groups, and irrespective of their delegate choice. We do find some difference in *relative* but not in *absolute* performance expectations depending on delegate choice, but this difference alone cannot fully account for the observed differences in delegation preferences.

Result 5b. *Differences in performance expectations between the human delegate and the computer delegate, albeit sensitive to delegate choice in relative terms, cannot explain the observed delegation preferences.*



Note: Distribution of individual differences in the number of expected correct answers between the computer delegate and the human delegate, depending on delegate choice; calculated by subtracting human delegate performance expectations from computer delegate performance expectations. Dashed vertical lines indicate the mean differences in the number of expected correct answers, 0.73 and 1.11.

Figure 1.12: Expected Difference: Computer vs Human Delegate

3.4 Determinants of Automation Reliance

We have shown that performance expectations alone cannot explain the high proportion of human delegate choices or the significant differences in delegation preferences between some of our treatment groups. Next, we perform OLS regressions for a comprehensive analysis of the determinants of automation reliance in our experiment.

Our dependent variable is a dummy that takes the value of 1 for subjects who chose the human delegate, and a value of 0 for subjects who chose the computer delegate. Explanatory variables include two treatment dummies indicating the nature of the delegate's task and whether or not delegators received programming information, as well as their interaction term. We further included differences in performance expectations as defined above, demographic characteristics and attitudes towards and knowledge of computers.

Overall, the three OLS regressions reported in Table 1.4 support our results thus far. Subjects who were offered the HF Task were, *ceteris paribus*, significantly more likely to choose the human delegate; the same is true for subjects who received information on the computer delegate's programming. However, the interaction term between the two treatment manipulations reveals that the effect of the HF Task on the likelihood of choosing the human delegate was, indeed, moderated by information on the programming of the computer delegate: in (1), HF-NPI corresponds to a coefficient of 0.38 that decreases to 0.18 when programming information is introduced in HF-PI; conversely, CF-NPI corresponds to a coefficient of 0 that increases to 0.18 when programming information is introduced in CF-PI. In other words, in the HF Task, programming information encouraged automation reliance, whereas in the CF Task, it discouraged automation reliance.

Differences in performance expectations in favour of the computer delegate, on the other hand, had a marginally significant but comparatively small, negative effect on the likelihood of choosing the human delegate; this confirms the weak effect of performance expectations that we have found previously.

Burton et al. (2020) posit that “a human decision-maker rarely, if ever, confronts an algorithm with a blank slate” but instead approaches automation reliance decisions against the backdrop of expectations and experiences with automated agents in the real world. For this reason, we also collected data on a subject’s perception of the overall influence of computers on their life and of the importance of information about a computer delegate’s programming in the post-experiment survey. We find that a positive overall influence of computers on a subject’s life is associated with a lower likelihood of choosing the human delegate and vice-versa for a high degree of importance placed on information about a computer delegate’s programming¹⁶.

Women were significantly more likely to choose the human delegate in all treatment groups save for the BASELINE, but all of our effects are robust to the inclusion of controls for gender and age (see also Fig. A.1 in Appendix A.1.2).

One possible explanation for the weak effect of performance expectations that we have yet to explore is that performance expectations were affected by our treatment manipulations. In other words, rather than treatment manipulations driving delegation preferences without much of an effect of performance expectations, it might be the case that, instead, treatment manipulations influenced performance expectations, which, in turn, drove delegation preferences.

We therefore test for between-treatment differences in performance expectations. Table 1.5¹⁷ shows that, holding constant the nature of the delegate’s task, we find no significant differences between treatment groups with and without information on the computer delegate’s programming, except for one case: in the CF Task, programming information significantly decreased performance expectations for the computer delegate. Firstly, this matches the decrease in the proportion of human delegate choices as we move from CF-NPI to CF-PI; secondly, it lends credence to our earlier theory that revealing information about the computer delegate’s programming might have inadvertently drawn attention to its fallibility and negatively affected willingness to adopt it as a delegate. We also theorised that, similarly, revealing the computer delegate’s programming in the HF Task might have increased reliance on the computer

¹⁶In addition to this, we collected data on a subject’s self-reported computer proficiency, their feelings about delegating decisions to computers, and their perceived risk of automation replacement (see Appendix A.2.2) but found all of them to have no significant effect on delegate choices.

¹⁷Note that the mean performance expectations reported in Table 1.5 correspond to those reported in Table 1.3, but are compared between rather than within treatments for different types of delegates.

delegate because it highlighted its ability to solve the HF Task; however, performance expectations do not support this. A possibility unrelated to performance is that, in the HF Task, the computer delegate’s programming revealed that it drew on human behaviour via Google Searches and therefore rendered it more similar to the preferred human delegate. In summary, it does not appear to be the case that our treatment manipulations affected delegation preferences via performance expectations alone.

	HUMAN DELEGATE CHOICE		
	(1)	(2)	(3)
HF TASK	0.38*** (0.09)	0.37*** (0.09)	0.39*** (0.09)
PROGRAMMING INFO	0.18** (0.09)	0.17* (0.09)	0.17** (0.09)
HF TASK $\times$ PROGRAMMING INFO	−0.38*** (0.12)	−0.37*** (0.12)	−0.37*** (0.12)
EXPECTED PERFORMANCE DIFFERENCE		−0.05* (0.03)	−0.05* (0.03)
FEMALE			0.15** (0.06)
AGE			0.00 (0.00)
COMPUTER INFLUENCE ON LIFE			−0.08* (0.05)
PROGRAMMING IMPORTANCE			0.07* (0.04)
CONSTANT	0.42*** (0.06)	0.47*** (0.07)	0.40 (0.25)
Observations	240	240	239
Prob > F	0.0003	0.0003	0.0001
R-squared	0.0758	0.0868	0.1323
Root MSE	0.4732	0.4714	0.4638

Note: BASELINE observations are not included. EXPECTED PERFORMANCE DIFFERENCE is calculated by subtracting performance expectations for the human from those for the computer delegate. COMPUTER INFLUENCE ON LIFE is a score from 1 to 5: 1 corresponds to ‘extremely negative’ and 5 to ‘extremely positive’. PROGRAMMING IMPORTANCE is a score from 1 to 4: 1 corresponds to ‘not at all important’ and 4 to ‘extremely important’. Standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.4: OLS Regressions: Human Delegate Choice

	μ_1	μ_2	p (H_0 : EQUAL EXP) <i>Mann-Whitney U Test</i>
EXPECTATION: HUMAN DELEGATE			
HF-NPI – CF-NPI	2.62	3.05	0.0251
HF-PI – CF-PI	2.53	2.98	0.0010
HF-NPI – HF-PI	2.62	2.53	0.5534
CF-NPI – CF-PI	3.05	2.98	0.8876
EXPECTATION: COMPUTER DELEGATE			
HF-NPI – CF-NPI	3.44	4.02	0.0041
HF-PI – CF-PI	3.40	3.77	0.0118
HF-NPI – HF-PI	3.44	3.40	0.5072
CF-NPI – CF-PI	4.02	3.77	0.0541

Note: We have 300 performance expectations for each type of delegate; μ_1 denotes the mean number of expected correct answers for the indicated type of delegate in the first treatment group in a given row, and μ_2 denotes the mean number of expected correct answers for the indicated type of delegate in the second treatment group in a given row.

Table 1.5: Treatment Comparisons of Performance Expectations

4 Discussion

We set out to investigate if and how the nature of a delegate’s task and information on the computer delegate’s programming influence delegation preferences for human versus computer delegates. We believe that both dimensions are important to consider in the context of the persistent underreliance on automation, separately and in conjunction with one another. Robo-advisors and self-driving cars are only two examples of the continuing rise of full or partial automation of decision environments that were once reserved for human agents. Previous research has shown that delegation preferences for human and computer delegates are sensitive to the nature of the task but, to the best of our knowledge, the interaction of this sensitivity with information on the computer delegate’s programming has so far not been explored. Considering the much discussed black-box character of automation as a possible deterrent of automation reliance, it is important to investigate if and how making information on a computer delegate’s programming available to potential delegators affects their willingness to adopt it.

Our results suggest that whether or not people are willing to adopt a computer over a human delegate is highly dependent on the perceived computer-friendliness of the delegate’s task and on whether or not information on how the computer delegate would approach it is readily available. Crucially, we observe opposing effects of information on the computer delegate’s programming depending on the nature of the delegate’s task: in case of a task that was ex-ante perceived to be more human-friendly, the proportion of computer delegate choices increased with programming information; in case of a task that was ex-ante perceived to be more computer-friendly, it decreased. As stated above, a popular theory regarding the preference for human over computer agency is the black-box character of a computer delegate, which suggests that reducing the uncertainty surrounding its programming may increase willingness to adopt it. However, our results suggest that, if a computer delegate is ex-ante perceived to be well-suited to the task, information about its programming could also draw attention to its fallibility and have the opposite effect instead. Given our task-specific set-up, we cannot differentiate between an effect of the *availability* of programming information itself and an effect of the *discrepancy* between programming information and what a subject might naturally expect a computer can or will do: in other words, our effects may, at least in part, result from the conflict that the specific programming information we provided introduced. In either case, however, our results demonstrate that programming information may matter when it differs from prior expectations.

In line with previous research, our subjects furthermore exhibit an overall preference for the human delegate in spite of consistently higher performance expectations for the computer delegate. We do find some difference in relative performance expectations between the human and the computer delegate depending on a subject’s delegate choice; however, these small differences could reflect ex-post rationalisation of delegate

choices, and the majority of subjects who chose the human delegate still reported higher performance expectations for the computer delegate. Consequently, differences in performance expectations alone cannot account for the overall higher proportion of human delegate choices that we observe.

Let us therefore explore a few theories as to why delegation preferences are not congruent with performance expectations in our study. One possible explanation that we briefly touched upon are other-regarding preferences; these could play a role if subjects held the belief that choosing the computer delegate would lead to the human delegate missing out on payment. We attempted to eliminate this as a motivator for choosing the human delegate by informing subjects that, regardless of which delegate they chose, the other delegate would still perform the task and, if human, receive payment for it. In addition to this, the delegation decision was preceded by two control questions that were designed to reinforce and ensure the subjects' understanding of this feature. Subjects had to answer both questions correctly before they were able to move on, and around 94% succeeded in doing so on their first or second try. It is therefore unlikely that misplaced other-regarding preferences can explain our findings.

Alternatively, performance expectations could have changed from the delegation task to the prediction task for subjects who performed the delegate's task for the previous cohort in between the two. Our cohort structure was such that subjects performed the opposite task to the one featured in their delegation and prediction tasks, which allowed for treatment-based asymmetry in the effect of the delegate's task on performance expectations. If this was the case, we would also expect performance expectations in Cohort 1 to differ from performance expectations in Cohorts 2–5 as Cohort 1 did not perform the delegate's task; however, we find no evidence of this.

Lastly, a potential factor that we are unable to test for is a subject's confidence in their performance expectations: subjects reported higher performance expectations for computer delegates, but we did not measure their confidence in the respective accuracy of these expectations. Burton et al. (2020) point out that human decision-making, like automated decision-making, operates through a black-box of its own: intuition. It may well be the case that performance expectations were, on average, higher for computer delegates, but subjects, intuitively, felt more certain of their performance expectations of human delegates, perhaps due to their shared humanity. If we assume some degree of risk and ambiguity aversion in our subjects, this could, at least in part, explain their overall preference for the human delegate.

Nevertheless, our results suggest that future research into delegation preferences for human versus computer delegates should focus on decision environment characteristics that go beyond mere differences in performance expectations, including the nature of the delegate's task and information on the computer delegate's programming.

References

- Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C., & Sugden, R. (2009). *Experimental Economics: Rethinking the Rules*. Princeton University Press.
- Bardsley, N., Mehta, J., Starmer, C., & Sugden, R. (2010). Explaining Focal Points: Cognitive Hierarchy Theory versus Team Reasoning. *The Economic Journal*, 120(543), 40–79.
- Bartling, B., Fehr, E., & Herz, H. (2014). The Intrinsic Value of Decision Rights. *Econometrica*, 82(6), 2005–2039.
- Ben-David, D. & Sade, O. (2021). Robo-Advisor Adoption, Willingness to Pay, and Trust — Before and at the Outbreak of the Covid-19 Pandemic. *Available at SSRN: <https://ssrn.com/abstract=3361710> or <http://dx.doi.org/10.2139/ssrn.3361710>*.
- Bigman, Y. E. & Gray, K. (2018). People Are Averse to Machines Making Moral Decisions. *Cognition*, 181, 21–34.
- Bobadilla-Suarez, S., Sunstein, C. R., & Sharot, T. (2017). The Intrinsic Value of Choice: The Propensity to Under-delegate in the Face of Potential Gains and Losses. *Journal of Risk and Uncertainty*, 54(3), 187–202.
- Bohnet, I., Greig, F., Herrmann, B., & Zeckhauser, R. (2008). Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review*, 98(1), 294–310.
- Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A Systematic Review of Algorithm Aversion in Augmented Decision Making. *Journal of Behavioral Decision Making*, 33(2), 220–239.
- Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, 56(5), 809–825.
- Charness, G., Gneezy, U., & Halladay, B. (2016). Experimental Methods: Pay One or Pay All. *Journal of Economic Behavior & Organization*, 131, 141–150.
- Christin, A. (2017). Algorithms in Practice: Comparing Web Journalism and Criminal Justice. *Big Data & Society*, 4(2).
- Cubitt, R. P., Starmer, C., & Sugden, R. (1998). On the Validity of the Random Lottery Incentive System. *Experimental Economics*, 1, 115–131.
- Cubitt, R. P. & Sugden, R. (2003). Common knowledge, salience and convention: A reconstruction of david lewis’ game theory. *Economics & Philosophy*, 19(2), 175–210.
- Danz, D., Kübler, D., Mechtenberg, L., & Schmid, J. (2015). On the Failure of Hindsight-Biased Principals to Delegate Optimally. *Management Science*, 61(8), 1938–1958.

- Deese, J. (1959). On the Prediction of Occurrence of Particular Verbal Intrusions in Immediate Recall. *Journal of Experimental Psychology*, 58(1), 17.
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors*, 44(1), 79–94.
- Eastwood, J., Snook, B., & Luther, K. (2012). What People Want from Their Professionals: Attitudes Toward Decision-Making Strategies. *Journal of Behavioral Decision Making*, 25(5), 458–468.
- Fehr, E., Herz, H., & Wilkening, T. (2013). The Lure of Authority: Motivation and Incentive Effects of Power. *American Economic Review*, 103(4), 1325–59.
- Gogoll, J. & Uhl, M. (2018). Rage Against the Machine: Automation in the Moral Domain. *Journal of Behavioral and Experimental Economics*, 74, 97–103.
- Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why Are We Averse towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion. *Research Papers (168)*, available at: https://aisel.aisnet.org/ecis2020_rp/168.
- Kaufmann, E., Chacon, A., Kausel, E. E., Herrera, N., & Reyes, T. (2023). Task-Specific Algorithm Advice Acceptance: A Review and Directions for Future Research. *Data and Information Management*, (pp. 100040).
- Lee, M. K. (2018). Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management. *Big Data & Society*, 5(1).
- Lewis, D. (1969). *Convention: A Philosophical Study*. Harvard University Press.
- Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022). What Influences Algorithmic Decision-Making? A Systematic Literature Review on Algorithm Aversion. *Technological Forecasting and Social Change*, 175, 121390.
- Mehta, J., Starmer, C., & Sugden, R. (1994a). Focal Points in Pure Coordination Games: An Experimental Investigation. *Theory and Decision*, 36, 163–185.
- Mehta, J., Starmer, C., & Sugden, R. (1994b). The Nature of Salience: An Experimental Investigation of Pure Coordination Games. *The American Economic Review*, 84(3), 658–673.
- Neri, C. & Rommeswinkel, H. (2017). Decision Rights: Freedom, Power, and Interference. Available at SSRN: <https://ssrn.com/abstract=2485107> or <http://dx.doi.org/10.2139/ssrn.2485107>.
- Niszczoła, P. & Kaszás, D. (2020). Robo-Investment Aversion. *PLoS One*, 15(9).
- Owens, D., Grossman, Z., & Fackler, R. (2014). The Control Premium: A Preference for Payoff Autonomy. *American Economic Journal: Microeconomics*, 6(4), 138–61.

- Palan, S. & Schitter, C. (2018). Prolific.ac — A subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance*, 17, 22–27.
- Rai, A. (2020). Explainable AI: From Black Box to Glass Box. *Journal of the Academy of Marketing Science*, 48, 137–141.
- Schelling, T. (1960). *The Strategy of Conflict*. Harvard University Press.

CHAPTER 2

Non-Realised Values and Agency: Positive Framing Can Increase Acceptance of Offers of Automated Assistance No Matter Who Makes the Offer[†]

1 Introduction

Persistent underreliance of human decision-makers on automated agents has proven to be a critical barrier to reaping the benefits of ongoing technological advancements: evidence of this spans a wide variety of topics, including but not limited to economic forecasting (Abeliuk et al., 2020; Dietvorst et al., 2018, 2015), autonomous vehicles (Goldbach et al., 2019; Bonnefon et al., 2016), and healthcare recommendations (Longoni et al., 2019).

Many of the potential benefits of automation reliance have been attributed to the fact that automated agents are not as prone to cognitive biases as their human counterparts (see e.g. Liaudinskas, 2022; Agrawal et al., 2019; Miyoshi & Matsubara, 2018). Indeed, human-computer interactions are often characterised by higher levels of rationality, strategic reasoning, and efficiency (March, 2021) and have also been shown to render humans less prone to emotional and social responses (Chugunova & Sele, 2022) than comparable human-human interactions. However, there is a large body of research that documents the similarities between human behaviour towards other humans and human behaviour towards computers (see e.g. Kirchkamp & Strobel, 2019; Lee & Liang, 2015; Lammer et al., 2014). In fact, Krämer et al. (2012), following a comparison of theories of human-human interactions to empirical evidence of human-computer interactions, go so far as to suggest that “it seems unlikely that one will be able to establish rules for human-agent/robot-interaction which radically depart from what humans know from and use in their everyday interactions.”

[†]This chapter is based on joint work with Robin Cubitt and Chris Starmer. I am grateful to NIBS (Network for Integrated Behavioural Science) for funding this research. I also thank CeDEx members and participants at the 2022 CCC conference for comments and feedback. The study received ethical approval from the Nottingham School of Economics Research Ethics Committee on 18/05/2022.

Nass & Moon (2000) conceptualised this idea in their *Computers As Social Actors* (CASA) paradigm: in a series of experiments, it is shown that humans ‘mindlessly’ apply human social categories such as gender and racial stereotyping (Nass et al., 1997; Lee et al., 2000) or overlearned social behaviours such as politeness, reciprocity and reciprocal self-disclosure (Fogg & Nass, 1997; Moon, 2000) to computers¹. If computers are, in fact, treated as social actors rather than machines, this opens up a new avenue of investigation into the determinants of underreliance on automation: we can, in theory, design unbiased computers, but the cognitive biases humans commonly exhibit in their interactions with other humans may carry over into their interactions with computers.

Several studies support the prevalence of social norms in human-computer interactions. Subjects in Fogg & Nass (1997) perform significantly more tasks for a computer if that same computer had proven helpful to them in a previous task than they do when matched with a different computer that had not yet helped them. Similarly, Sandoval et al. (2016) report equal levels of reciprocity towards humans and computers in a Repeated Prisoner’s Dilemma Game as well as in an Ultimatum Game. Kim et al. (2021) furthermore provide evidence of not just the existence but also the complexity of reciprocity towards automated agents: subjects in their study perceive the automated agent as less agreeable and less attractive as a social interactant when the automated agent’s help is contingent on *their* reciprocating said help. Mol et al. (2020) investigate cheating behaviour in a Virtual Reality adaption of the mind game from Jiang (2013). An active virtual observer does not significantly reduce cheating behaviour relative to a control condition in which behaviour is not observed; crucially, however, their findings demonstrate that an active virtual observer, relative to a passive virtual observer, can intensify reputational concerns and significantly reduce cheating behaviour. In other words, much like with a human observer, the behaviour of a virtual observer influences an individual’s likelihood to engage in cheating behaviour.

In summary, human-computer interactions appear to bear a similar complexity to human-human interactions that has, however, been explored to a much lesser degree. In line with the unconscious application of social behaviours and expectations under CASA, Sandoval et al. (2016) propose the Media Equation Theory (Reeves & Nass, 1996) as an explanation for the similarities between human-human and human-computer interactions in their study: humans are frequently shown to treat objects, including automated agents, as they would sentient human beings (Epley et al., 2007). According to Kim & Sundar (2012), this type of behaviour is not contingent on anthropomorphic cues that trigger conscious attributions of humanness but rather the result of unconscious applications of anthropomorphism towards human and non-human entities alike.

¹Nass & Moon (2000) use the term ‘mindless’ as the opposite of ‘mindful’; in order to avoid confusion due to its alternate meanings of ‘thoughtless’ or even ‘stupid’, we hereafter replace their terminology with ‘conscious’ and ‘unconscious’.

Conversely, it has also been suggested that a computer’s intentionality is derived from that of its human programmer (Searle, 1980; Searle et al., 1983; Nanay, 2006)². Johnson (2006) argues that automated agents are not moral agents in their own right because “they have meaning and significance only in relation to human beings” but still inherit the intentionality of their human programmer and should thus not be dismissed from the realm of morality.

Intentionality has been shown to be a decisive element of many human-human interactions. McCabe et al. (2003) compare cooperative behaviour in voluntary and involuntary trust games, and show that, holding constant the action of the first mover, second-mover cooperation depends on whether or not the first mover had an outside option available to them. Furthermore, Falk et al. (2003) have famously demonstrated that non-realised values are a critical factor in whether or not identical offers in an Ultimatum Game are accepted. Given that there is some evidence to suggest that people do, in fact, ascribe a moderate degree of agency to automated agents (Gray et al., 2007), we could imagine a similar effect for an automated offer of assistance. Much like its human equivalent, an automated offer of assistance may be perceived as more or less helpful as well as more or less kind depending on one’s beliefs about the computer’s capabilities and intentions, which could prove pivotal for the decision of whether or not to accept it.

We set out to investigate this in an experiment in which subjects had to decide whether to rely, for successful completion of a perceptual task, on their own performance or that of an objectively superior computer. In our study, computers performed the task with a specific success probability that was, on average, greater than that of a human decision-maker; however, we varied the way in which this success probability was presented to the subjects in order to examine non-realised values and agency as causal determinants of underreliance on automation. Specifically, our treatment groups varied (1) whether the computer success probability was chosen from a set and, if so, whether it constituted the ‘good’ or the ‘bad’ outlier within this set, and (2) whether agency over the choice of the success probability was ascribed to the computer itself or to the experimenter. Together, these manipulations allow us to explore the effect of non-realised values on the acceptance of otherwise identical offers of automated assistance, and whether or not it differed depending on who was framed as the responsible party in making the offer. Furthermore, we measured emotional reactions to and kindness perceptions of the offer of automated assistance as potential mechanisms for these effects.

²Similar arguments have been made in the context of ‘object betrayal’, where an aversion to betrayal by an inanimate object may simply reflect feelings towards the designer or supplier of the object (Koehler & Gershoff, 2003; Cubitt et al., 2017).

Our results suggest that automation reliance is sensitive to non-realised values and that it may be possible to decrease underreliance by framing an algorithm as *superior* to others which might have been available. We furthermore find that this effect was linked to positive emotional reactions and, to a lesser degree, positive kindness perceptions. Conversely, we find no significant effect of whether agency over the choice of the success probability was ascribed to the computer itself or to the experimenter: subjects exhibited small differences in emotional reactions and kindness perceptions, which did, however, not translate into differences in automation reliance. Overall, our results are consistent with the unconscious application of social behaviours and expectations to humans and computers alike.

The paper proceeds as follows: in Section 2, we describe the design of the experiment and outline our research hypotheses; Section 3 contains the results of our data analysis, while Section 4 discusses these results and concludes.

2 Experimental Design

Our main variable of interest is a subject’s automation reliance decision in a variation of the perceptual task from Hollard et al. (2016); we will hereafter refer to this task as the Dot Task. Subjects performed ten rounds of the Dot Task after choosing whether to have their payoff depend on their own performance or that of a computer in a randomly selected round of the task (for evidence on the validity of this incentivisation method, see Chapter 6 in Bardsley et al., 2009; Cubitt et al., 1998; Charness et al., 2016).

Let us first establish two key concepts that we rely on throughout our analysis: *automation reliance* and *underreliance on automation*. In the context of our experiment, automation reliance refers to a subject’s decision to have their payoff depend on the performance of the computer in a randomly selected round of the Dot Task. Conversely, a subject’s decision to have their payoff depend on their own performance will be labelled *self-reliance*. *Underreliance* can be defined as: (1) non-reliance on a computer that would, on average, outperform a human subject, or (2) non-reliance on a computer that a human subject believed would outperform them³. We hereafter distinguish the two definitions as (1) *objective underreliance* (OUR) and (2) *subjective underreliance* (SUR). We focus primarily on the former, but also elicited individual expectations of a subject’s own performance and that of the computer to garner insights into the latter.

In each round of the Dot Task, subjects were shown two circles and had to select the one containing the higher number of dots: the *winning circle*. We constructed the circles with the aim of setting the average human success probability q to be below the implemented computer success probability p , such that any instance of non-reliance on the computer’s performance would constitute a case of objective underreliance in line with the above definition. Details on the Dot Task will be provided in Section 2.2.

2.1 Treatment Groups

Subjects were randomly assigned to a BASELINE condition or one of four treatment groups. In the BASELINE, the computer success probability was simply set to $p = 0.60$. In the four remaining treatment groups, the computer success probability was selected from a set in which p was either the bad outlier (BO) or the good outlier (GO), and the selection of the computer success probability was framed as either the computer’s choice (CC) or the experimenter’s choice (EC).

³Note that our interpretations of underreliance are performance-based and, as such, implicitly assume that there was no intrinsic utility to be derived from a subject relying on their own performance in the Dot Task. Self-reliance might, for some subjects, generate utility beyond that arising from material consequences; however, our interpretations follow public discourse of real-world examples like autonomous vehicles, which focuses on the conflict between favourable performance expectations of and the unwillingness to utilise automation.

Fig. 2.1 illustrates the resulting treatment partition. In BO-CC and BO-EC, p constituted the bad outlier in a set comprising the following five possible computer success probabilities: $\{0.60; 0.67; 0.68; 0.69; 0.70\}$. In GO-CC and GO-EC, on the other hand, p constituted the good outlier in a set comprising the following five possible computer success probabilities: $\{0.50, 0.51, 0.52, 0.53, 0.60\}$. In BO-CC and GO-CC, the selection of the computer success probability was framed as the computer’s choice; in BO-EC and GO-EC, it was framed as the experimenter’s choice.

Whenever the computer success probability was selected from one of the above sets, we used a non-uniform random device that selected p with a probability of 0.96 and each of the four alternative computer success probabilities with a probability of 0.01. Subjects were informed that every one of the five success probabilities in the set was possible but not equally likely to be chosen. In doing so, we ensured that the vast majority of subjects were presented with the same computer success probability, p , without invalidating any of the non-realised values as possible success probabilities. Subjects for whom a computer success probability other than p was drawn completed the experiment but were afterwards excluded from our analysis; hence, we hereafter use p to denote the implemented computer success probability. Consequently, our first treatment manipulation varied whether the implemented computer success probability was selected from a set and, if so, whether it was a ‘good’ or a ‘bad’ offer relative to the other, non-realised values.

	P = 0.60 BAD OUTLIER	P = 0.60 GOOD OUTLIER
P NOT CHOSEN	BASELINE	
P CHOSEN BY COMPUTER	BO-CC	GO-CC
P CHOSEN BY EXPERIMENTER	BO-EC	GO-EC

Note: BO denotes the Bad Outlier, GO the Good Outlier; CC indicates that the selection of the computer success probability was framed as the computer’s choice, whereas EC indicates that it was framed as the experimenter’s choice.

Figure 2.1: Treatment Groups

Our second treatment variation, on the other hand, varied who was framed as the responsible party in making this offer, enabling us to test whether automation reliance differed depending on whether the selection of p was ascribed to the computer itself or to the experimenter.

Subjects in the BASELINE were simply told that “the computer uses an algorithm that will, in each separate round, select the winning circle with a 60% chance of success”. Meanwhile, subjects in BO-CC and GO-CC were informed that “there are five algorithms for **the computer** to choose from, each with a different chance of success” and were then presented with the respective set of success probabilities as outlined above. After clicking a button to reveal the algorithm, subjects were told that “**the computer** has chosen the algorithm that will, in each separate round, select the winning circle with a 60% chance of success”⁴. Conversely, subjects in BO-EC and GO-EC were informed that “there are five algorithms for **the experimenter** to choose from, each with a different chance of success” and that “**the experimenter** has chosen the algorithm that will, in each separate round, select the winning circle with a 60% chance of success” (see Appendix B.2 for detailed instructions).

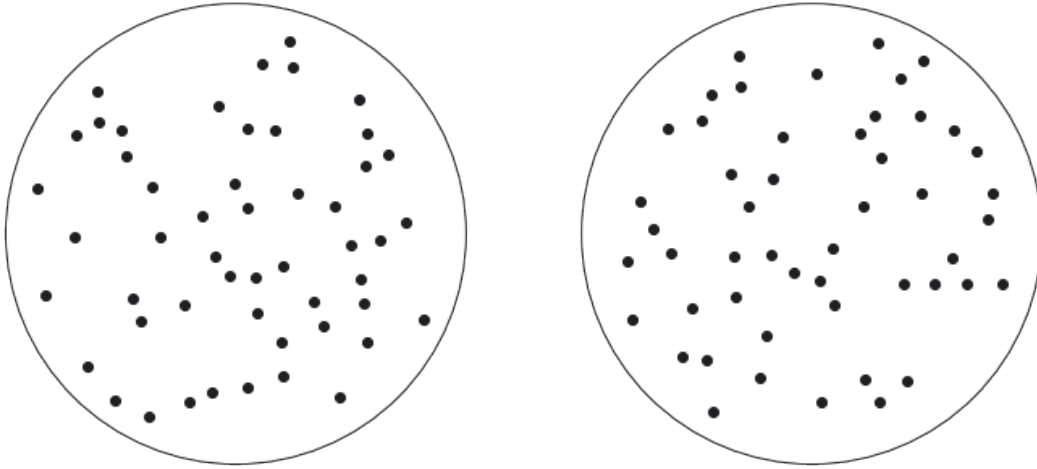
2.2 The Dot Task

The Dot Task worked as follows: in each round, subjects were shown two circles for five seconds, one of which always contained a higher number of dots than the other. After the dots disappeared, subjects were asked to select the circle that contained the higher number of dots, which was referred to as the *winning circle* for that round. We wanted our subjects to infer that each round of the Dot Task would be of similar difficulty, so they were furthermore informed that: one, the winning circle contained the same number of dots in each round, and two, the difference in the number of dots between the two circles was also the same in each round.

Unbeknownst to the subjects, the number of dots was fixed at 51 dots in the winning circle versus 50 dots in the other circle. Hence, identifying the circle with the higher number of dots was not trivial and, given the time limit of five seconds, it was furthermore impossible for subjects to count the numbers of dots in both circles to determine the winning circle; in other words, the best a subject could do was choose a circle at random. We aimed to thereby set the average human success probability close or equal to $q = 0.50$ such that, given the implemented computer success probability p , any instance of non-reliance on the computer’s performance constituted a case of objective underreliance in line with the previous definition.

⁴This is, of course, applicable only to the majority of cases in which p was selected; when a different computer success probability was selected, the message reflected this.

We chose the Dot Task for two reasons. Firstly, it allowed for a high degree of control of human success probabilities via the total number of dots in each circle and the difference in the number of dots between the two circles; specifically, it allowed us to set the average human success probability in each round of the task close or equal to q . Secondly, we expected most of our subjects to be able to correctly infer the average human success probability to be close or equal to q even if we did not explicitly state it⁵.



Note: The left-hand circle contains 51 dots, whereas the right-hand circle contains only 50 dots, making the left-hand circle the winning circle in this pair.

Figure 2.2: Dot Task based on Hollard et al. (2016)

Our experiment consisted of two stages. In Stage 1, subjects performed a series of ten unincentivised practice rounds of the Dot Task to familiarise themselves with the task and recognise the average human success probability to be close or equal q . Subjects did, however, not receive any feedback on their performance in these practice rounds. In Stage 2, subjects performed another series of ten rounds of the Dot Task. This time, however, they first received the treatment-specific information about the underlying process that determined the computer success probability. Subjects were then asked to make their automation reliance decision in choosing whether to have their payoff depend on their own performance or that of the computer in a randomly selected round of the Dot Task.

⁵Due to the inevitable variance in human performance, of course individual performances could and did differ from q ; however, we designed the circles such that, if a human were to perform the Dot Task indefinitely, we would expect their success probability to converge to q . In Section 3.2, we test whether we succeeded in achieving an average human success probability close or equal to q and whether average human performance expectations reflected an understanding of this.

Our reasoning for not providing feedback in the practise rounds is twofold. Firstly, q was an *expectation* of the average human success probability; in the unlikely event that the average human success probability had, instead, turned out to be equal to or even exceed p , feedback would have removed performance-based incentives for automation reliance in the Dot Task for a majority of subjects. Secondly, we hypothesised that leaving subjects to estimate their own success probabilities would generate the necessary variation in performance expectations to create two distinct groups: subjects who engaged in subjective underreliance and subjects who did not.

If a subject chose to rely on their own performance and selected the winning circle in the payoff-relevant round, they received a bonus of £1; if they did not select the winning circle, they received nothing. Conversely, if a subject chose to rely on the computer’s performance and the computer selected the winning circle in the payoff-relevant round, they received a bonus of £1; if the computer did not select the winning circle, they received nothing. Finally, subjects were not given any information in advance regarding whether they would receive feedback on their own performance or that of the computer⁶.

Freer et al. (2023) establish the reduction of decision cost as a significant motive for delegating financial decisions to another human, which we suspected might apply to our decision environment as well. Irrespective of a subject’s automation reliance decision, they thus knew that they would always perform the Dot Task alongside the computer; this was done so that subjects did not choose to rely on the computer’s performance just to avoid performing the task themselves. It is also worth noting that Lobeck (2023) suggests that subjects may choose to rely on their own, inevitable performance to give it meaning, which is a possibility we have to take into account when interpreting absolute levels of self-reliance; it should, however, not affect treatment comparisons. Having subjects perform the Dot Task irrespective of their automation reliance decision furthermore ensured that we collected performance data from all subjects rather than just the ones who chose not to rely on the computer.

2.3 Additional Measurements

In between each subject’s automation reliance decision and the first round of the Dot Task, we collected six additional measurements to provide us with insights into possible motivations for choosing (not) to rely on the computer. We collected these measurements *before* the performance of the Dot Task because we are interested in expectations, emotional reactions and kindness perceptions at the time of making the automation reliance decision, and performing the Dot Task might have distorted these.

⁶We thereby hoped to avoid — in as far as this was possible — introducing regret aversion as an influencing factor for the automation reliance decision.

First, we elicited a subject’s expectations of the computer’s performance as well as their own: we had the subjects guess in how many of the ten rounds they expected to select the winning circle, and in how many of the ten rounds they expected the computer to do so. Conveniently, these performance expectations not only allow us to connect individual automation reliance decisions with differences in relative payoff expectations, but we can also use them to test whether subjects did, in fact, expect their success probability to be below that of the computer. We incentivised these elicitations via a bonus of £0.50 for each guess that matched a subject’s own or the computer’s actual performance. To avoid order effects, the order in which we elicited the two expectations was randomised.

Next, we measured emotional reactions to the revelation of the computer’s algorithm and perceptions of its kindness using two sets of five-point Likert scales consisting of eight items each. Reminding subjects of the moment in which p was revealed to them, we asked them to evaluate: (1) the extent to which they felt HAPPY, CONTENT, GRATEFUL, CALM, ANGRY, ANNOYED, DISAPPOINTED, and SAD, and (2) the extent to which they considered the computer’s algorithm KIND, FAIR, FRIENDLY, GENEROUS, MEAN, UNFAIR, HOSTILE, and UNCHARITABLE⁷.

We anticipated that risk and ambiguity preferences might also influence automation reliance decisions: the computer success probability was known to the subjects, but they received no feedback on their own performance in the practise rounds of the Dot Task. In other words, evaluations of reliance on the computer’s performance were governed by risk preferences, whereas relying on one’s own performance likely depended on ambiguity preferences as well. For this reason, we measured ambiguity preferences using the parsimonious *Ambiguity Preference Survey Module* (Cavatorta & Schröder, 2019) and risk preferences using the *General Risk Question* (Dohmen et al., 2011).

Finally, we measured hypothetical willingness-to-pay (WTP) to maintain one’s choice in the automation reliance decision as an indicator of strength of preference. We forced subjects to choose between self-reliance and automation reliance, so we need to account for the possibility of indifference between the two, which WTP values close or equal to zero would point towards. Subjects indicated their hypothetical WTP using a slider that allowed them to choose any value between £0.00 to £1.00 (the minimum and maximum earnings that could result from the automation reliance decision) in 0.01 increments (see Appendix B.2 for detailed instructions).

⁷Our measures are modelled after the PANAS (Watson et al., 1988). We consider four dimensions: positive and negative emotional reactions, as well as positive and negative kindness perceptions; each measured via four similar but not identical affective states that we later aggregate into a robust score. Further details will be provided in Section 3.3.

2.4 Hypotheses

Frequent findings in delegation experiments are the preference for control over one’s payoff (Owens et al., 2014; Bobadilla-Suarez et al., 2017) as well as the reluctance to give up one’s decision rights (Fehr et al., 2013; Bartling et al., 2014; Neri & Rommeswinkel, 2017). While our study differed significantly from most of these delegation experiments, it shares one of their central features in that there was a possibility of self-reliance bias. Given that previous research into automation reliance furthermore suggests an overall aversion to algorithms (Burton et al., 2020; Jussupow et al., 2020), we expected to observe a significant level of underreliance across all treatment groups including the BASELINE. If this was the case, self-reliance bias would need to be taken into account when interpreting absolute levels of underreliance on automation; it should, however, not affect comparisons *between* treatment groups.

Hypothesis 1. *We observe objective underreliance when p is not chosen from a set.*

Our first treatment manipulation was designed to influence a subject’s beliefs about where p ranked in the range of the computer’s performance capabilities. Simply put, when p was the good outlier, subjects were encouraged to believe that the computer would perform the Dot Task to the best of its ability; conversely, when p was the bad outlier, subjects were encouraged to believe that the computer would perform the Dot Task to the worst of its ability. Following the establishment of the importance of non-realised values in human-human interactions (Falk et al., 2003) and the large body of evidence supporting the CASA paradigm (see e.g. Fogg & Nass, 1997; Nass et al., 1997; Lammer et al., 2014; Lee & Liang, 2015; Sandoval et al., 2016), we expected automation reliance decisions to reflect this difference.

Hypothesis 2. *Objective underreliance is higher when p is the bad outlier than when p is the good outlier.*

Our second treatment manipulation addresses the issue of agency: we varied whether the selection of the computer’s algorithm was framed as the computer’s choice or that of the experimenter. CASA predicts the unconscious application of social behaviours to computers and humans alike; consequently, under CASA, we would expect similar levels of automation reliance between the two differing ascriptions of agency. If a computer’s intentionality is, indeed, derived from that of its human programmer (Searle, 1980; Searle et al., 1983; Johnson, 2006; Nanay, 2006), the relative rank of p within the set of algorithms might, however, interact with the ascription of agency: to be precise, emphasising the role of the experimenter could intensify the effect of the relative rank of p within the set of algorithms as outlined below.

Hypothesis 3. *When p is the bad outlier, objective underreliance is higher when we present it as the experimenter’s choice rather than that of the computer; when p is the good outlier, objective underreliance is lower when we present it as the experimenter’s choice rather than that of the computer.*

We investigate emotional reactions to and kindness perceptions of the automated offer of assistance as underlying mechanisms for the hypothesised treatment effects. If non-realised values played a role in how an offer of automated assistance is received, we would first of all expect p to be associated with more positive emotional reactions and kindness perceptions when it was the good outlier than when it was the bad outlier. Secondly, as before, under CASA we would expect to see similar emotional reactions and kindness perceptions no matter who was framed as the responsible party in making this offer. If a computer’s intentionality is, indeed, derived from that of its human programmer, we might, however, expect emotional reactions and kindness perceptions to be more pronounced when the selection of p was framed as the experimenter’s choice rather than the computer’s choice. In line with this theory, Schniter et al. (2020) find significant differences in social emotions like guilt, gratitude and anger between robot and human trustees: interactions with human trustees triggered more intense social emotions compared to interactions with robots.

Hypothesis 4a. *Emotional reactions and kindness perceptions are more positive when p is the good outlier than when p is the bad outlier.*

Hypothesis 4b. *Emotional reactions and kindness perceptions are more pronounced when we present p as the experimenter’s choice rather than that of the computer.*

Finally, we expected to be able to establish a direct link between emotional reactions to and kindness perceptions of the offered algorithm and observed levels of underreliance on automation.

Hypothesis 5a. *Positive emotional reactions to p are negatively correlated with objective underreliance; negative emotional reactions to p are positively correlated with objective underreliance.*

Hypothesis 5b. *Positive kindness perceptions of p are negatively correlated with objective underreliance; negative kindness perceptions to p are positively correlated with objective underreliance.*

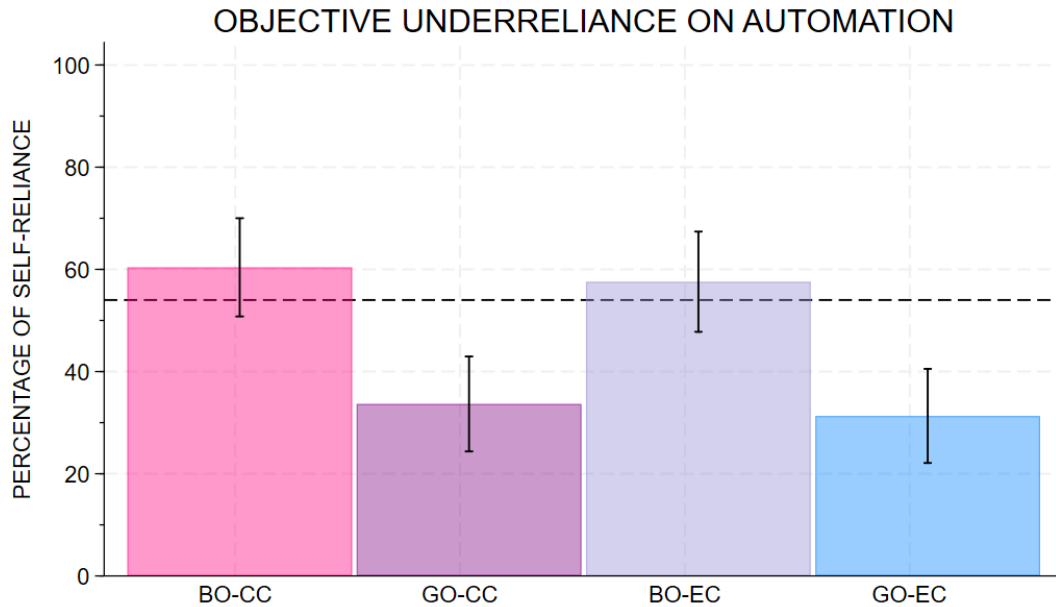
Note that we have formulated all of our hypotheses for objective underreliance. Where possible, we performed the same analyses for the narrower definition of subjective underreliance, but were, as we will later see, in some cases restricted by sample size.

3 Results

The experiment was programmed using Qualtrics software and conducted via Prolific (Palan & Schitter, 2018) in June 2022. The non-uniform random device, by design, presented a small number of subjects with a computer success probability other than p ; we recruited a total of 520 subjects, but excluded the 16 subjects for whom this was the case from our analysis. Of the remaining 504 subjects, 306 were female and the average age was 40.63 years⁸. All subjects received a participation fee of £3.00 and could earn a bonus payment of up to £2.00; the median completion time was 15 minutes and 37 seconds and the average payment, including bonuses, was £3.83.

3.1 Automation Reliance Decision

Our subjects were evenly divided across the BASELINE and the four remaining treatment groups, with 104 subjects in the BASELINE, 101 each in BO-CC and GO-CC, and 99 each in BO-EC and GO-EC. Overall, 239 subjects (47.42%) chose to rely on their own performance rather than that of the computer. In the BASELINE, 56 subjects (53.85%) chose to do so, which supports **Hypothesis 1**.



Note: ‘Objective Underreliance on Automation’ is defined as non-reliance on the computer. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark of 53.85%.

Figure 2.3: Proportions of Objective Underreliance on Automation

⁸Demographic sample characteristics and a treatment breakdown thereof can be found in Tables B.1 and B.2 in Appendix B.1.

Result 1. *Subjects exhibited significant levels of objective underreliance both overall and when p was not chosen from a set.*

Fig 2.3 illustrates the treatment-specific proportions of objective underreliance, with whiskers indicating the 95% confidence intervals and a dashed line marking the BASELINE proportion of 53.85% as a benchmark; Table 2.1 shows the corresponding between-treatment comparisons of the proportions of objective underreliance. In line with **Hypothesis 2**, we find that there is a significant difference in the proportion of objective underreliance between the treatment groups in which p was the bad outlier and those in which p was the good outlier: subjects relied on the computer’s performance significantly less often when its success probability was inferior to the non-realised values in the set than when it was superior, even if, from a pecuniary perspective, these non-realised values should not have affected automation reliance decisions.

Notably, however, there is also a significant difference between the BASELINE and GO-CC as well as between the BASELINE and GO-EC; conversely, we do not observe a significant difference between the BASELINE and BO-CC nor between the BASELINE and BO-EC. In other words, the proportion of underreliance when p was not chosen from a set was similar to the proportion of underreliance when it was the bad outlier but not when it was the good outlier.

We consider three possible interpretations of this result. One, in the absence of a set to judge p by, subjects formed their own priors about the true capabilities of the computer, and these priors were such that p was considered a bad outlier in the BASELINE just like it was in BO-CC and BO-EC. Two, the perception of p as a bad outlier was exclusive to BO-CC and BO-EC, but — unlike the perception of p as a good outlier — it did not affect automation reliance decisions. Three, the perception of p as a bad outlier in BO-CC and BO-EC *did* affect automation reliance decision, but this effect was small and we were underpowered to detect its statistical significance.

Our experiment does not allow us to formally test these three interpretations against one another. However, our first interpretation is consistent with Burton et al. (2020), who posit that “a human decision-maker rarely, if ever, confronts an algorithm with a blank slate” but instead approaches automation reliance decisions against the backdrop of expectations and experiences with automated agents in the real world. We do, on the other hand, see a small increase in the proportion of objective underreliance in BO-CC and BO-EC relative to the BASELINE, which is in line with our third interpretation. Either way, our results strongly suggest that it might be possible to reduce underreliance on automation via the positive framing of the computer’s success probability.

Result 2. *Underreliance was significantly lower when p was the good outlier than when p was the bad outlier or when p was not chosen from a set.*

	OUR ₁	OUR ₂	p (H_0 : OUR ₁ = OUR ₂) <i>Two-Sample Test of Proportion</i>
BASELINE — BO-CC	0.54	0.60	0.3435
BASELINE — GO-CC	0.54	0.34	0.0036
BASELINE — BO-EC	0.54	0.58	0.5929
BASELINE — GO-EC	0.54	0.31	0.0012
BO-CC — GO-CC	0.60	0.34	0.0001
BO-EC — GO-EC	0.58	0.32	0.0002
BO-CC — BO-EC	0.60	0.58	0.6851
GO-CC — GO-EC	0.34	0.31	0.7227

Note: OUR₁ and OUR₂ denote the two proportions of objective underreliance in each row for whose between-treatment equality we test using Two-Sided Two-Sample Tests of Proportion.

Table 2.1: Treatment Comparisons of Objective Underreliance (OUR)

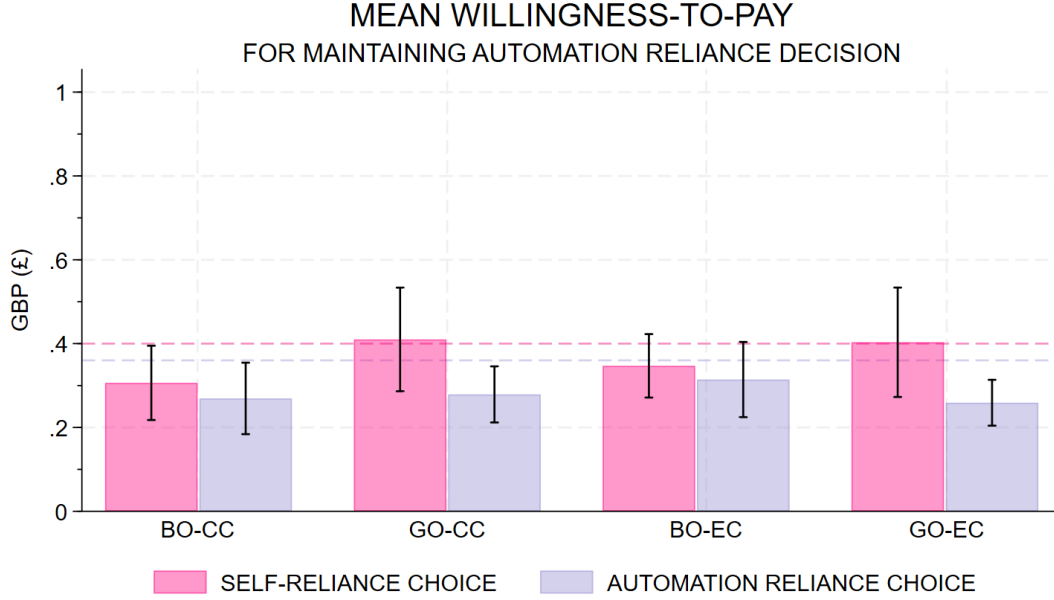
Hypothesis 3, on the other hand, is not supported: the proportion of objective underreliance when p was presented as the experimenter’s choice did not significantly differ from the proportion of objective underreliance when it was presented as that of the computer. In other words, it does not appear to matter who was framed as the responsible party in selecting p , which is consistent with the unconscious application of social behaviours and expectations to humans and computers alike⁹.

Result 3. *Underreliance was not significantly affected by whether p was presented as the experimenter’s choice or that of the computer itself.*

Lastly, subjects were forced to rely either on their own performance or that of the computer with no outside option; we therefore elicited hypothetical WTP to maintain their chosen option as a measure of strength of preference. As shown by the dashed lines in Fig 2.4, average WTP in the BASELINE was £0.40 for those who chose to rely on their own performance and £0.36 for those who chose to rely on the computer’s performance. Average WTP was significantly higher than zero in all treatment groups and regardless of automation reliance decisions ($p < 0.01$, One-Sided Sign Test). We recognise that this might, in part, reflect an endowment effect, but the consistent pattern of higher mean WTP for self-reliance than for automation reliance suggests that the significant levels of objective underreliance we observe were not just the result of random tie-breaker choices in the face of indifference between automation reliance and self-reliance¹⁰.

⁹Another possible explanation for this is the subtlety of our treatment manipulation, which relied solely on differences in wording between treatment groups. Given a situation in which the ascription of agency is more salient, it may still affect the level of underreliance: in an experiment that allows for endogenously determined success probabilities, one could, for example, have another subject choose the computer’s success probability and visualise this choice for the affected party or parties.

¹⁰In Section 3.4, we furthermore perform an econometric analysis of the treatment effects that controls for demographic characteristics, and revisit **Results 1 – 3**.

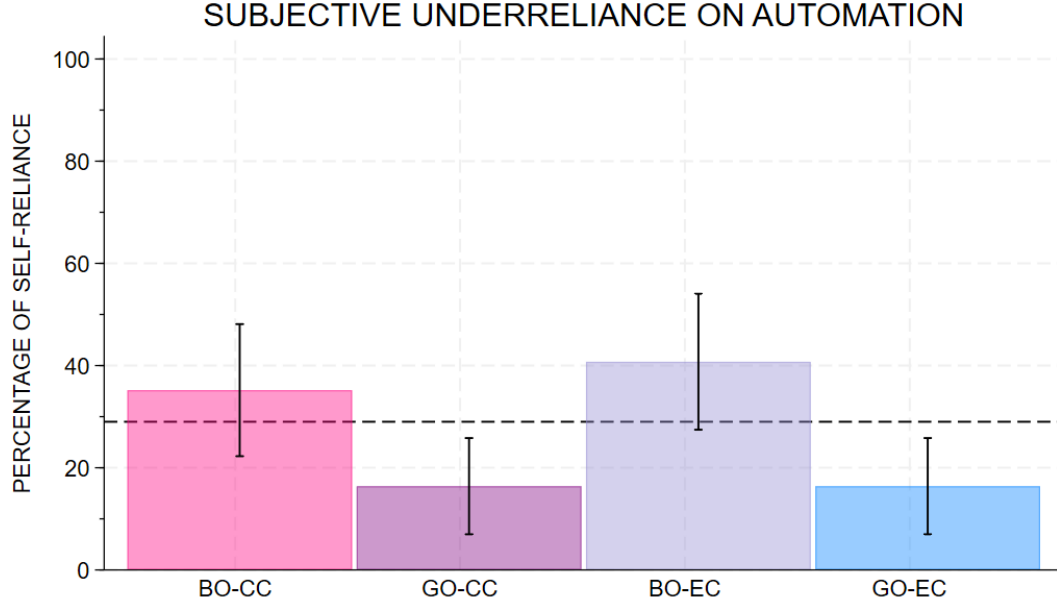


Note: £0 and £1 are minimum and maximum earnings from the automation reliance decision. Whiskers indicate 95% confidence intervals; dashed lines indicate the BASELINE benchmarks, 0.40 and 0.36.

Figure 2.4: Mean WTP for Maintaining Automation Reliance Decision

So far, we have looked at *objective* underreliance as per our earlier definition, which was, however, contingent on the understanding that the computer would, on average, outperform its human counterpart. In the absence of this (e.g. due to overconfidence in one’s own ability to perform the Dot Task or an inability to correctly interpret the implemented computer success probability), there might have been cases of non-reliance on the computer that could be ‘justifiable’ – even from a pecuniary perspective – given subjective payoff expectations from self-reliance versus subjective payoff expectations from automation reliance. We previously labelled these cases *subjective* underreliance: non-reliance on a computer that a human subject *believed* would outperform them.

We examine performance and performance expectations in detail in Section 3.2 but for now focus on the automation reliance decisions of the subsample of 288 subjects whose expectations of the computer’s performance exceeded their expectations of their own performance. Fig. 2.5 – 2.6 and Table 2.2 report analogous tests and illustrations to Fig. 2.3 – 2.4 and Table 2.1 for this subsample. First of all, we note that, of these 288 subjects, 78 subjects chose to rely on their own performance rather than that of the computer: in other words, 15.48% of the original sample and 27.08% of the subsample engaged in not just objective but also subjective underreliance on automation. This matches findings of a general aversion to algorithms in studies with similar set-ups (Burton et al., 2020; Jussupow et al., 2020) and is a convincing indicator of a preference for human agency that cannot be explained by performance expectations alone.



Note: ‘Subjective Underreliance on Automation’ is defined as a subject’s non-reliance on the computer when reporting higher performance expectations for the computer than for their own performance which, in total, 288 subjects did. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark of 29%.

Figure 2.5: Proportions of Subjective Underreliance on Automation

	SUR ₁	SUR ₂	p (H ₀ : SUR ₁ = SUR ₂) <i>Two-Sample Test of Proportion</i>
BASELINE — BO-CC	0.29	0.35	0.5059
BASELINE — GO-CC	0.29	0.16	0.0926
BASELINE — BO-EC	0.29	0.41	0.2045
BASELINE — GO-EC	0.29	0.16	0.0926
BO-CC — GO-CC	0.35	0.16	0.0206
BO-EC — GO-EC	0.41	0.16	0.0036
BO-CC — BO-EC	0.35	0.41	0.5519
GO-CC — GO-EC	0.16	0.16	1.0000

Note: SUR₁ and SUR₂ denote the two proportions of subjective underreliance in each row for whose between-treatment equality we test using Two-Sided Two-Sample Tests of Proportion.

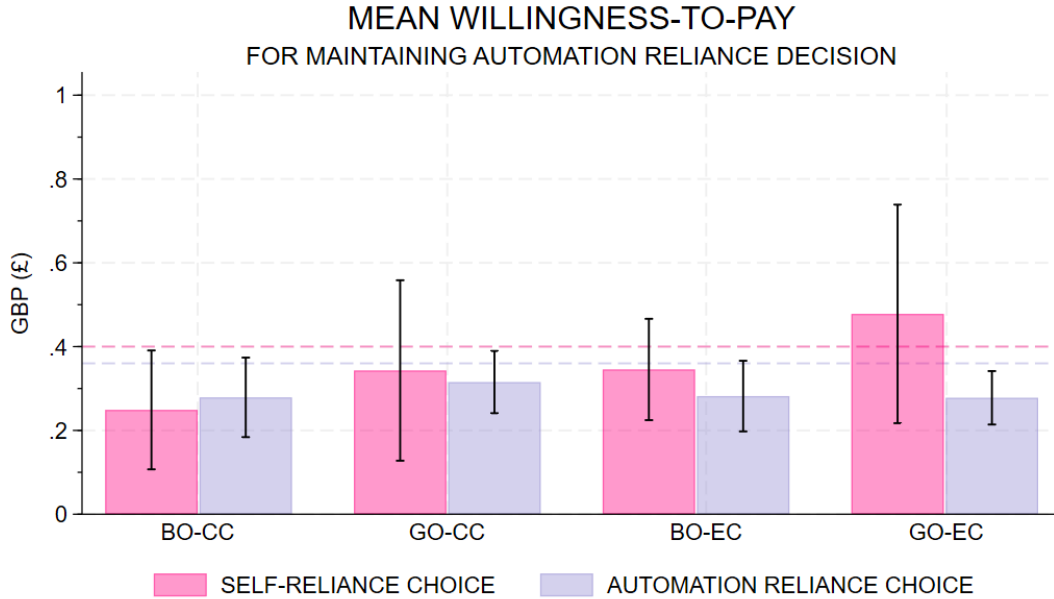
Table 2.2: Treatment Comparisons of Subjective Underreliance (SUR)

Fig 2.5 illustrates the treatment-specific proportions of subjective underreliance, with whiskers indicating the 95% confidence intervals and a dashed line marking the BASELINE proportion as a benchmark; Table 2.2 reports the results of the corresponding between-treatment comparisons of the proportions of subjective underreliance. While the proportion of subjective underreliance was consistently lower than that of objective underreliance in all treatment groups, **Results 1 — 3** also apply to the subsample.

Like in the case of objective underreliance, we find that there is a significant difference in the proportion of subjective underreliance between the treatment groups in which p was the bad outlier and those in which p was the good outlier. Again, there is also a significant difference in the level of subjective underreliance between the BASELINE and GO-CC as well as between the BASELINE and GO-EC; conversely, we do not observe a significant difference between the BASELINE and BO-CC nor between the BASELINE and BO-EC. In other words, here as well, the proportion of underreliance when p was not chosen from a set is similar to the proportion of underreliance when it was the bad outlier but not when it was the good outlier.

In addition to this, the proportion of subjective underreliance when p was presented as the experimenter's choice did not significantly differ from the proportion of subjective underreliance when it was presented as that of the computer. Again consistent with the unconscious application of social behaviours and expectations to humans and computers alike, it does not appear to matter who was framed as responsible for selecting p .

Finally, Fig. 2.6 confirms that, in the subsample, average WTP was also significantly higher than zero in all treatment groups and regardless of automation reliance decisions ($p < 0.01$, One-Sided Sign Test); this indicates that the significant levels of subjective underreliance we observe were not just the result of random tie-breaker choices in the face of indifference between automation reliance and self-reliance.

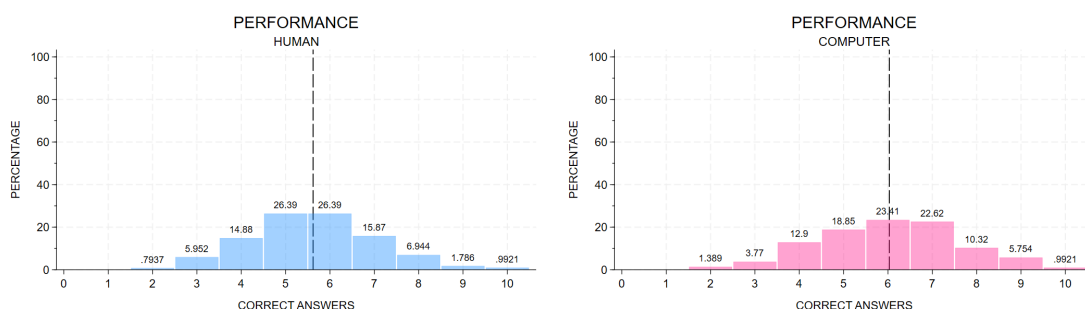


Note: WTP in the subsample of the 288 subjects who expected the computer to outperform them. £0 and £1 are minimum and maximum earnings from the automation reliance decision. Whiskers indicate 95% confidence intervals; dashed lines indicate the BASELINE benchmarks.

Figure 2.6: Mean WTP for Maintaining Automation Reliance Decision

3.2 Performance and Performance Expectations

So far, we have classed any instance of non-reliance on the computer’s performance as objective underreliance because the computer should, on average, have outperformed humans in the Dot Task. Fig. 2.7 shows that this was the case: while the mean number of correct answers for humans (5.62) was significantly higher than the expected mean of 5 ($p < 0.01$, Two-Sided One Sample z-Test), the mean number of correct answers for computers (6.03) was significantly higher still ($p < 0.01$, Wilcoxon Signed Rank Test) and — as was to be expected given $p = 0.60$ — not significantly different from 6 (see Fig. B.1 – B.5 in Appendix B.1 for treatment-specific illustrations of performance).

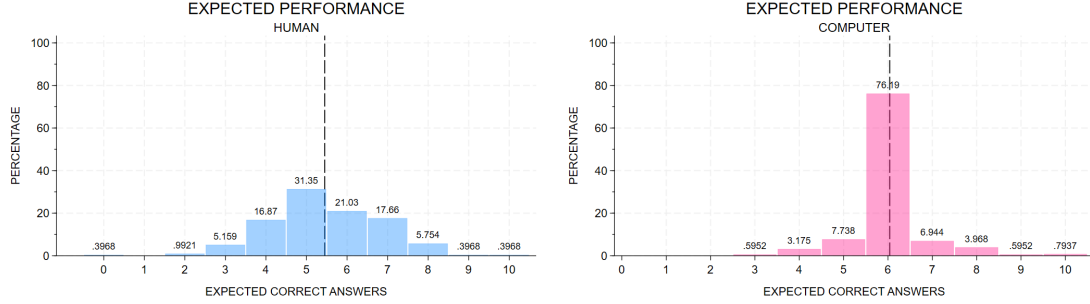


Note: Distribution of the number of correct answers in the ten rounds of the Dot Task. Dashed vertical lines indicate the mean numbers of correct answers, 5.62 and 6.03.

Figure 2.7: Performance

Next, the majority of subjects (76.19%) did translate p into an expectation of 6 correct answers for the computer, but only 54.77% reported an expectation of 5 or fewer correct answers for their own performance (see Fig. 2.8). At the aggregate level, performance expectations did, however, match actual performance: the mean number of expected correct answers for humans (5.45) was significantly higher than 5 ($p < 0.01$, Two-Sided One Sample z-Test), but the mean number of expected correct answers for computers (6.04) was significantly higher still ($p < 0.01$, Wilcoxon Signed Rank Test) and again not significantly different from 6. Indeed, there was no significant difference in computer performance and computer performance expectations. Human performance and human performance expectations, on the other hand, did differ ($p < 0.10$, Wilcoxon Signed Rank Test), but the subjects’ slight underconfidence in their own performance means that expectations of the human success probability were even closer to our target of q^{11} . In conclusion, not only did computers outperform our human subjects, but subjects — at the aggregate level — expected them to (see Fig. B.6 – B.10 in Appendix B.1 for treatment-specific illustrations of performance expectations).

¹¹Note, however, that if we compare own performance and own performance expectations at the individual level, we find that the two are uncorrelated; this is, perhaps, not surprising in a task that was designed to have humans choose one of two options at random.



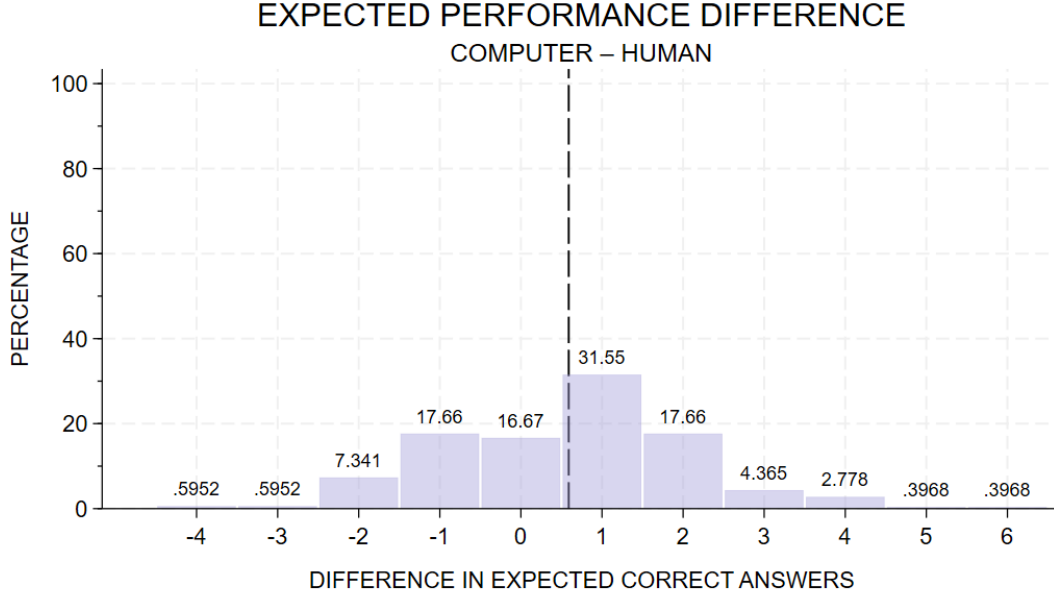
Note: Distribution of the number of expected correct answers in the ten rounds of the Dot Task. Dashed vertical lines indicate the mean numbers of expected correct answers, 5.45 and 6.04.

Figure 2.8: Performance Expectations

Turning our attention towards individual *differences* in performance expectations, Fig. 2.9 shows that a significantly higher proportion of subjects (57.14%) expected the computer to outperform them than vice-versa (26.19%) ($p < 0.01$, Two-Sided Two-Sample Test of Proportion). Furthermore, the proportion of subjects who chose to rely on the computer was significantly higher among those whose computer performance expectation exceeded their own (72.92%) than among those where the reverse was the case (15.15%) ($p < 0.01$, Two-Sided Two-Sample Test of Proportion). Similarly, the proportion of subjects who chose to rely on their own performance was significantly higher among those whose own performance expectation exceeded that of the computer (84.85%) than among those where the reverse was the case (27.08%)¹². In spite of the high levels of both objective and subjective underreliance that we observe, individual differences in performance expectations between humans and computers thus appear to have played an important role for automation reliance.

This is not unexpected, but it highlights a possible alternative explanation for the observed treatment effects: if performance expectations were affected by whether or not p was the good or the bad outlier and subjects, in turn, based their automation reliance decision on differences in said performance expectations, then our treatment effect might reflect a change in expectations due to the non-realised values within the set. There are two main reasons for why we need to account for this possibility. Firstly, while this channel could exist alongside our proposed channels of emotional reactions and kindness perceptions, it would suggest that the role played by emotional reactions and kindness perceptions is weaker than we might initially assume. Secondly, it could be the case that treatment differences in performance expectations reflect subjects — consciously or subconsciously — adjusting their expectations to match their automation reliance decisions, which would suggest that automation reliance decisions *caused* differences in performance expectations rather than the other way around.

¹²Note that this is the proportion of subjective underreliance that we discussed in Section 3.1.



Note: Distribution of individual differences in the number of expected correct answers; calculated by subtracting own performance expectations from computer performance expectations. Positive (negative) values indicate higher (lower) computer performance expectations, zero indicates equal performance expectations. Dashed vertical line indicates the mean difference in the number of expected correct answers of 0.59.

Figure 2.9: Performance Expectation Differences

However, comparing differences in performance expectations across treatment groups, we can rule out these differences as a primary cause of our treatment effect. With the exception of the BASELINE and GO-EC, where the degree to which subjects expected the computer to outperform them was marginally lower in the former than it was in the latter ($p < 0.10$, Mann-Whitney U Test), we are, in all other cases, unable to reject the null hypothesis of equal differences in performance expectations (see Fig. B.11 – B.15 in Appendix B.1 for treatment-specific illustrations of performance expectation differences).

3.3 Emotional Reactions and Kindness Perceptions

Next, we test whether emotional reactions and kindness perceptions were affected by where p ranked in the set of possible success probabilities or who was framed as the responsible party in selecting p , and whether either was a determinant of automation reliance. We therefore computed an Emotions Score (ES) and a Kindness Score (KS) that positively valued positive emotional reactions and positive kindness perceptions, and negatively valued negative emotional reactions and negative kindness perceptions.

Recall that we used two sets of five-point Likert scales consisting of eight items each, according to which subjects evaluated (1) the extent to which they felt ANGRY, ANNOYED, DISAPPOINTED, SAD, HAPPY, CONTENT, GRATEFUL, and CALM, and (2) the extent to which they considered the computer’s algorithm KIND, FAIR, FRIENDLY, GENEROUS, MEAN, UNFAIR, HOSTILE, and UNCHARITABLE. Our five-point Likert scales for each of these items were coded 1 – 5, with 1 indicating “not at all” and 5 indicating “extremely” (see Appendix B.2 for detailed instructions).

It follows that, if a subject indicated the maximum value for all positive emotional reactions or kindness perceptions and the minimum value for all negative emotional reactions or kindness perceptions, they would obtain a total score of $ES = 16$ or $KS = 16$. Conversely, if a subject indicated the maximum value for all negative emotional reactions or kindness perceptions and the minimum value for all positive emotional reactions or kindness perceptions, they would obtain a total score of $ES = -16$ or $KS = -16$. Consequently, we normalised ES and KS as follows:

$$(1) ES = \frac{(\text{happy} + \text{content} + \text{grateful} + \text{calm}) - (\text{angry} + \text{annoyed} + \text{disappointed} + \text{sad})}{16}$$

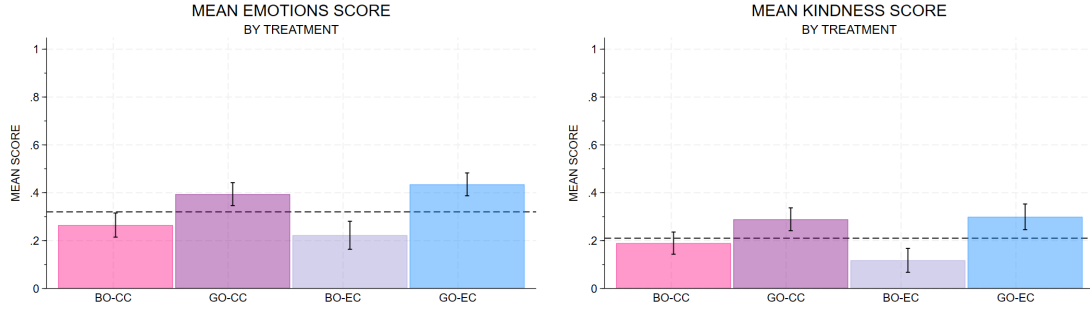
$$(2) KS = \frac{(\text{kind} + \text{fair} + \text{friendly} + \text{generous}) - (\text{mean} + \text{unfair} + \text{hostile} + \text{uncharitable})}{16}$$

Then, $ES, KS \in [-1, 1]$, where $ES, KS = -1$ indicate the maximum negative emotional reactions and kindness perceptions and $ES, KS = 1$ the maximum positive emotional reactions and kindness perceptions¹³.

Fig 2.10 illustrates mean ES and KS, with whiskers indicating the 95% confidence intervals and a dashed line marking ES and KS averages in the BASELINE as a benchmark. Note first of all that both ES and KS were significantly higher than zero in all treatment groups ($p < 0.01$, Two-Sided Wilcoxon Single Sample t-test). In other words, our subjects experienced stronger positive than negative emotional reactions and reported stronger positive than negative kindness perceptions. Furthermore, this difference was the most pronounced in GO-CC and GO-EC, which matches our earlier result that underreliance decreased when p was the good outlier¹⁴.

¹³Cf. Prah & Van Swol (2017), who, in a study on the aversion to algorithmic advice, employ a similar approach using six positive emotions (APPRECIATIVE, HAPPY, GRATEFUL, THANKFUL, SATISFIED, GLAD) and four negative emotions (MAD, FRUSTRATED, ANNOYED, IRRITATED) which were combined into indices of positive and negative emotions.

¹⁴For a breakdown of individual emotions and kindness measures, see Appendix B.1: Figures B.16 – B.31 show summary statistics, Tables B.3 and B.4 show median scores across treatment groups, and Tables B.5 – B.9 report intercorrelations (Spearman’s ρ) and internal consistency reliabilities (Cronbach’s α).



Note: ES and KS range from -1 to 1 , with $ES, KS = -1$ indicating the maximum negative emotional reactions and kindness perceptions and $ES, KS = 1$ the maximum positive emotional reactions and kindness perceptions. Whiskers indicate 95% confidence intervals; dashed lines indicate the BASELINE benchmarks, 0.32 and 0.21.

Figure 2.10: Average Emotions Score (ES) and Kindness Score (KS)

Tables 2.3 and 2.4 also paint a similar picture to what we have already observed for automation reliance decisions: both ES and KS were significantly higher when p was the good outlier than when p was the bad outlier. However, unlike before, we now also observe *some* treatment differences reflecting our manipulation of the ascription of agency in choosing p . Specifically, both ES and KS were significantly higher in the BASELINE than in BO-EC but not in BO-CC. Consequently, relative to the BASELINE, we see a significant drop in positive emotional reactions and kindness perceptions for the bad outlier only when it was presented as the experimenter’s choice. KS was also significantly higher in BO-CC than in BO-EC, so, if p was the bad outlier, positive kindness perceptions were significantly lower when it was presented as the experimenter’s rather than the computer’s choice; this was, however, not the case for ES and also did not apply in either case when p was the good outlier. Unlike **Hypothesis 4a**, **Hypothesis 4b** is therefore only partially supported.

Result 4a. *Emotional reactions and kindness perceptions were more positive when p was the good outlier.*

Result 4b. *Kindness perceptions were more pronounced for the experimenter’s than the computer’s choice of the bad outlier, but the same was not true of emotional reactions.*

So far, we have shown that whether p was the good or the bad outlier affected automation reliance as well as emotional reactions and kindness perceptions, but we have yet to establish a direct link between them. In line with **Hypothesis 5a**, ES is negatively correlated with objective underreliance (Spearman’s $\rho = -0.20$, $p < 0.01$) and was significantly lower among those who chose to rely on their own performance than among those who chose to rely on the computer’s ($p < 0.01$, Mann-Whitney U Test). However, this difference is once again driven by positive emotions: except for DISAPPOINTED, none of the negative emotions are significantly positively correlated with objective underreliance, whereas all positive emotions are significantly negatively correlated with objective underreliance (see Table B.10 in Appendix B.1).

	ES ₁	ES ₂	p (H ₀ : ES ₁ = ES ₂) <i>Mann-Whitney U Test</i>
BASELINE — BO-CC	0.3221	0.2649	0.1138
BASELINE — GO-CC	0.3221	0.3942	0.0839
BASELINE — BO-EC	0.3221	0.2222	0.0159
BASELINE — GO-EC	0.3221	0.4350	0.0021
BO-CC — GO-CC	0.2649	0.3942	0.0005
BO-EC — GO-EC	0.2222	0.4350	0.0000
BO-CC — BO-EC	0.2649	0.2222	0.3159
GO-CC — GO-EC	0.3942	0.4350	0.1520

Note: ES₁ and ES₂ denote the first and second mean emotions scores in each row for whose between-treatment equality we test.

Table 2.3: Mean Emotions Scores and Treatment Comparisons

	KS ₁	KS ₂	p (H ₀ : KS ₁ = KS ₂) <i>Mann-Whitney U Test</i>
BASELINE — BO-CC	0.2121	0.1894	0.7548
BASELINE — GO-CC	0.2121	0.2890	0.0320
BASELINE — BO-EC	0.2121	0.1174	0.0204
BASELINE — GO-EC	0.2121	0.2992	0.0072
BO-CC — GO-CC	0.1894	0.2890	0.0100
BO-EC — GO-EC	0.1174	0.2992	0.0000
BO-CC — BO-EC	0.1894	0.1174	0.0440
GO-CC — GO-EC	0.2890	0.2992	0.4735

Note: KS₁ and KS₂ denote the first and second mean kindness scores in each row for whose between-treatment equality we test.

Table 2.4: Mean Kindness Scores and Treatment Comparisons

In line with **Hypothesis 5b**, KS is negatively correlated with objective underreliance (Spearman's $\rho = -0.13$, $p < 0.01$) and was significantly lower among those who chose to rely on their own performance than among those who chose to rely on the computer's ($p = 0.0027$, Mann-Whitney U Test). Again, this difference is driven by positive kindness perceptions: negative kindness perceptions were not significantly positively correlated with objective underreliance, whereas all positive emotions were significantly negatively correlated with objective underreliance (see Table B.11 in Appendix B.1).

Result 5a. *Positive emotional reactions are significantly negatively correlated with objective underreliance, but negative emotional reactions are largely uncorrelated with objective underreliance.*

Result 5b. *Positive kindness perceptions are significantly negatively correlated with objective underreliance, but negative kindness perceptions are uncorrelated with objective underreliance.*

3.4 Determinants of Underreliance

Finally, we perform OLS regressions for a comprehensive analysis of the determinants of objective underreliance in our experiment. Our dependent variable is a dummy that takes the value of 1 for subjects who chose to rely on their own performance and of 0 for those who chose to rely on the computer’s performance in the Dot Task. Explanatory variables include two dummies indicating whether p was the good or the bad outlier and whether the selection of p was presented as the experimenter’s or the computer’s choice, as well as their interaction term. We furthermore included differences in performance expectations, ES, KS as well as gender, risk preferences, and ambiguity preferences.

Overall, the three OLS regressions reported in Table 2.5 support our results thus far. Subjects for whom p was the good outlier, *ceteris paribus*, were significantly less likely to engage in objective underreliance, whereas who was framed as the responsible party in selecting p had no significant effect — neither in isolation nor in conjunction with the relative rank of p within the set that it was selected from.

In addition to this, expected performance differences are a significant predictor of objective underreliance: while we have previously shown that there was a non-negligible number of cases of subjective underreliance (i.e. non-reliance on the computer in spite of higher performance expectations for the computer than oneself), higher performance expectations for the computer than oneself generally decreased the likelihood of engaging in objective underreliance. It is, however, worth noting that the effect of the good outlier is not only robust to the inclusion of expected performance differences but also larger in size. Positive ES are also significantly negatively associated with objective underreliance, while KS, like we have shown, do not appear to play a significant role.

All effects are robust to the inclusion of controls for gender, risk preferences, and ambiguity preferences. Women were significantly more likely to engage in objective underreliance than men. Subjects reported a mean score of 0.4724 in the *General Risk Question* (Dohmen et al., 2011), with 0 corresponding to the most risk-averse and 1 to the most risk-seeking behaviour; risk preferences do, however, not appear to have significantly influenced the likelihood of objective underreliance. We also calculated a mean matching probability of 0.4480 via the *Ambiguity Preference Survey Module* (Cavatorta & Schröder, 2019), classing 330 subjects (65.48%) as ambiguity-averse and 174 (34.52%) subjects as ambiguity-seeking (see Tables B.12 and B.13 in Appendix B.1)¹⁵. Ambiguity-seeking behaviour is significantly positively associated with objective underreliance, which, given the ambiguous nature of one’s own performance compared to the risky nature of the computer’s performance, is what we would expect to see.

¹⁵Note that, for ease of comparison with ambiguity scores, we normalised risk scores from a range of 0 – 10 to a range of 0 – 1.

	OBJECTIVE UNDERRELIANCE		
	(1)	(2)	(3)
GOOD OUTLIER	−0.27*** (0.07)	−0.22*** (0.06)	−0.23*** (0.06)
EXPERIMENTER CHOICE	−0.03 (0.07)	−0.03 (0.06)	−0.03 (0.06)
GOOD OUTLIER × EXPERIMENTER CHOICE	0.00 (0.10)	0.04 (0.09)	0.05 (0.09)
EXPECTED PERFORMANCE DIFFERENCE		−0.13*** (0.01)	−0.14*** (0.01)
EMOTIONS SCORE (ES)		−0.27*** (0.10)	−0.26** (0.10)
KINDNESS SCORE (KS)		0.03 (0.11)	0.03 (0.10)
FEMALE			0.09* (0.05)
RISK SCORE			0.00 (0.01)
AMBIGUITY SCORE			0.32* (0.17)
CONSTANT	0.60*** (0.05)	0.73*** (0.05)	0.52*** (0.10)
Observations	400	400	396
Prob > F	0.0000	0.0000	0.0000
R-squared	0.0714	0.2515	0.2629
Root MSE	0.4825	0.4348	0.4328

Note: BASELINE observations are not included. Dependent variable OBJECTIVE UNDERRELIANCE is a dummy that takes the value of 1 for subjects who chose to rely on their own performance rather than that of the computer. EXPECTED PERFORMANCE DIFFERENCE is calculated by subtracting own performance expectations from computer performance expectations. ES and KS range from −1 to 1, with higher values indicating more positive emotional reactions and kindness perceptions. RISK SCORE is a score from 0 to 1, where 0 corresponds to the most risk-averse and 1 to the most risk-seeking behaviour. AMBIGUITY SCORE is a probability where 0 corresponds to the most ambiguity-averse and 1 to the most ambiguity-seeking behaviour. Standard errors in parentheses, * p<0.1, ** p<0.05, *** p<0.01.

Table 2.5: OLS Regression: Underreliance on Automation

4 Discussion and Conclusion

Building on the seminal CASA paradigm and subsequent evidence that humans treat computers as social actors rather than machines, we set out to investigate non-realised values and the ascription of agency as determinants of underreliance on automation. We therefore conducted an online experiment in which subjects had to decide whether to rely on their own performance or that of a computer whose success probability was, on average, greater than their own. Our treatment groups varied (1) whether the computer success probability was chosen from a set and, if so, whether it constituted the ‘good’ or the ‘bad’ outlier within this set, and (2) whether agency over the choice of the computer success probability was ascribed to the computer itself or to the experimenter.

Our results suggest that automation reliance was sensitive to non-realised values: we find a significant decrease in objective as well as subjective underreliance on automation when the computer success probability was the good outlier — that is, when subjects were encouraged to believe that the computer performed to the best of its ability. Crucially, when the computer success probability was the bad outlier — that is, when subjects were encouraged to believe that the computer performed to the worst of its ability — we do not observe the opposite effect but rather a reversion to our BASELINE, in which the computer success probability was not chosen from a set. Setting aside the possibility that there was a small effect of the perception of p as a bad outlier on automation reliance decisions that we were underpowered to detect, we propose two possible explanations for this finding: one, in the absence of non-realised values to judge p by, subjects either formed their own priors compared with which p *was* a bad outlier, or two, the perception of p as a bad outlier did not affect automation reliance decisions.

We also link these effects to emotional reactions to the revelation of the computer success probability and, to a lesser degree, to perceptions of its kindness. When the computer success probability was the good outlier, subjects reported significantly more positive emotional reactions to and kindness perceptions of it. Furthermore, both positive emotional reactions and positive kindness perceptions, unlike negative ones, are significantly negatively correlated with objective underreliance; however, only the former survive the inclusion of controls for performance expectations, gender, risk attitudes, and ambiguity attitudes in a parametric analysis. We need to take into account that we collected our measures of emotional reactions and kindness perceptions *after* subjects made their automation reliance decisions; consequently, it is possible that the above links are not causal but, at least in part, the results of the ex-post rationalisation of automation reliance decisions. Nevertheless, our findings point towards emotional reactions as a potential channel through which framing an algorithm as superior to others which might have been available can increase automation reliance.

Conversely, we find no significant effect of whether agency over the selection of the computer success probability was ascribed to the computer itself or to the experimenter: subjects exhibited small differences in emotional reactions and kindness perceptions depending on who was framed as the responsible party, which did, however, not translate into differences in automation reliance decisions. Overall, our results are consistent with the unconscious application of social behaviours and expectations to humans and computers alike.

Nevertheless, future research into differences in the ascription of agency is necessary to analyse how the possible derivation of a computer's intentionality from that of its programmer factors into this. We initially hypothesised that, unlike explanations based on the CASA paradigm, the derivation of a computer's intentionality from that of its programmer may result in more pronounced emotional reactions and kindness perceptions when the role of the experimenter is made prominent. We find no evidence of this, but acknowledge that our treatment manipulations — which relied exclusively on verbal framing — might have been too subtle. Alternatively, it is also conceivable that subjects in our experiment were acutely aware of the experimenter's role as the computer's programmer even when it is not explicitly mentioned, and we therefore see no effect of our second treatment manipulation.

Finally, individual differences in performance expectations have been shown to play a significant, but comparatively small role for automation reliance decisions in our study; instead, the results confirm that the (un)willingness to accept an automated offer of assistance, just like has been shown to be the case in similar human-human interactions, goes beyond concerns of material consequences, and, in some cases, persisted even when an individual expected the computer to outperform them. Still, our findings indicate that it may be possible to shift priors about the capabilities of an automated agent via non-realised values in such a way that people are encouraged to believe that it is performing to the best of its ability — and are therefore more likely to rely on it, regardless of who is framed as responsible for setting its underlying success probability.

References

- Abeliuk, A., Benjamin, D. M., Morstatter, F., & Galstyan, A. (2020). Quantifying Machine Influence Over Human Forecasters. *Scientific Reports*, 10(1), 1–14.
- Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Exploring the Impact of Artificial Intelligence: Prediction versus Judgment. *Information Economics and Policy*, 47, 1–6.
- Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C., & Sugden, R. (2009). *Experimental Economics: Rethinking the Rules*. Princeton University Press.
- Bartling, B., Fehr, E., & Herz, H. (2014). The Intrinsic Value of Decision Rights. *Econometrica*, 82(6), 2005–2039.
- Bobadilla-Suarez, S., Sunstein, C. R., & Sharot, T. (2017). The Intrinsic Value of Choice: The Propensity to Under-delegate in the Face of Potential Gains and Losses. *Journal of Risk and Uncertainty*, 54(3), 187–202.
- Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The Social Dilemma of Autonomous Vehicles. *Science*, 352(6293), 1573–1576.
- Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A Systematic Review of Algorithm Aversion in Augmented Decision Making. *Journal of Behavioral Decision Making*, 33(2), 220–239.
- Cavatorta, E. & Schröder, D. (2019). Measuring Ambiguity Preferences: A New Ambiguity Preference Survey Module. *Journal of Risk and Uncertainty*, 58(1), 71–100.
- Chugunova, M. & Sele, D. (2022). We and It: An Interdisciplinary Review of the Experimental Evidence on How Humans Interact with Machines. *Journal of Behavioral and Experimental Economics*, 99, 101897.
- Cubitt, R., Gächter, S., & Quercia, S. (2017). Conditional Cooperation and Betrayal Aversion. *Journal of Economic Behavior & Organization*, 141, 110–121.
- Cubitt, R. P., Starmer, C., & Sugden, R. (1998). On the Validity of the Random Lottery Incentive System. *Experimental Economics*, 1, 115–131.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General*, 144(1), 114.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*, 64(3), 1155–1170.

- Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences. *Journal of the European Economic Association*, 9(3), 522–550.
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review*, 114(4), 864.
- Falk, A., Fehr, E., & Fischbacher, U. (2003). On the Nature of Fair Behavior. *Economic Inquiry*, 41(1), 20–26.
- Fehr, E., Herz, H., & Wilkening, T. (2013). The Lure of Authority: Motivation and Incentive Effects of Power. *American Economic Review*, 103(4), 1325–59.
- Fogg, B. & Nass, C. (1997). How Users Reciprocate to Computers: An Experiment That Demonstrates Behavior Change. In *CHI'97 Extended Abstracts on Human Factors in Computing Systems* (pp. 331–332).
- Freer, M., Friedman, D., & Weidenholzer, S. (2023). Motives for Delegating Financial Decisions. Available at: <https://doi.org/10.48550/arXiv.2309.03419>.
- Goldbach, C., Kayar, D., Pitz, T., & Sickmann, J. (2019). Transferring Decisions to an Algorithm: A Simple Route Choice Experiment. *Transportation Research Part F: Traffic Psychology and Behaviour*, 65, 402–417.
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science*, 315(5812), 619–619.
- Hollard, G., Massoni, S., & Vergnaud, J.-C. (2016). In Search of Good Probability Assessors: An Experimental Comparison of Elicitation Rules for Confidence Judgments. *Theory and Decision*, 80(3), 363–387.
- Jiang, T. (2013). Cheating in mind games: The subtlety of rules matters. *Journal of Economic Behavior & Organization*, 93, 328–336.
- Johnson, D. G. (2006). Computer Systems: Moral Entities But Not Moral Agents. *Ethics and Information Technology*, 8(4), 195–204.
- Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why Are We Averse towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion. *Research Papers (168)*, available at: https://aisel.aisnet.org/ecis2020_rp/168.
- Kim, J., Sah, Y. J., & Song, H. (2021). Agreeableness of a Virtual Agent: Effects of Reciprocity and Need for Help. In *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)* (pp. 1–6).: IEEE.
- Kim, Y. & Sundar, S. S. (2012). Anthropomorphism of Computers: Is It Mindful or Mindless? *Computers in Human Behavior*, 28(1), 241–250.

- Kirchkamp, O. & Strobel, C. (2019). Sharing Responsibility with a Machine. *Journal of Behavioral and Experimental Economics*, 80, 25–33.
- Koehler, J. J. & Gershoff, A. D. (2003). Betrayal Aversion: When Agents of Protection Become Agents of Harm. *Organizational Behavior and Human Decision Processes*, 90(2), 244–261.
- Krämer, N. C., von der Pütten, A., & Eimler, S. (2012). Human-Agent and Human-Robot Interaction Theory: Similarities to and Differences from Human-Human Interaction. *Human-Computer Interaction: The Agency Perspective*, (pp. 215–240).
- Lammer, L., Huber, A., Weiss, A., & Vincze, M. (2014). Mutual Care: How Older Adults React When They Should Help Their Care Robot. In *AISB2014: Proceedings of the 3rd International Symposium on New Frontiers in Human-Robot Interaction* (pp. 1–4).
- Lee, E. J., Nass, C., & Brave, S. (2000). Can Computer-Generated Speech Have Gender? An Experimental Test of Gender Stereotype. In *CHI'00 Extended Abstracts on Human Factors in Computing Systems* (pp. 289–290).
- Lee, S. A. & Liang, Y. (2015). Reciprocity in Computer-Human Interaction: Source-Based, Norm-Based, and Affect-Based Explanations. *Cyberpsychology, Behavior, and Social Networking*, 18(4), 234–240.
- Liaudinskas, K. (2022). Human vs. Machine: Disposition Effect Among Algorithmic and Human Day Traders. In *Proceedings of the EUROFIDAI-ESSEC Paris December Finance Meeting 2022*, available at SSRN: <https://ssrn.com/abstract=4260869>.
- Lobeck, M. (2023). Motivating Beliefs in a Just World. Available at SSRN: <https://ssrn.com/abstract=4369242> or <http://dx.doi.org/10.2139/ssrn.4369242>.
- Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, 46(4), 629–650.
- March, C. (2021). Strategic Interactions Between Humans and Artificial Intelligence: Lessons from Experiments with Computer Players. *Journal of Economic Psychology*, 87, 102426.
- McCabe, K. A., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization*, 52(2), 267–275.
- Miyoshi, T. & Matsubara, S. (2018). Dynamically Forming a Group of Human Forecasters and Machine Forecasters for Forecasting Economic Indicators. In *IJCAI* (pp. 461–467).

- Mol, J. M., van der Heijden, E. C., & Potters, J. J. (2020). (Not) Alone in the World: Cheating in the Presence of a Virtual Observer. *Experimental Economics*, 23, 961–978.
- Moon, Y. (2000). Intimate Exchanges: Using Computers to Elicit Self-Disclosure from Consumers. *Journal of Consumer Research*, 26(4), 323–339.
- Nanay, B. (2006). Symmetry Between the Intentionality of Minds and Machines? The Biological Plausibility of Dennett’s Account. *Minds and Machines*, 16(1), 57–71.
- Nass, C. & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1), 81–103.
- Nass, C., Moon, Y., & Green, N. (1997). Are Machines Gender-Neutral? Gender-Stereotypic Responses to Computers with Voices. *Journal of Applied Social Psychology*, 27(10), 864–876.
- Owens, D., Grossman, Z., & Fackler, R. (2014). The Control Premium: A Preference for Payoff Autonomy. *American Economic Journal: Microeconomics*, 6(4), 138–61.
- Palan, S. & Schitter, C. (2018). Prolific.ac — A subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance*, 17, 22–27.
- Prahl, A. & Van Swol, L. (2017). Understanding Algorithm Aversion: When Is Advice from Automation Discounted? *Journal of Forecasting*, 36(6), 691–702.
- Reeves, B. & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People*. Center for the Study of Language and Information, Cambridge University Press.
- Sandoval, E. B., Brandstetter, J., Obaid, M., & Bartneck, C. (2016). Reciprocity in Human-Robot Interaction: A Quantitative Approach Through the Prisoner’s Dilemma and the Ultimatum Game. *International Journal of Social Robotics*, 8(2), 303–317.
- Schniter, E., Shields, T. W., & Sznycer, D. (2020). Trust in Humans and Robots: Economically Similar But Emotionally Different. *Journal of Economic Psychology*, 78, 102253.
- Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417–424.
- Searle, J. R., Willis, S., et al. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology*, 54(6), 1063.

CHAPTER 3

Preferences for Normative over Preferential Algorithms With(out) Feedback: Experimental Evidence from a Repeated Lottery Selection Task[†]

1 Introduction

“Errare humanum est” — to err is human, and human error is, by and large, forgiveable. Evidence suggests that this adage does not, however, extend to algorithms. Suboptimal automation reliance is an increasingly well-documented phenomenon in the experimental and behavioural economic literature. So far, however, research is inconclusive as to its underlying causes (see e.g. Bonnefon et al., 2016; Abeliuk et al., 2020; Liehner et al., 2021). A promising yet underexplored avenue is the sensitivity to algorithmic errors: it appears that human decision-makers tend to overweight algorithmic errors relative to comparable human errors when deciding whether or not to rely on an automated agent (Shariff et al., 2017).

Studies have found that the root of the problem lies not in misperceptions of algorithmic error rates (see e.g. Dzindolet et al., 2002; Gogoll & Uhl, 2018) but in the disproportionate, adverse reaction of human decision-makers to observing such errors. Initially, people often hold high expectations of novel forms of automation without corroborating evidence (Dzindolet et al., 2003). Observable errors, however, appear to destroy the illusion of infallibility and trigger a rapid dissolution of trust in automation (Hoff & Bashir, 2015), which may be prevented by more realistic expectations in the case of erring humans (Madhavan & Wiegmann, 2007). Highhouse (2008) further suggests that, in the context of employee selection, the resistance to analytical decision aids — including but not limited to algorithms — is exacerbated by the mistaken belief that human intuition categorically corrects imperfections of analytical decision aids but not vice-versa.

[†]This chapter is based on joint work with Robin Cubitt, Chris Starmer, and Irenaeus Wolff. I am grateful to CeDEx for funding this research. I also thank CeDEx members for comments and feedback. The study received ethical approval from the Nottingham School of Economics Research Ethics Committee on 02/05/2023.

Dietvorst et al. (2015) show that subjects in their study do, indeed, consistently forego algorithmic forecasters in favour of inferior human forecasters as a consequence of observing algorithmic errors, and that this effect persists even when subjects witness the overall superior performance of the algorithm. In support of Highhouse (2008), automation reliance also increases when human forecasters are given the opportunity to modify imperfect algorithmic ones (Dietvorst et al., 2018). Prahla & Van Swol (2017) do not find evidence of algorithm aversion, but report that, after receiving bad algorithmic advice, the utilisation of algorithmic advice decreases significantly more than the utilisation of human advice does after receiving comparably bad human advice.

Until now, research into suboptimal automation reliance has, however, predominantly focused on environments in which algorithms are *designed* to correct human errors — that is, environments in which “right” and “wrong” decision(s) are objective. In contrast, the preferential choice domain has yet to be explored in spite of the fact that people have long since been exposed to automated assistance in such environments: think, for instance, of algorithmic advice on complex financial products like insurance or investment products, where there is rarely, if ever, a single “right” decision but optimal actions depend on individual characteristics such as risk preferences. Consequently, in the preferential choice domain, algorithmic as well as human advisors and delegates are faced with a choice between two different approaches: a *normative* paradigm based on a set of rules that prescribe optimal actions on a general level, or a *preferential* paradigm that attempts to predict preferred actions on an individual level.

In light of the human sensitivity to algorithmic errors, the effect of algorithm type — normative versus preferential — on automation reliance is furthermore likely to interact with feedback on the outcome of algorithmic decisions. Overall, the preferential choice domain renders errors harder to define than environments in which “right” and “wrong” decision(s) are objective. However, preferential algorithms may be particularly susceptible to the identification of and, as a consequence, adverse reactions to algorithmic errors if humans believe themselves to possess an inherent understanding of their own preferences: simply put, it may be easier to assess whether a preferential algorithm is successful in predicting one’s own choices than whether a normative algorithm produces results that are superior to the results that one’s own choices would have produced.

We believe that investigating the perception and reception of normative versus preferential algorithms, with and without feedback, could generate insights not only into the unique challenges of the preferential choice domain, but also for the wider goal of understanding the general determinants and deterrents of automation reliance in complex decision environments.

We therefore conducted an experiment that investigates automation reliance in a repeated lottery selection task. Subjects could choose, in each round of the task, whether to select a lottery themselves or have an algorithm select a lottery for them. We designed the lotteries in such a way that, under the assumption of well-defined risk preferences, the best choice for each subject depended on their risk preferences but was difficult to identify due to the complexity of the menu of available lotteries in each round of the task. Treatment groups varied (1) whether subjects were offered a normative algorithm based on Expected Value Theory or a preferential algorithm based on Rank Dependent Utility Theory, and (2) whether or not they received outcome feedback regarding the selected lottery in each round. We programmed the preferential algorithm based on the results of a prior experiment with the same subjects, in which we followed the procedures in Murad et al. (2016) — which are based on Fehr-Duda et al. (2006) and Bruhin et al. (2010) — to measure individual risk preferences.

Our results show significantly lower levels of reliance on the preferential algorithm than on its normative counterpart. We furthermore find that satisfactory feedback reinforced a subject’s previous decision to select the lottery themselves or have an algorithm select it for them, whereas unsatisfactory feedback led to changes from their previous decision; an effect which was significantly stronger for the preferential than for the normative algorithm. Crucially, we observe significantly lower levels of reliance on the preferential algorithm in every round *including* the first round, where subjects had yet to receive feedback on its selection behaviour or the resulting outcomes. In other words, the observed aversion to the preferential algorithm in our setting cannot be attributed to negative experience alone; instead, we hypothesise that unfavourable ex-ante beliefs about whether or not the preferential algorithm would be able to deliver on its promises lowered initial willingness to rely on it, which was then exacerbated when unsatisfactory feedback confirmed these beliefs.

The paper proceeds as follows: in Section 2, we outline the design of the experiment and our research hypotheses; Section 3 presents the results of our data analysis, while Section 4 discusses these results and concludes.

2 Experimental Design

Our main experiment investigates automation reliance in a repeated lottery selection task that will be explained in detail in Section 2.1. In each round of the task, a lottery had to be selected from a menu of lotteries. Lottery menus were constructed in such a way that, given the time limit for which the menu was visible, comparing lotteries to identify the one that best aligned with one’s preferences was not trivial. Consequently, subjects were offered a choice between selecting the lottery themselves or having a computer, which acted according to a specific algorithm, select it for them.

Our treatment groups manipulated: (1) whether subjects were offered a computer that used what we label a ‘normative’ algorithm based on Expected Value Theory (EV) or a computer that used what we label a ‘preferential’ algorithm based on Rank Dependent Utility Theory (RDU), and (2) whether or not subjects received feedback on the outcome of the selected lottery after each round. In short, the EV algorithm would, in each round of the task, compare the expected values of all lotteries within the menu and select the one for which this was highest; the RDU algorithm would, in each round of the task and for each subject, compare estimates of their individual, rank dependent utilities for all lotteries within the menu and select the one for which this was highest.

We acknowledge that labelling the two models as ‘normative’ and ‘preferential’ is reductive as it oversimplifies their respective characteristics and differences between them. It is therefore important to note that we make no claims about whether or not Expected Value Theory *should* be considered a normative model of individual risk preferences, nor about how its performance compares to that of Rank Dependent Utility Theory. Instead, we use the labels ‘normative’ and ‘preferential’ to differentiate between algorithms that define optimality in a prescriptive sense and algorithms that define optimality in an individual, predictive sense. Expected Value Theory uses a mathematical rule that assumes collective risk neutrality and thus prescribes the same optimal lottery choice for every subject. Rank Dependent Utility Theory, on the other hand, allows for non-linearity and differences in individual risk preferences that can lead to between-subject variations in their preferred lottery choice. Finally, both models are widely established in behavioural and experimental research as well as comparatively parsimonious, which made them a good fit for the purpose of this study.

Unlike programming the EV algorithm, programming the RDU algorithm relied on the estimation of individual risk preferences prior to the main experiment. Consequently, we conducted our study in two parts: first, we collected data that allowed us to estimate individual RDU parameters and programmed the RDU algorithm accordingly; then, we reinvited the same sample of subjects to complete the main experiment three days later.

2.1 Main Experiment

Subjects completed 8 rounds of the lottery selection task. In each round of the task, one lottery needed to be selected from a menu of 8 lotteries; the outcome of the selected lottery would then be determined by the roll of a fair, 10-sided die. Fig 3.1 shows a sample lottery menu¹ with experimental currency units, hereafter ‘Points’, that were converted into monetary payoffs at the following conversion rate: 50 Points = £1.00. We employed a random incentive mechanism to determine which of the 8 rounds was relevant for each subject’s payoff (for evidence on the validity of this method, see Chapter 6 in Bardsley et al., 2009; Cubitt et al., 1998; Charness et al., 2016).

In each round, *prior* to being shown the lottery menu, subjects needed to decide on one of two selection mechanisms: they could either select the lottery themselves (‘self-reliance’) or have an algorithm select it for them (‘automation reliance’). In either case, the menu would be visible for a duration of 10 seconds. Once it disappeared, if a subject had decided to select the lottery themselves, they had 10 seconds to do so; otherwise they needed to wait for 10 seconds for the algorithm to select a lottery so as to avoid time concerns as a motivation for automation reliance. Subjects also learned that they would be making this decision at the start of every round, and that their decision in one round would not affect the proceedings in any other round.

	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>
	1	2	3	4	5	6	7	8	9	10
LOTTERY A	0	0	175	0	170	0	165	0	0	0
LOTTERY B	105	110	0	110	0	100	110	0	105	100
LOTTERY C	0	0	170	0	170	0	170	0	0	0
LOTTERY D	105	105	0	105	0	105	105	0	105	105
LOTTERY E	0	0	175	0	175	0	175	0	0	0
LOTTERY F	110	110	0	110	0	110	110	0	110	110
LOTTERY G	110	110	0	0	0	110	110	0	110	110
LOTTERY H	0	0	175	0	175	0	0	0	0	0

Note: 50 Points = £1.00.

Figure 3.1: Sample Lottery Menu

¹Details on the characteristics and construction of the lottery menus will be provided in Section 2.1.2.

After completing all 8 rounds of the lottery selection task, subjects were asked which of the two selection mechanisms they would choose if they had to commit to one of them for a further 8 rounds, and why. It was made clear to the subjects that this decision was purely hypothetical and would have no bearing on their payoff. We included it as an additional measure to determine whether, after completing the lottery selection task, subjects emerged with a preference for self-reliance or for automation reliance.

2.1.1 Treatment Groups

Subjects were randomly assigned to one of four treatment groups, which varied across two dimensions as shown in Fig. 3.2: (1) whether subjects were offered the normative algorithm (EV) or the preferential algorithm (RDU), and (2) whether subjects received outcome feedback regarding the selected lottery (F) or not (NF) after each round.

Specifically, subjects in EV-NF and EV-F received the following information: if, for a given task, they decided to have the computer select a lottery for them, its choice would be determined by an algorithm *that uses mathematics to work out how much each lottery would yield on average, and selects the lottery for which this is highest*. Subjects in RDU-NF and RDU-F, on the other hand, were told that: if, for a given task, they decided to have the computer select a lottery for them, its choice would be determined by an algorithm *that uses modern decision science and their choices in Part 1 of the study to model their behaviour, and selects the lottery that it predicts they would select*².

Fig 3.1 illustrates that the lottery selection task was designed to mimic the complexity of many decision environments in the real world, in which decision-makers are confronted with multiple, uncertain options. In principle, the optimal choice in such an environment depends on risk preferences, but subjects may nevertheless struggle to identify it (see literature on decision confidence and imprecise preferences, e.g. Cubitt et al., 2015; Loomes & Pogrebna, 2014) — especially under time pressure. Consequently, we expected that both the EV algorithm and the RDU algorithm would, for different reasons, appeal as alternatives to self-reliance in our set-up: the EV algorithm was designed to infallibly select the lottery that, on average, would yield the highest payoff but ignore the possibility of individual, non-linear risk preferences; the RDU algorithm, on the other hand, was designed to tailor its selection in each round to estimations of individual risk preferences. Note that, as will be explained in Section 2.2, the estimations underlying the programming of the RDU algorithm were based on choices in a series of comparatively simpler tasks, in which subjects should have been able to demonstrate their risk preferences without confounding factors.

²Further details on the programming the RDU algorithm and how it compared to the EV algorithm will be provided in Section 2.2. Within the experiment, we referred to the RDU parameter elicitation as Part 1 and the main experiment featuring the lottery selection task as Part 2 of the study.

	NO FEEDBACK	FEEDBACK
EV ALGORITHM	EV-NF	EV-F
RDU ALGORITHM	RDU-NF	RDU-F

Note: EV denotes the EV algorithm, RDU the RDU algorithm; NF indicates that the subjects did not receive feedback on the outcome of the selected lottery, whereas F indicates that subjects did receive feedback on the outcome of the selected lottery.

Figure 3.2: Treatment Groups

Crucially, our feedback-centric treatment manipulation focused on *outcome feedback* as opposed to *selection feedback*. Subjects in EV-NF and RDU-NF were informed that, after completing a task, regardless of whether they decided to select the lottery themselves or have the computer select it for them, they would be shown the selected lottery. Subjects in EV-F and RDU-F, on the other hand, would be shown the outcome of the die roll and the Points they would earn if that task were chosen for payoff in addition to the selected lottery. In both cases, the feedback screen only included the selected lottery and never the entire lottery menu.

Our reasons for focusing on outcome feedback are threefold. Firstly, we believe that feedback in the real world typically features the outcome of an automation reliance decision and we wanted to replicate this in our experiment: consider, for instance, returns to investment in the context of algorithmic trading or the discrepancy between expected and realised arrival times when using a sophisticated navigation system.

Secondly, feedback on the processes that facilitated a specific outcome — say, which portfolio of stocks money was invested in or which traffic data was used to plan a route — even if available, is likely less important than the outcome itself for ex-post evaluations of automation reliance. In fact, there is ample empirical evidence of the human tendency for outcome bias following the seminal work of Baron & Hershey (1988). Recently, Gillenkirch & Velthuis (2023) have shown, across three experiments, that principals hold agents accountable not for their actions but for outcomes of their actions that they have no control over (see also Brownback & Kuhn, 2019; Gurdal et al.,

2013). Similarly, König-Kersting et al. (2021) find evidence of outcome bias to the point where principals, in the context of risky financial decisions, report higher satisfaction with their agent after random successes that resulted from going against the principal’s recommendations than after a failed investment that was in line with the principal’s explicit requests.

Thirdly, from a practical perspective, self-reliance in our experiment implicitly ‘revealed’ the selected lottery because subjects would make the selection themselves. Consequently, we decided to hold *selection feedback* constant across treatment groups as well as automation reliance decisions, which then allowed us to isolate and focus on the effect of *outcome feedback*.

Lastly, after each round of the lottery selection task, subjects were asked to rate their satisfaction with the lottery selection on a scale of 1 to 5 stars, where 1 star meant “not satisfied at all” and 5 stars meant “completely satisfied”. Note that we inquired about the satisfaction with lottery *selection* rather than lottery *outcome* to avoid introducing an additional difference between treatment groups who received outcome feedback and those who did not. However, König-Kersting et al. (2021) demonstrate that outcome bias persists even in cases in which the separation of outcome evaluation and decision process evaluation is clearly emphasised; given that, in our experiment, this separation was not explicit, we expected a strong correlation between lottery outcome and reported satisfaction with lottery selection. In other words, we hypothesise that, in EV-NF and RDU-NF, in the absence of outcome feedback, we did measure satisfaction with the lottery *selection*, whereas in EV-F and RDU-F, where feedback on the lottery selection and its outcome were presented together, this measurement likely captured satisfaction with lottery *outcome* instead³.

2.1.2 Lottery Menu Construction

Unbeknownst to the subjects, all lottery menus were constructed using one of the two patterns illustrated in Fig. 3.3 and Fig. 3.4: every menu contained exactly one ‘P-Bet’, one ‘\$-Bet’, and four strictly dominated variations of either. P-Bets are characterised by a comparatively low monetary outcome that occurs with a comparatively high probability, whereas \$-Bets are characterised by a comparatively high monetary outcome that occurs with a comparatively low probability; both were popularised in the context of preference reversal phenomena (see e.g. Lichtenstein & Slovic, 1971; Grether & Plott, 1979; Loomes et al., 1989; Tversky et al., 1990; Cubitt et al., 2004).

³In addition to this, minimum and maximum values of a given lottery can presumably be interpreted as negative and positive outcomes, respectively. However, some lotteries could lead to a variety of outcomes *between* their minimum and maximum values, whose interpretation is less straight-forward – in these cases, satisfaction ratings also serve as an easily interpretable proxy for lottery outcome.

	DIE ROLL = 1	DIE ROLL = 2	DIE ROLL = 3	DIE ROLL = 4	DIE ROLL = 5	DIE ROLL = 6	DIE ROLL = 7	DIE ROLL = 8	DIE ROLL = 9	DIE ROLL = 10
LOTTERY A	X	X	X	0	0	0	0	0	0	0
LOTTERY B	X	X-5	X-10	0	0	0	0	0	0	0
LOTTERY C	X-5	X-5	X-5	0	0	0	0	0	0	0
LOTTERY D	X	X	0	0	0	0	0	0	0	0
LOTTERY E	Y	Y	Y	Y	Y	Y	Y	0	0	0
LOTTERY F	Y	Y-5	Y-10	Y	Y-5	Y-10	Y	0	0	0
LOTTERY G	Y-5	Y-5	Y-5	Y-5	Y-5	Y-5	Y-5	0	0	0
LOTTERY H	Y	Y	Y	Y	Y	Y	0	0	0	0

Note: X denotes the non-zero outcome of the \$-Bet (here: Lottery A) that occurs with a probability of 0.3, Y denotes the non-zero outcome of the P-Bet (here: Lottery E) that occurs with a probability of 0.7. Lotteries B, C, and D are the dominated variations of the \$-Bet, Lotteries F, G, and H are the dominated variations of the P-Bet.

Figure 3.3: Lottery Menu Construction (Pattern 1)

	DIE ROLL = 1	DIE ROLL = 2	DIE ROLL = 3	DIE ROLL = 4	DIE ROLL = 5	DIE ROLL = 6	DIE ROLL = 7	DIE ROLL = 8	DIE ROLL = 9	DIE ROLL = 10
LOTTERY A	X	X	0	0	0	0	0	0	0	0
LOTTERY B	X	X-10	0	0	0	0	0	0	0	0
LOTTERY C	X-5	X-5	0	0	0	0	0	0	0	0
LOTTERY D	X	0	0	0	0	0	0	0	0	0
LOTTERY E	Y	Y	Y	Y	Y	Y	Y	Y	0	0
LOTTERY F	Y	Y-10	Y	Y-10	Y	Y-10	Y	Y-10	0	0
LOTTERY G	Y-5	Y-5	Y-5	Y-5	Y-5	Y-5	Y-5	Y-5	0	0
LOTTERY H	Y	Y	Y	Y	Y	Y	Y	0	0	0

Note: X denotes the non-zero outcome of the \$-Bet (here: Lottery A) that occurs with a probability of 0.2, Y denotes the non-zero outcome of the P-Bet (here: Lottery E) that occurs with a probability of 0.8. Lotteries B, C, and D are the dominated variations of the \$-Bet, Lotteries F, G, and H are the dominated variations of the P-Bet.

Figure 3.4: Lottery Menu Construction (Pattern 2)

X denotes the non-zero outcome of the \$-Bet, which occurred with a probability of 0.3 or 0.2; Y denotes the non-zero outcome of the P-Bet, which occurred with a probability of 0.7 or 0.8. Presented as above, it is easy to see that the remaining six lotteries are strictly dominated: Lotteries B and C take small amounts off some of the non-zero outcomes of the \$-Bet and Lottery D reduces one of its outcomes to zero; Lotteries F, G, and H do the same for the P-Bet. For an individual whose preferences satisfy monotonicity, the choice would therefore boil down to the two dominating lotteries and depend on their risk preferences. However, the preference reversal literature has established that preferences over these two types of lotteries are not robust across elicitation procedures, which is an indication that some individuals struggle to articulate a clear preference for one over the other.

In addition to this, we did not present the lottery menus as shown in Fig. 3.3 and Fig. 3.4 but generated ‘scrambled’ versions of them: we first of all shuffled the order of rows, then horizontally shifted the P-Bet so as to reduce the number of columns containing only zeros, and, finally, shuffled the order of columns. Fig. 3.5 illustrates this using the lottery menu from Fig. 3.1 as an example.

Consequently, given the time pressure that the subjects in our study were under, what first appears to be a simple task turned out to be fairly complex: firstly, we made it difficult for subjects to realise that, under the assumption of monotonicity of preferences, each round of the lottery selection task was, effectively, a choice between one P-Bet and one \$-Bet; secondly, even if we had presented these two lotteries as a pairwise choice, past research has established that, for many individuals, articulating their preferences over P-Bets and \$-Bets is not trivial in and of itself.

Furthermore, the juxtaposition of P-Bets and \$-Bets ensured that differences in risk preferences were likely to lead to differences in lottery selections that only one of our two algorithms was programmed to account for: our EV algorithm selected lotteries based on expected values without heed for the specific composition of probabilities and outcomes determining these values; in contrast, our RDU algorithm selected the lotteries it predicted a subject would select based on estimations of their individual utility and probability weighting functions, which would take these factors into account.

Lastly, the inclusion of dominated lotteries meant that it was possible for humans to make errors in the form of monotonicity violations that both of our algorithms were programmed not to make. In the context of suboptimal automation reliance, both of our algorithms therefore had an advantage over humans from a performance-centric perspective and made for a good benchmark to compare human lottery selections to.

	<i>DIE ROLL</i> 1	<i>DIE ROLL</i> 2	<i>DIE ROLL</i> 3	<i>DIE ROLL</i> 4	<i>DIE ROLL</i> 5	<i>DIE ROLL</i> 6	<i>DIE ROLL</i> 7	<i>DIE ROLL</i> 8	<i>DIE ROLL</i> 9	<i>DIE ROLL</i> 10
<i>LOTTERY A</i>	175	175	175	0	0	0	0	0	0	0
<i>LOTTERY B</i>	175	170	165	0	0	0	0	0	0	0
<i>LOTTERY C</i>	170	170	170	0	0	0	0	0	0	0
<i>LOTTERY D</i>	175	175	0	0	0	0	0	0	0	0
<i>LOTTERY E</i>	110	110	110	110	110	110	110	0	0	0
<i>LOTTERY F</i>	110	105	100	110	105	100	110	0	0	0
<i>LOTTERY G</i>	105	105	105	105	105	105	105	0	0	0
<i>LOTTERY H</i>	110	110	110	110	110	110	0	0	0	0

⇓ SHUFFLE ROWS ⇓

	<i>DIE ROLL</i> 1	<i>DIE ROLL</i> 2	<i>DIE ROLL</i> 3	<i>DIE ROLL</i> 4	<i>DIE ROLL</i> 5	<i>DIE ROLL</i> 6	<i>DIE ROLL</i> 7	<i>DIE ROLL</i> 8	<i>DIE ROLL</i> 9	<i>DIE ROLL</i> 10
<i>LOTTERY A</i>	175	170	165	0	0	0	0	0	0	0
<i>LOTTERY B</i>	110	105	100	110	105	100	110	0	0	0
<i>LOTTERY C</i>	170	170	170	0	0	0	0	0	0	0
<i>LOTTERY D</i>	105	105	105	105	105	105	105	0	0	0
<i>LOTTERY E</i>	175	175	175	0	0	0	0	0	0	0
<i>LOTTERY F</i>	110	110	110	110	110	110	110	0	0	0
<i>LOTTERY G</i>	110	110	110	110	110	110	0	0	0	0
<i>LOTTERY H</i>	175	175	0	0	0	0	0	0	0	0

⇓ SHIFT P-BET AND ITS DOMINATED VARIATIONS ⇓

	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>
	1	2	3	4	5	6	7	8	9	10
LOTTERY A	175	170	165	0	0	0	0	0	0	0
LOTTERY B	0	0	110	105	100	110	105	100	110	0
LOTTERY C	170	170	170	0	0	0	0	0	0	0
LOTTERY D	0	0	105	105	105	105	105	105	105	0
LOTTERY E	175	175	175	0	0	0	0	0	0	0
LOTTERY F	0	0	110	110	110	110	110	110	110	0
LOTTERY G	0	0	110	110	110	110	110	110	0	0
LOTTERY H	175	175	0	0	0	0	0	0	0	0

⇓ SHUFFLE COLUMNS ⇓

	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>	<i>DIE ROLL</i>
	1	2	3	4	5	6	7	8	9	10
LOTTERY A	0	0	175	0	170	0	165	0	0	0
LOTTERY B	105	110	0	110	0	100	110	0	105	100
LOTTERY C	0	0	170	0	170	0	170	0	0	0
LOTTERY D	105	105	0	105	0	105	105	0	105	105
LOTTERY E	0	0	175	0	175	0	175	0	0	0
LOTTERY F	110	110	0	110	0	110	110	0	110	110
LOTTERY G	110	110	0	0	0	110	110	0	110	110
LOTTERY H	0	0	175	0	175	0	0	0	0	0

Note: Cf. Fig. 3.1: in the final, scrambled menu, the \$-Bet is Lottery E and the P-Bet is Lottery F; Lotteries A, C, and H are the dominated variations of the \$-Bet, Lotteries B, D, and G are the dominated variations of the P-Bet. Note that this corresponds to Menu G in Fig. 3.6.

Figure 3.5: Lottery Menu Scrambling Procedure

Fig. 3.6 illustrates the EVs of the P-Bets and \$-Bets in each of the 8 lottery menus. Note that the table summarises only the unperturbed P-Bets and \$-Bets in each menu but not any of their dominated variations, and that the order in which these menus were presented to the subjects was individually randomised to avoid order effects. In MENUS A–D, the \$-Bet has a higher EV than the corresponding P-Bet; in MENUS A and B the EV of the \$-Bet is approximately 30% larger, whereas in MENUS C and D this difference is only around 5%. In MENUS E–H, the P-Bet has a higher EV than the corresponding \$-Bet; in MENUS E and F the EV of the P-Bet is approximately 30% larger, whereas in MENUS G and H the EV of the P-Bet is approximately 50% larger. It follows that the EV algorithm would select the P-Bet half of the time and the \$-Bet otherwise, irrespective of individual risk preferences or the magnitude of the difference in EVs, which the RDU algorithm would take into account.

In summary, we created a decision environment that mimics real-world situations in which humans might elect to transfer their decision rights to an automated agent: given infinite time to evaluate the available information, subjects might have been able to avoid dominated lotteries and identify which of the remaining two options better aligned with their preferences; however, as time was scarce, as it typically is in our daily lives, there was scope for automation reliance to improve decision outcomes.

	X (\$-BET)	Y (P-BET)	$P(X)$	$P(Y)$	EV (\$-BET) vs EV (P-BET)
MENU A	170	55	0.3	0.7	51.00 >> 38.50
MENU B	210	40	0.2	0.8	42.00 >> 32.00
MENU C	220	90	0.3	0.7	66.00 > 63.00
MENU D	190	45	0.2	0.8	38.00 > 36.00
MENU E	205	115	0.3	0.7	61.50 << 80.50
MENU F	195	65	0.2	0.8	39.00 << 52.00
MENU G	175	110	0.3	0.7	52.50 <<< 77.00
MENU H	225	85	0.2	0.8	45.00 <<< 68.00

Note: X denotes the non-zero outcome of the \$-Bet, Y the non-zero outcome of the P-Bet; $P(X)$ indicates the probability with which X occurred, $P(Y)$ the probability with which Y occurred. In the final column, > indicates an EV difference $\approx 5\%$ of the smaller EV, >> and << indicate an EV difference $\approx 30\%$ of the smaller EV, and <<< indicates an EV difference $\approx 50\%$ of the smaller EV.

Figure 3.6: EV Comparison of P-Bets and \$-Bets in Lottery Menus

2.2 RDU Parameter Estimation

Let us now look at how we programmed the RDU algorithm. Three days before the main experiment, all subjects participated in a prior experiment in which we adopted the methodology established in Fehr-Duda et al. (2006) that has been shown to capture individual utility and probability weighting function parameters using a choice elicitation task (see e.g. Bruhin et al., 2010; Fehr-Duda et al., 2010; Epper et al., 2011). We used the resulting RDU parameter estimates only for subjects who were randomly assigned to RDU-NF or RDU-F in the main experiment, but having all subjects complete the prior experiment prevented an additional, undesirable difference between EV and RDU treatment groups, and furthermore ensured that we collected RDU parameter estimates for all subjects.

We closely followed the approach in Murad et al. (2016), which focuses exclusively on tasks in the gains domain and thus does not require the estimation of loss aversion parameters. To keep incentives consistent across our two-part study without sacrificing sample size, we scaled down the payoffs from Murad et al. (2016) by a factor of five and converted them into Points that, like in the main experiment, would be converted into monetary payoffs at the rate: 50 Points = £1.00.

Subjects completed 25 tables in random order, each of which required them to make a series of choices between a two-outcome lottery and a guaranteed amount of money. An example of such a table can be seen in Fig. 3.7: moving down the rows, the two-outcome lottery remained unchanged, while the guaranteed amount of money gradually decreased from the high outcome of the lottery towards the low outcome of the lottery. For each table, a subject’s certainty equivalent could then be approximated as the average between the guaranteed amounts before and after which they switched to the two-outcome lottery. Note that we enforced a single switch point in the “right” direction, that is, from the guaranteed amount towards the two-outcome lottery; thus, we did not need to account for the possibility of multiple switch points nor switches in the “wrong” direction, that is, from the two-outcome lottery towards the guaranteed amount. While this did, of course, restrict the ways in which subjects could express their preferences over the two options, the estimation of RDU parameters following Fehr-Duda et al. (2006) worked best with clear estimates of all 25 certainty equivalents⁴.

Fig. 3.8 summarises the 25 tables via the high lottery outcome Z_1 , the low lottery outcome Z_2 , as well as the probability $P(Z_1)$ with which the high lottery outcome occurred in the corresponding lottery.

⁴Furthermore, the accuracy of the estimation of a subject’s risk preferences was not our main focus. Our aim was to program an RDU algorithm that, unlike the EV algorithm, tailored its lottery selections to the estimation of a subject’s risk preferences — whether it performed well, objectively and in the eyes of the subjects, is a different issue that we address in Sections 3.2 and 3.3.

Bruhin et al. (2010) choose the underlying functional forms for the utility and the probability weighting functions based on parsimony, flexibility, as well as the ability to account for individual heterogeneity in risk preferences. Following their lead, we therefore use (1) the power form of the utility function, which has been shown to be the best compromise between parsimony and goodness of fit in the context of Cumulative Prospect Theory (Stott, 2006), and (2) the Goldstein & Einhorn (1987) linear-in-log-odds probability weighting function, which provides easily interpretable parameters and captures individual heterogeneity well (Wu et al., 2004):

$$U(x) = x^\alpha \tag{1}$$

$$w(p) = \frac{\beta p^\gamma}{\beta p^\gamma + (1 - p)^\gamma} \tag{2}$$

Consequently, our RDU estimation comprises parameter α , which controls curvature of the utility function, and parameters β and γ , which control the elevation and curvature of the probability weighting function. β governs the degree of ‘optimism’: in the gains domain, higher values of β elevate the probability weighting function, which places a higher weight on every probability and therefore values the corresponding prospects in a more optimistic sense. γ , on the other hand, measures sensitivity towards changes in probability and, for $\gamma < 1$, lower values indicate a stronger deviation from linearity.

Subjects were not told, at this stage, that their choices in the prior experiment would be used to estimate their risk preferences and program the RDU algorithm for the main experiment; they only learned that their choices in “Part 1 of the study” might affect what happens in “Part 2 of the study”, but received no further information on the main experiment itself or *how* their choices might affect it. Subjects received a flat fee of £2.00 for completing the RDU parameter estimation independently of their return for the main experiment. In an effort to minimise attrition, bonus payments, which ranged from £0.00 to £6.00 and were determined by a randomly selected row of a randomly selected table, were, however, paid out only upon completion of the main experiment.

Finally, it is worth noting that, in the choice elicitation task, in each row of a table, subjects were confronted with a trade-off between the probability and the size of the best outcome — not unlike the trade-off created by the juxtaposition of P-Bets and \$-Bets in each of the lottery menus in the main experiment. However, given the lack of time pressure and distractors in the form of dominated lotteries, we believe that the choice elicitation task made it easier for subjects to express their risk preferences.

	OPTION A	OPTION B	
	LOTTERY	GUARANTEED AMOUNT	
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	20
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	19
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	18
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	17
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	16
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	15
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	14
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	13
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	12
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	11
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	10
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	9
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	8
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	7
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	6
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	5
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	4
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	3
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	2
50% CHANCE OF 20 & 50% CHANCE OF 0	☐	☐	1

Figure 3.7: RDU Parameter Estimation Sample Table

	Z_1	Z_2	$P(Z_1)$		Z_1	Z_2	$P(Z_1)$
TABLE 1	40	0	0.05	TABLE 14	100	0	0.5
TABLE 2	80	20	0.05	TABLE 15	100	40	0.5
TABLE 3	100	40	0.05	TABLE 16	300	0	0.5
TABLE 4	300	100	0.05	TABLE 17	40	0	0.75
TABLE 5	20	0	0.1	TABLE 18	80	20	0.75
TABLE 6	40	20	0.1	TABLE 19	100	40	0.75
TABLE 7	100	0	0.1	TABLE 20	20	0	0.9
TABLE 8	40	0	0.25	TABLE 21	40	20	0.9
TABLE 9	80	20	0.25	TABLE 22	100	0	0.9
TABLE 10	100	40	0.25	TABLE 23	40	0	0.95
TABLE 11	20	0	0.5	TABLE 24	80	20	0.95
TABLE 12	40	20	0.5	TABLE 25	100	40	0.95
TABLE 13	80	20	0.5				

Note: 50 Points = £1.00. Z_1 denotes the high outcome and Z_2 denotes the low outcome of the lottery in each table; $P(Z_1)$ indicates the probability with which Z_1 occurred.

Figure 3.8: RDU Parameter Estimation Tables based on Murad et al. (2016)

2.3 Research Hypotheses

In line with previous studies documenting self-reliance bias (see e.g. Fehr et al., 2013; Bartling et al., 2014; Owens et al., 2014) and algorithm aversion (for an overview, see Burton et al., 2020; Jussupow et al., 2020), we expected to see an overall preference for self-reliance in the lottery selection task manifesting as an average proportion of automation reliance below 50% and a number of choices favouring automation reliance across all rounds of the lottery selection task that is, on average, below 4.

Hypothesis 1. *Subjects exhibit an overall preference for self-reliance.*

To the best of our knowledge, differences in the reception of normative versus preferential algorithms have yet to be studied. Consequently, it was unclear how or, indeed, *if* automation reliance is sensitive to whether a subject was offered the EV algorithm or the RDU algorithm as an alternative to self-reliance in our lottery selection task. Subjects could prefer the RDU algorithm because it tailors its selection to an estimation of their individual risk preferences; conversely, subjects could prefer the EV algorithm due to its consistent selection of the lottery that, on average, would yield the highest payoff — for example if they recognise their own tendency for excessive risk-taking or overcautiousness. At the same time, subjects might doubt the ability of the RDU algorithm to capture complex, human preferences or the alignment of the EV algorithm with their preferences. We therefore hypothesised a difference in automation reliance between the two types of algorithms without specifying its direction.

Hypothesis 2. *Automation reliance differs depending on whether subjects are offered the EV algorithm or the RDU algorithm.*

Following evidence of the sensitivity to algorithmic errors (Dietvorst et al., 2015; Hoff & Bashir, 2015; Prahll & Van Swol, 2017), we expected outcome feedback to affect subsequent automation reliance. Specifically, we expected satisfactory outcome feedback to reinforce previous automation reliance or self-reliance and unsatisfactory outcome feedback to induce changes from a subject’s previous decision. In addition to this, the overweighting of algorithmic relative to comparable human errors led us to hypothesise that the effect of unsatisfactory outcome feedback would be stronger when the lottery selection was made by the computer algorithm rather than the subject themselves.

Due to the relatively high likelihood of outcomes of zero as a result of the specific structure of our lottery menus, we therefore also expected automation reliance to decline over time for subjects in EV-F and RDU-F, who did receive outcome feedback. For subjects in in EV-NF and RDU-NF, who did not receive outcome feedback, we expected learning effects to either increase *or* decrease automation reliance over time, depending on whether subjects felt more or less confident in selecting a lottery themselves after performing a few rounds of the task. While any such effect would equally apply to subjects in EV-F and RDU-F, we expected it to be overshadowed by the effects of outcome feedback and the sensitivity to algorithmic errors.

Hypothesis 3a. *Automation reliance declines over time for subjects who receive outcome feedback.*

Hypothesis 3b. *Satisfactory feedback reinforces previous automation reliance or previous self-reliance, unsatisfactory feedback leads to changes from a subject’s previous decision.*

Hypothesis 3c. *Unsatisfactory feedback has a stronger effect on subsequent automation reliance or self-reliance when the lottery selection was made by the computer algorithm rather than by the subject themselves.*

Lastly, we also expected our subjects to sometimes end up selecting dominated lotteries as a result of self-reliance, which the EV algorithm and the RDU algorithm were programmed not to do: the EV algorithm compared the expected values of all lotteries within the menu and selected the one for which this was highest, which would always be the P-Bet or the \$-Bet but never any of their dominated variations; the RDU algorithm allowed for non-linear risk preferences but not for dominance violations and would therefore also only ever select the P-Bet or the \$-Bet.

Hypothesis 4. *Self-reliance, unlike automation reliance, sometimes results in the selection of dominated lotteries.*

3 Results

The experiment was programmed using Qualtrics software and conducted via Prolific (Palan & Schitter, 2018) in July 2023. We recruited a UK sample with a total of 399 subjects for the RDU parameter estimation, of which 377 (94.49%) returned for the main experiment. RDU parameter estimates did not significantly differ between those who returned and those who did not, so we will hereafter focus exclusively on the 377 subjects who returned for the main experiment.

Of these 377 subjects, 184 were female and the average age was 43.57 years⁵. In the RDU parameter estimation, subjects received a participation fee of £2.00 and could earn a bonus payment of up to £6.00 conditional on completion of the main experiment; among those who returned for the main experiment, the median completion time in the RDU parameter estimation was 19 minutes and 18 seconds and the average payment, including bonuses, was £3.27. In the main experiment, subjects received a participation fee of £1.00 and could earn a bonus payment of up to £4.50; the median completion time was 8 minutes and 28 seconds and the average payment, including bonuses, was £2.14.

Subjects were evenly divided across the four treatment groups, with 93 subjects in EV-NF, 94 subjects in EV-F, and 95 subjects each in RDU-NF and RDU-F. RDU parameter estimates did not significantly differ between treatment groups, with the following exceptions: γ estimates in RDU-NF were significantly lower than those in EV-NF ($p < 0.01$, Mann-Whitney U Test) and significantly lower than those in RDU-F ($p < 0.05$, Mann-Whitney U Test). As a consequence, the average number of \$-Bet choices of the RDU algorithm in RDU-NF (6.60) was slightly higher than in EV-NF (6.27) and RDU-F (6.21), but this difference is not statistically significant⁶. In other words, the RDU algorithm, if chosen, would have behaved similarly in all treatment groups. Where possible, we furthermore controlled for the effects of RDU parameter estimates on automation reliance and found them to be largely insignificant.

Consequently, Section 3.1 analyses the results of the main experiment — taking RDU parameter estimates as given — but we will revisit the results of the RDU parameter estimation and their consequences for the main experiment in Sections 3.2 and 3.3.

⁵Demographic sample characteristics and a treatment breakdown thereof can be found in Tables C.1 and C.2 in Appendix C.1.

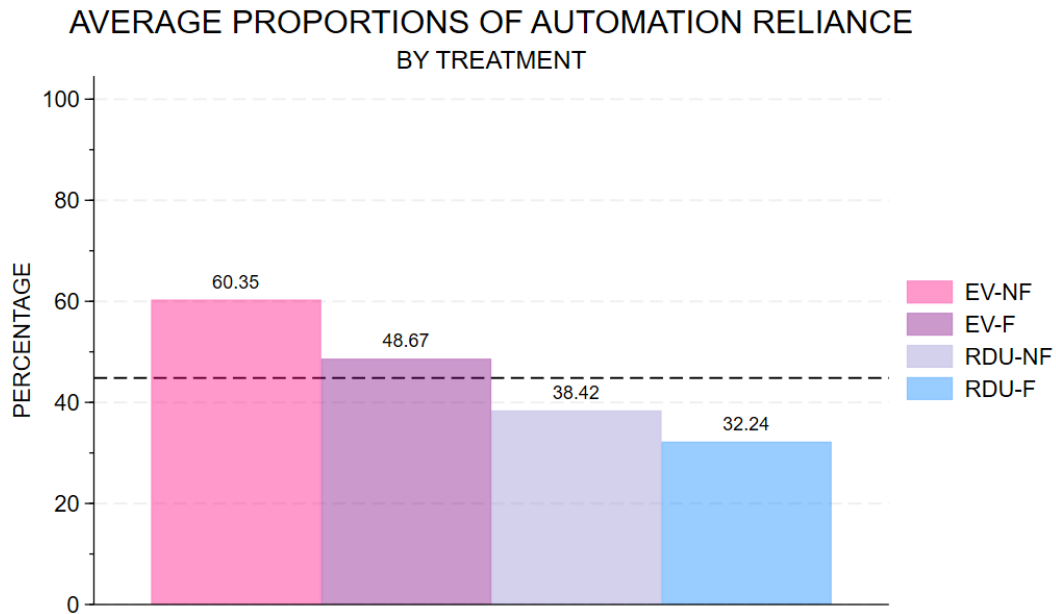
⁶Here, we are referring to the total number of \$-Bets that the RDU algorithm *would* have selected in each treatment group, *if* the RDU algorithm had been available to a subject and they had decided to rely on it in every round. Consequently, this number is hypothetical in EV-NF due to the unavailability of the RDU algorithm, and assumes consistent automation reliance in RDU-NF and RDU-F.

3.1 Main Experiment

Overall Automation Reliance

Subjects were shown the lottery menus in random order, so we need to distinguish between `MENUS` and `ROUNDS`, where the former refers to `MENUS A–H` as labelled in Fig. 3.6 and the latter captures the individual, temporal order in which a subject saw the lottery menus. Before we analyse the evolution of automation reliance over time in the different treatment groups, however, let us take a look at automation reliance aggregated across all rounds of the lottery selection task. Fig. 3.9 reports the average proportions of automation reliance across rounds in each of the treatment groups.

Hypothesis 1 posits that subjects exhibit an overall preference for self-reliance. Indeed, the proportion of automation reliance across rounds and treatment groups, indicated by the dashed line in Fig. 3.9, was 44.83% — so our subjects chose to have the computer algorithm select a lottery for them in less than half of the cases. However, this primarily reflects low average proportions of automation reliance in `RDU-NF` and `RDU-F`; in `EV-NF` and `EV-F`, these proportions were substantially higher than and close to 50%, respectively. In other words, we already see pronounced treatment differences: average proportions of automation reliance were generally higher for the `EV` than for the `RDU` algorithm, as well as among subjects who did not receive outcome feedback than among those who did.

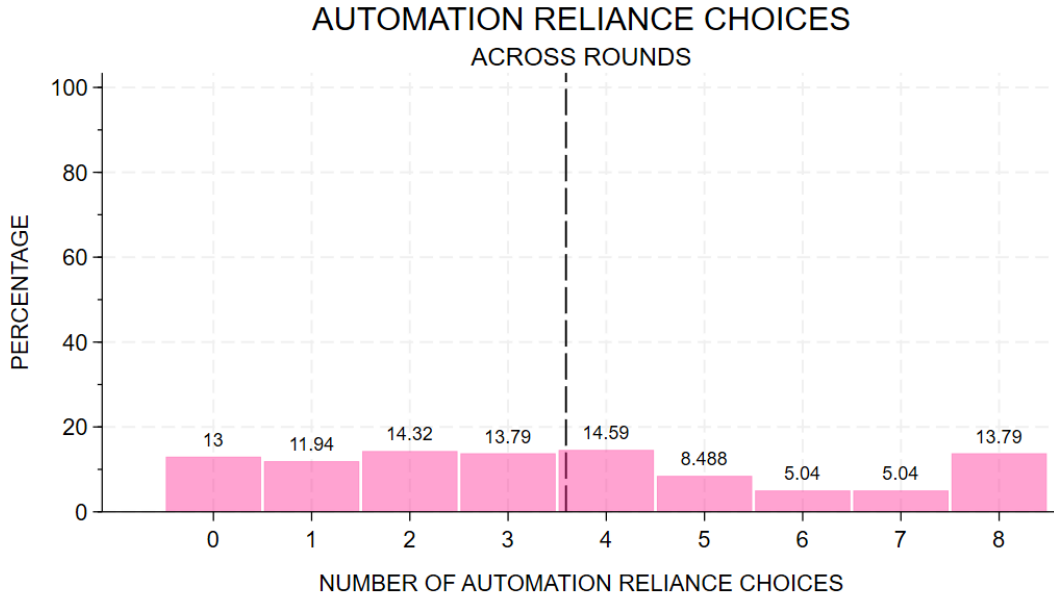


Note: Average proportions of automation reliance across all rounds of the lottery selection task in each of the treatment groups, with the dashed line indicating the proportion of automation reliance across rounds and treatment groups of 44.83%.

Figure 3.9: Proportions of Automation Reliance Across Rounds

Next, we examine the number of choices favouring automation reliance across rounds, that is, how many times — out of the maximum of 8 — a subject chose to have the computer algorithm select the lottery for them instead of selecting the lottery themselves. Fig. 3.10 shows that, at 13% and 13.79%, the proportions of subjects who always chose self-reliance and those who always chose automation reliance were similar. Nevertheless, over half of the subjects (53.05%) chose self-reliance more often than automation reliance, whereas the reverse was the case for less than a third (32.36%). It follows that the average number of choices favouring automation reliance across rounds was 3.59, which is significantly lower than 4 ($p < 0.01$, Two-Sided One Sample t-test). We conclude that **Hypothesis 1** is supported.

Result 1. *Subjects exhibited an overall preference for self-reliance.*



Note: Distribution of the number of choices favouring automation reliance across all rounds of the lottery selection task, with the dashed vertical line indicating the average number of choices favouring automation reliance of 3.59.

Figure 3.10: Number of Choices Favouring Automation Reliance

Round 1

ROUND 1 of our lottery selection task differed from all subsequent rounds because it was the only round in which subjects made their automation reliance decision without having experienced the lottery selection task, their own performance at it, or that of their treatment-specific algorithm. We therefore analyse ROUND 1 separately from ROUNDS 2 – 8.

Table 3.1 reports the results of Two-Sided Two-Sample Tests of Proportion for between-treatment equality of automation reliance in ROUND 1. Holding constant the computer algorithm that was offered to the subjects, there is no significant difference between treatment groups with and without outcome feedback, which is what we would expect in ROUND 1. Holding constant whether or not subjects would receive outcome feedback or not, we do, however, see a significant difference between treatments in which subjects were offered the EV algorithm and treatments in which subjects were offered the RDU algorithm: the proportion of automation reliance in EV-NF was significantly higher than that in RDU-NF and the proportion of automation reliance in EV-F was significantly higher than that in RDU-F.

	AR ₁	AR ₂	p (H ₀ : AR ₁ = AR ₂) <i>Two-Sample Test of Proportion</i>
ROUND 1	0.57		
EV-NF – EV-F	0.71	0.64	0.2979
RDU-NF – RDU-F	0.42	0.49	0.3081
EV-NF – RDU-NF	0.71	0.42	0.0001
EV-F – RDU-F	0.64	0.49	0.0465

Note: AR₁ and AR₂ denote the two proportions of automation reliance in each row for whose between-treatment equality we test using Two-Sided Two-Sample Tests of Proportion. Note that, in each round, we aggregate automation reliance across different lottery menus as these were presented to subjects in individually randomised order.

Table 3.1: Treatment Comparisons of Automation Reliance: Round 1

We also perform OLS regressions that control for RDU parameter estimates and demographic characteristics. Our dependent variable is a dummy that takes the value of 1 for subjects who, in the first round of the task, chose to have the computer algorithm select the lottery for them and a value of 0 for those who, in the first round of the task, chose to select the lottery themselves. Explanatory variables include two dummies indicating whether subjects were offered the RDU algorithm or the EV algorithm and whether they received outcome feedback or not, as well as their interaction term.

Overall, the three OLS regressions reported in Table 3.2 support our results thus far. Subjects who were offered the RDU algorithm, *ceteris paribus*, were significantly less likely to have it select a lottery for them than those who were offered the EV algorithm; this effect is robust to the inclusion of RDU parameter estimates as well as demographic controls for gender, age, education, income, and programming skills. Conversely, whether or not a subject received outcome feedback — as was to be expected in ROUND 1 — had no significant effect on automation reliance, neither in isolation nor in conjunction with the type of algorithm that was offered. Similarly, once we include demographic controls, individual RDU parameter estimates also had no significant effect on automation reliance.

In summary, in ROUND 1, prior to experiencing the lottery selection task, subjects exhibited a significant preference for the EV algorithm over the RDU algorithm.

	AUTOMATION RELIANCE		
	(1)	(2)	(3)
RDU ALGORITHM	−0.29*** (0.07)	−0.28*** (0.07)	−0.27*** (0.08)
FEEDBACK	−0.07 (0.07)	−0.07 (0.07)	−0.04 (0.07)
RDU ALGORITHM × FEEDBACK	0.15 (0.10)	0.12 (0.10)	0.08 (0.11)
ALPHA		0.05* (0.03)	0.05 (0.03)
BETA		0.04 (0.05)	0.08 (0.08)
GAMMA		0.04 (0.07)	0.03 (0.07)
CONSTANT	0.71*** (0.05)	0.59*** (0.08)	0.16 (0.17)
Demographic Controls	No	No	Yes
Observations	377	367	340
Prob > F	0.0002	0.0003	0.0013
R-squared	0.0528	0.0671	0.0878
Root MSE	0.4851	0.4842	0.4828

Note: Dependent variable AUTOMATION RELIANCE is a dummy that takes the value of 1 for subjects who chose to have the computer algorithm select the lottery for them in the first round of the lottery selection task. Demographic controls include gender, age, education, income, and programming skills. Standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.2: OLS Regressions: First Round Automation Reliance

Rounds 2 – 8

Turning our attention to ROUNDS 2 – 8, Table 3.3 reports the results of Two-Sided Two-Sample Tests of Proportion for between-treatment equality of automation reliance⁷. Here as well, in all but one case – the comparison of EV-F and RDU-F in ROUND 4 – the proportion of automation reliance in EV-NF was significantly higher than that in RDU-NF and the proportion of automation reliance in EV-F was significantly higher than that in RDU-F.

⁷See Table C.3 in Appendix C.1 for treatment comparisons of automation reliance in each lottery menu.

	AR ₁	AR ₂	p (H ₀ : AR ₁ = AR ₂) <i>Two-Sample Test of Proportion</i>
ROUND 2	0.55		
EV-NF – EV-F	0.66	0.57	0.2525
RDU-NF – RDU-F	0.52	0.44	0.3094
EV-NF – RDU-NF	0.66	0.52	0.0512
EV-F – RDU-F	0.57	0.44	0.0688
ROUND 3	0.49		
EV-NF – EV-F	0.62	0.56	0.4049
RDU-NF – RDU-F	0.40	0.37	0.6546
EV-NF – RDU-NF	0.62	0.40	0.0022
EV-F – RDU-F	0.56	0.37	0.0071
ROUND 4	0.40		
EV-NF – EV-F	0.56	0.41	0.0485
RDU-NF – RDU-F	0.31	0.34	0.6411
EV-NF – RDU-NF	0.56	0.31	0.0004
EV-F – RDU-F	0.41	0.34	0.2679
ROUND 5	0.44		
EV-NF – EV-F	0.65	0.45	0.0065
RDU-NF – RDU-F	0.40	0.26	0.0451
EV-NF – RDU-NF	0.65	0.40	0.0008
EV-F – RDU-F	0.45	0.26	0.0083
ROUND 6	0.40		
EV-NF – EV-F	0.57	0.43	0.0484
RDU-NF – RDU-F	0.34	0.26	0.2678
EV-NF – RDU-NF	0.57	0.34	0.0013
EV-F – RDU-F	0.43	0.26	0.0188
ROUND 7	0.40		
EV-NF – EV-F	0.58	0.44	0.0482
RDU-NF – RDU-F	0.37	0.21	0.0164
EV-NF – RDU-NF	0.58	0.37	0.0036
EV-F – RDU-F	0.44	0.21	0.0009
ROUND 8	0.35		
EV-NF – EV-F	0.48	0.39	0.2136
RDU-NF – RDU-F	0.33	0.20	0.0480
EV-NF – RDU-NF	0.48	0.33	0.0277
EV-F – RDU-F	0.39	0.20	0.0036

Note: AR₁ and AR₂ denote the two proportions of automation reliance in each row for whose between-treatment equality we test using Two-Sided Two-Sample Tests of Proportion. Note that, in each round, we aggregate automation reliance across different lottery menus as these were presented to subjects in individually randomised order.

Table 3.3: Treatment Comparisons of Automation Reliance: Rounds 2 – 8

Now, however, we also start to see some differences in automation reliance between subjects who received outcome feedback and those who did not: in ROUNDS 4 – 7, the proportion of automation reliance is significantly higher in EV-NF than in EV-F; in ROUNDS 5, 7 and 8, the proportion of automation reliance is significantly higher in RDU-NF than in RDU-F. Consequently, over time, it appears that receiving outcome feedback did have a negative – albeit less consistent – effect on automation reliance for both types of algorithm.

In summary, we consistently observe a higher level of automation reliance for the EV algorithm than for the RDU algorithm, which lends support to **Hypothesis 2**. It is also worth noting that, while perhaps exacerbated by it, this difference does not appear to be *caused* by positive or negative experience with either algorithm: even when focusing only on ROUND 1, we saw that reliance on the EV algorithm was significantly higher than reliance on the RDU algorithm⁸.

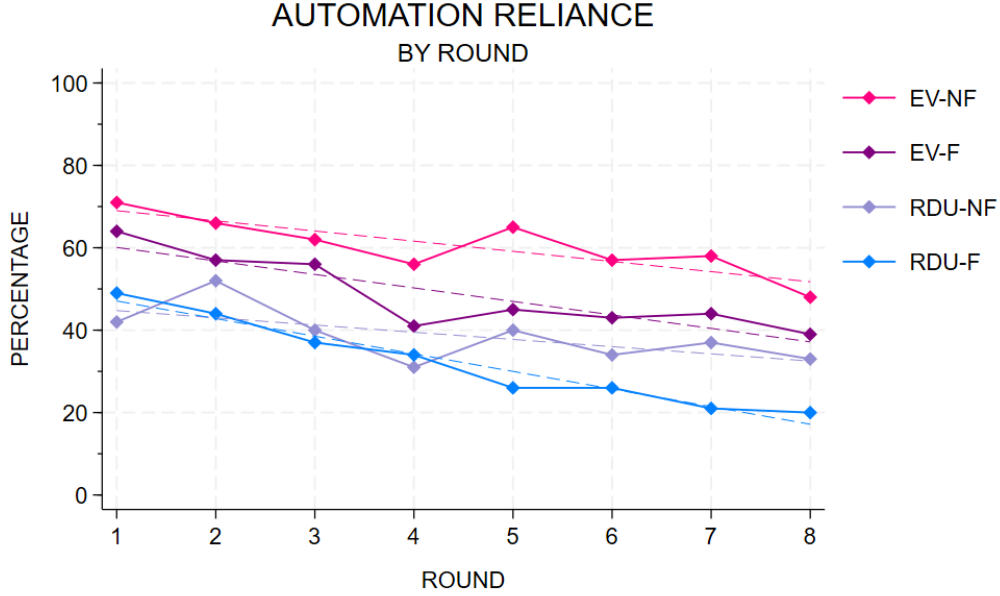
Result 2. *Automation reliance was significantly higher among subjects who were offered the EV algorithm than among those who were offered the RDU algorithm, even in the first round of the lottery selection task.*

Fig. 3.11 illustrates the steady decline of automation reliance over time. In ROUND 1, 71% of the subjects in EV-NF, 64% of the subjects in EV-F, 42% of the subjects in RDU-NF, and 49% of the subjects in RDU-F chose to rely on the computer; by ROUND 8, these proportions had reduced to 48%, 39%, 33%, and 20%, respectively.

We furthermore plotted the predictions from linear regressions of automation reliance on the round of the task for each treatment group, represented by the dashed lines in Fig. 3.11: for EV-NF we obtain a coefficient of -2.46 (std. err.: 0.64 , $p < 0.01$), for EV-F we obtain a coefficient of -3.27 (std. err.: 0.71 , $p < 0.01$), for RDU-NF we obtain a coefficient of -1.75 (std. err.: 0.84 , $p < 0.10$), and for RDU-F we obtain a coefficient of -4.27 (std. err.: 0.35 , $p < 0.01$). In line with **Hypothesis 3a**, automation reliance declined over time for subjects who received outcome feedback. While it also declined for those who did not receive outcome feedback, the size of the linear regression coefficients indicate a faster rate of decline in EV-F than in EV-NF and in RDU-F than in RDU-NF.

Result 3a. *Automation reliance significantly declined over time in all treatment groups, but the rate of decline for subjects who did receive outcome feedback exceeded the rate of decline for those who did not.*

⁸We recognise that there is a multitude of possible explanations for why reliance on the RDU algorithm is significantly lower than reliance on the EV algorithm from the start, including beliefs about the RDU parameter estimation, and will discuss these further in Section 4.



Note: Dashed lines represent predictions from linear regressions of automation reliance on the round of the task for each treatment group. Note that, in each round, we aggregate automation reliance across different lottery menus as these were presented to subjects in individually randomised order.

Figure 3.11: Automation Reliance Over Time

Table 3.4 reports the average round-to-round transition frequencies underlying the evolution of automation reliance over time shown in Fig. 3.11. Firstly, we see that switching frequencies — indicating that subjects, from one round to the next, switched from self-reliance to automation reliance (SR-AR) or from automation reliance to self-reliance (AR-SR) — were generally lower than frequencies indicating that subjects, from one round to the next, stuck with self-reliance (SR-SR) or automation reliance (AR-AR).

In line with what we have seen so far, subjects were more likely to stick with self-reliance than with automation reliance. Notably, the only treatment group in which this was not the case is EV-NF — the treatment group in which automation reliance was consistently highest at above 50% until its drop to 48% in the last round. Furthermore, the gap between SR-SR and AR-AR is substantially larger in RDU-NF than in RDU-F. In line with previous observations, switches from automation reliance to self-reliance were more frequent than switches from self-reliance to automation reliance in all treatment groups.

In summary, the decline in automation reliance appears to have been caused by a higher frequency of switches from automation reliance to self-reliance than vice-versa, after which subjects frequently continued to choose self-reliance.

	SR–SR	AR–AR	SR–AR	AR–SR
EV-NF	20.74	41.63	17.20	20.43
EV-F	38.75	35.26	11.25	14.74
RDU-NF	43.16	20.30	17.59	18.95
RDU-F	55.04	18.79	10.98	15.19

Note: Average frequencies of subjects who, in each treatment group, from one round to the next, stuck with self-reliance (SR–SR), stuck with automation reliance (AR–AR), switched from self-reliance to automation reliance (SR–AR), or switched from automation reliance to self-reliance (AR–SR).

Table 3.4: Average Round-to-Round Transition Frequencies

An obvious explanation for why automation reliance declined in all treatment groups is, of course, learning: if subjects performed better at the lottery selection task over time, or at the very least believed themselves to perform better, this would have lowered incentives to rely on either algorithm. While we cannot make any conclusive statements about subjective beliefs regarding a subject’s own performance relative to that of either algorithm, we are able to examine objective performance in terms of lottery type selection, which we will discuss in Section 3.2. In addition to this, it is possible that the decline in automation reliance was a reaction to *selection* rather than *outcome* feedback; however, this would not explain the higher rates of decline in EV-F than in EV-NF and in RDU-F than in RDU-NF.

Next, we therefore directly examine the effect of outcome feedback on subsequent automation reliance. Our analysis includes satisfaction ratings of 1 – 5, where 1 indicated “not satisfied at all” and 5 indicated “completely satisfied” with the lottery selection that was made either by a subject themselves or by one of the two algorithms. Recall that, in spite of the fact that we enquired about satisfaction with lottery selection, we expected to capture satisfaction with lottery *outcome* rather than lottery *selection* in EV-F and RDU-F due to outcome bias.

In support of this assumption, outcome feedback and satisfaction with lottery selection are highly correlated (Spearman’s $\rho = 0.77$, $p < 0.01$). Consequently, satisfaction ratings in EV-F and RDU-F appear to be a suitable measure of satisfaction with lottery outcome. It is also worth noting that the correlation between outcome feedback and satisfaction ratings is stronger when subjects chose to have the computer algorithm select the lottery for them (Spearman’s $\rho = 0.81$, $p < 0.01$) than when they chose to select the lottery themselves (Spearman’s $\rho = 0.70$, $p < 0.01$); this indicates stronger reactions to outcome feedback related to the computer algorithm’s selection.

Table 3.5 reports correlations between automation reliance, outcome feedback, and satisfaction ratings. Firstly, we are interested in how outcome feedback and satisfaction ratings from the previous round (FB₋₁ and SF₋₁) affect the likelihood of automation

reliance in the current round (AR), depending on whether the lottery in the previous round was selected by the subject themselves ($AR_{-1} = \text{SELF}$) or by the computer algorithm ($AR_{-1} = \text{COMP}$). Secondly, we also look at the correlation of automation reliance in the current round (AR) with outcome feedback and satisfaction ratings for the current round (FB and SF) to see whether automation reliance led to significantly higher or lower outcome feedback or significantly higher or lower satisfaction ratings than self-reliance.

$AR_{-1} =$	FB_{-1} SELF	FB_{-1} COMP	SF_{-1} SELF	SF_{-1} COMP	FB	SF
AR						
EV-NF			-0.04	0.41***		-0.02
EV-F	-0.28***	0.20***	-0.32***	0.19***	-0.02	-0.13***
RDU-NF			-0.13**	0.46***		-0.31***
RDU-F	-0.18***	0.34***	-0.18***	0.40***	-0.18***	-0.29***

Note: AR = Automation Reliance (1 if “yes”, 0 if “no”), FB = Outcome Feedback (applicable only to EV-F and RDU-F), SF = Satisfaction Rating (1 to 5, where 1 indicates “not satisfied at all” and 5 indicates “completely satisfied”). Subscript $_{-1}$ indicates previous period. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.5: Correlations: Automation Reliance, Feedback & Satisfaction

Previous outcome feedback and satisfaction ratings are significantly correlated with current automation reliance in all treatment groups except for EV-NF. Moreover, again except for EV-NF, these correlations go in the expected directions: previous outcome feedback and satisfaction ratings are negatively correlated with current automation reliance when subjects selected the lottery themselves, and positively correlated with current automation reliance when the computer algorithm did. Simply put, supporting **Hypothesis 3b**, subjects who received satisfactory outcome feedback resulting from automation reliance in the previous round were more likely to again choose automation reliance and vice-versa for those who received unsatisfactory outcome feedback; subjects who received satisfactory outcome feedback resulting from self-reliance in the previous round were more likely to again choose self-reliance and vice-versa for those who received unsatisfactory outcome feedback.

Result 3b. *Satisfactory outcome feedback reinforced previous automation reliance or self-reliance, unsatisfactory outcome feedback led to changes from automation reliance or self-reliance.*

Given previous findings of the human sensitivity to algorithmic errors, we hypothesised that unsatisfactory outcome feedback would have a stronger effect on subsequent automation reliance decisions when the computer selected the lottery than when subjects selected

the lottery themselves. In RDU-F, it is indeed the case that ρ almost doubles in absolute value when the computer selected the lottery: in other words, the effect of outcome feedback from the previous round on automation reliance in the current round is twice as strong when the computer selected the lottery than when a subject selected the lottery themselves. However, this difference is reversed — albeit not as pronounced — in EV-F. Satisfaction ratings also reflect this: EV-F is the only treatment group in which the correlation between satisfaction ratings for the previous round and automation reliance in the current round is stronger for subjects who selected the lottery themselves; in all other treatment groups, the effect of satisfaction ratings for the previous round on automation reliance in the current round is over twice as strong when the computer selected the lottery in the previous round than when a subject selected the lottery themselves. **Hypothesis 3c** is therefore partially supported.

Result 3c. *Unsatisfactory outcome feedback has a stronger effect on subsequent automation reliance or self-reliance when the lottery selection was made by the RDU algorithm rather than by a subject themselves.*

Outcome feedback in the current round is furthermore negatively correlated with automation reliance in the current round in RDU-F but not in EV-F; this is likely a byproduct of the RDU algorithm’s frequent selection of \$-Bets, which generally had a much higher likelihood of zero outcomes than P-Bets⁹. Interestingly, satisfaction ratings for the current round, in all treatment groups except EV-NF, are significantly negatively correlated with automation reliance in the current round. In other words, when subjects chose to have the computer algorithm select the lottery for them, this generally resulted in lower satisfaction ratings even in EV-F, where this is not associated with significantly lower outcome feedback. Consequently, subjects were generally more satisfied with their own lottery selections than those of either computer algorithm.

Finally, we complement our previous OLS regressions of first round automation reliance with an analogous analysis for final round automation reliance to check whether our treatment differences survived experience with the lottery selection task. Like before, our dependent variable is a dummy that takes the value of 1 for subjects who, in the final round of the task, chose to have the computer algorithm select the lottery for them and a value of 0 for those who, in the final round of the task, chose to select the lottery themselves. Explanatory variables include two dummies indicating whether subjects were offered the RDU algorithm or the EV algorithm and whether they received outcome feedback or not, as well as their interaction term. We again controlled for RDU parameter estimates and demographic characteristics including gender, age, education, income, and programming skills.

⁹We explore the reasons for and consequences of the RDU algorithm’s frequent selection of \$-Bets in Sections 3.2 and 3.3.

	AUTOMATION RELIANCE		
	(1)	(2)	(3)
RDU ALGORITHM	−0.16** (0.07)	−0.14** (0.07)	−0.13* (0.07)
FEEDBACK	−0.09 (0.07)	−0.08 (0.07)	−0.08 (0.07)
RDU ALGORITHM × FEEDBACK	−0.04 (0.10)	−0.07 (0.10)	−0.07 (0.10)
ALPHA		−0.00 (0.03)	−0.02 (0.03)
BETA		0.05 (0.05)	0.05 (0.08)
GAMMA		0.08 (0.07)	0.07 (0.07)
CONSTANT	0.48*** (0.05)	0.42*** (0.08)	0.28 (0.17)
Demographic Controls	No	No	Yes
Observations	377	367	340
Prob > F	0.0004	0.0009	0.0262
R-squared	0.0471	0.0611	0.0634
Root MSE	0.4681	0.4671	0.4691

Note: Dependent variable AUTOMATION RELIANCE is a dummy that takes the value of 1 for subjects who chose to have the computer select the lottery for them in the final round of the task. Demographic controls include gender, age, education, income, and programming skills. Standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.6: OLS Regressions: Final Round Automation Reliance

Table 3.6 shows that automation reliance in the final round was still significantly lower among subjects who were offered the RDU instead of the EV algorithm; this effect is robust to the inclusion of RDU parameter estimates as well as demographic controls. Whether or not a subject received outcome feedback had no significant effect on automation reliance, neither in isolation nor in conjunction with the type of algorithm that was offered; however, this is likely due to the fact that our model specifications do not capture the gradual decline of automation reliance over time. Like in the first round, neither RDU parameters nor demographic characteristics had a significant effect on automation reliance in the final round.

Hypothetical Automation Reliance

Let us now briefly examine the hypothetical continuation of the lottery selection task. Recall that, by ROUND 8, automation reliance had dropped to 48% in EV-NF, 39% in EV-F, 33% in RDU-NF, and 20% in RDU-F; in the hypothetical continuation of the lottery selection task, these levels sank even further to 23% in RDU-NF and 15% in RDU-F, and stayed more or less the same at 47% in EV-NF and 41% in EV-F.

Mirroring previous results (cf. Tables 3.1 and 3.3), Table 3.7 again shows significantly higher levels of automation reliance among subjects who were offered the EV algorithm instead of the RDU algorithm, but no comparable effect with respect to our outcome feedback manipulation. Holding the offered algorithm constant, automation reliance was still lower when subjects received outcome feedback than when they did not, but this difference is not statistically significant¹⁰.

	AR ₁	AR ₂	p (H ₀ : AR ₁ = AR ₂) <i>Mann-Whitney U Test</i>
HYPOTHETICAL	0.32		
EV-NF — EV-F	0.47	0.41	0.4242
RDU-NF — RDU-F	0.23	0.15	0.1396
EV-NF — RDU-NF	0.47	0.23	0.0005
EV-F — RDU-F	0.41	0.15	0.0000

Note: AR₁ and AR₂ denote the two proportions of hypothetical automation reliance in each row for whose between-treatment equality we test using Two-Sided Two-Sample Tests of Proportion.

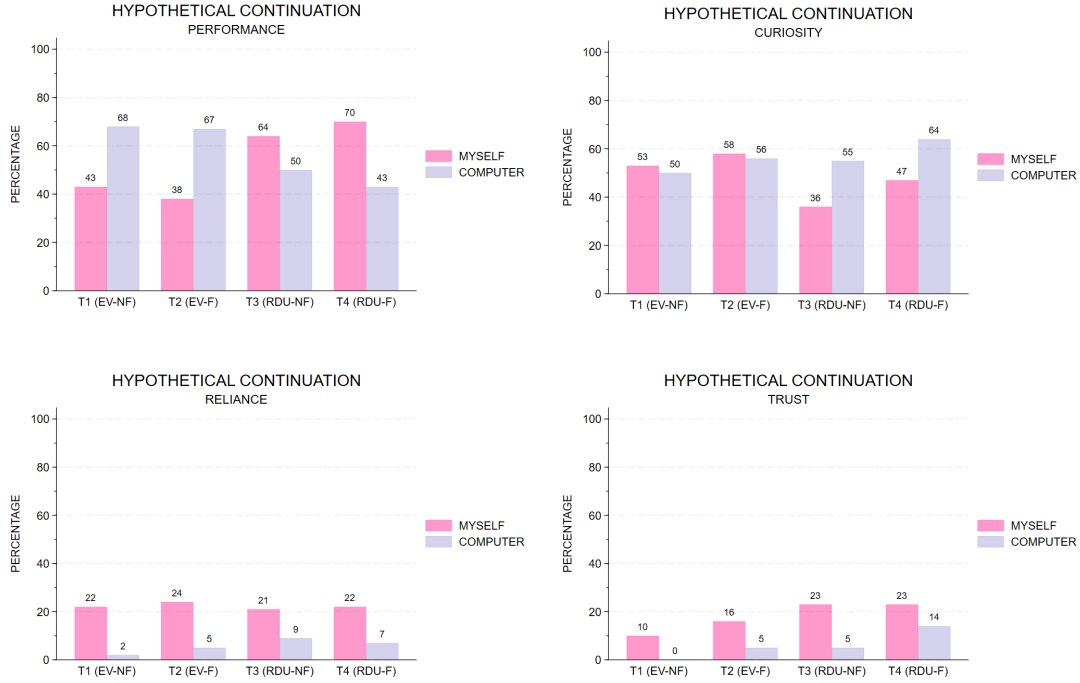
Table 3.7: Treatment Comparisons of Hypothetical Continuation

Next, we evaluate self-reported reasons for automation reliance or self-reliance in the hypothetical continuation of the lottery selection task. Subjects could select multiple reasons from a list and, as shown in Fig. 3.12, the most commonly reported reasons were PERFORMANCE (“I can do a better job than the computer/ the computer can do a better job than me”) and CURIOSITY (“I want to see how well I can do/ I want to see how well the computer can do”), followed by RELIANCE (“I don’t like relying on anyone but myself/ I don’t like relying on myself”) and TRUST (“I don’t trust the computer/ I don’t trust myself”) (see Table C.5 in Appendix C.1 for detailed summary statistics).

We find that performance-related reasons are significantly more common among those who indicated a preference for automation reliance in EV-NF ($p < 0.05$, Two-Sided Two-Sample Test of Proportion) as well as in EV-F ($p < 0.05$, Two-Sided Two-Sample Test of Proportion). Conversely, performance-related reasons are significantly more

¹⁰In line with this, Table C.4 in Appendix C.1 shows OLS regression results for the hypothetical continuation of the lottery selection task that are similar to those for ROUNDS 1 and 8.

common among subjects who indicated a preference for self-reliance in RDU-F ($p < 0.05$, Two-Sided Two-Sample Test of Proportion). For curiosity-related reasons we see the opposite pattern, but none of these differences are significant; this might, at least in part, be due to the small proportions of subjects who did indicate a preference for automation reliance in RDU-NF (23%) and RDU-F (15%), and for whom curiosity in the RDU algorithm appears to have been a prevalent reason for this preference.



Note: Proportions of subjects who selected PERFORMANCE, CURIOSITY, RELIANCE, and TRUST as a reason for their decision in the hypothetical continuation of the lottery selection task, depending on whether they expressed a preference for self-reliance or automation reliance.

Figure 3.12: Reasons for Decision in Hypothetical Continuation

Lastly, reliance-related reasons were significantly more common among those with a preference for self-reliance in EV-NF ($p < 0.01$, Two-Sided Two-Sample Test of Proportion) and EV-F ($p < 0.05$, Two-Sided Two-Sample Test of Proportion); the same applies for trust-related reasons in EV-NF ($p < 0.05$, Two-Sided Two-Sample Test of Proportion), EV-F ($p < 0.10$, Two-Sided Two-Sample Test of Proportion), and RDU-NF ($p < 0.10$, Two-Sided Two-Sample Test of Proportion). Ultimately, however, the low proportions of subjects who cited trust-related or reliance-related reasons for their preference of self-reliance over automation reliance support trust as an unlikely cause of suboptimal automation reliance (cf. Gogoll & Uhl, 2018; Schniter et al., 2020) and furthermore suggest that the frequently observed unwillingness to give up one's own decision rights in human-human delegation studies (see e.g. Fehr et al., 2013; Neri & Rommeswinkel, 2017; Bartling et al., 2014; Owens et al., 2014) might not be as influential in human-computer interactions either.

3.2 Lottery Type Selection

Next, due to the inclusion of P-Bets, \$-Bets, and their dominated variations in each of our lottery menus, we also need to examine lottery type selections resulting from self-reliance versus automation reliance — both from a performance-centric perspective as well as to account for the role differences in lottery type selection played in the evolution of automation reliance over time. Fig. 3.13 and 3.14 compare lottery type selections depending on whether a subject selected the lottery themselves, whether the EV algorithm did, or whether the RDU algorithm did. Fig. 3.13 illustrates lottery type selections in ROUNDS 1 — 8; Fig 3.14 illustrates lottery type selections in MENUS A — H.

As per our programming, both the EV algorithm and the RDU algorithm always selected either the P-Bet or the \$-Bet, but never any of their dominated variations. In line with **Hypothesis 4**, the proportion of subjects who selected dominated variations of the P-Bet, while smaller than the proportion of subjects who selected the P-Bet itself, fluctuated between 44% and 26%. Overall, subjects rarely selected \$-Bets or their dominated variations; however, the proportion of subjects who did select dominated variations of the \$-Bet fluctuated between 18% and 4% and was, on average, higher than the proportion of subjects who selected the \$-Bet itself. In summary, even in ROUND 8, after substantial experience with the task, 37% of the subjects who chose self-reliance still ended up selecting a dominated lottery.

Result 4. *Self-reliance, unlike automation reliance, frequently resulted in the selection of dominated lotteries.*

Another key difference between those who selected the lottery themselves and those who had the computer algorithm select the lottery for them is the proportion of \$-Bet selections: the EV algorithm and the RDU algorithm, when chosen, selected \$-Bets 51% and 78% of the time, respectively, whereas subjects themselves selected \$-Bets and their dominated variations only 11% of the time (see Fig. C.1 and C.2 in Appendix C.1 for numbers of algorithmic \$-Bet selections).

Fig. 3.14 shows that, in MENUS A — D, where the expected value of the \$-Bet was higher than that of the P-Bet, the two types of algorithm nearly always coincided in their lottery type selection; our subjects, on the other hand, were much more likely to select the P-Bet or one of its dominated variations. In MENUS E — H, we see a difference between the two types of algorithm: given that the expected value of the P-Bet was now higher than that of the \$-Bet, the EV algorithm consistently selected the P-Bet; the RDU algorithm, on the other hand, selected the P-Bet more frequently than in MENUS A — D but still less frequently than it selected the \$-Bet. Our subjects, again, were much more likely to select the P-Bet or one of its dominated variations, which is now in line with the behaviour of the EV algorithm but still very similar to

lottery type selections in MENUS A – D in spite of the reversal of the expected value difference between P-Bets and \$-Bets.

Given that Fig. 3.13 and 3.14 comprise only *realised* lottery type selections, it could be the case that, in RDU-NF and RDU-F, risk preferences between subjects who chose to select the lottery themselves and those who chose to have the computer algorithm select the lottery for them differed significantly; leading to a higher proportion of \$-Bet selections from the RDU algorithm and a higher proportion of P-Bet selections from subjects themselves. A look at *non-realised* lottery type selections for both algorithms – i.e. lottery selections that the two algorithms would have made if chosen instead of self-reliance – does, however, rule this out: Fig. 3.15 shows that realised and non-realised lottery type selections for both types of algorithm are similar across rounds and their proportions of \$-Bet selections heavily outweigh the proportion of \$-Bet selections resulting from self-reliance.

Result 5a. *Self-reliance resulted in the selection of \$-Bets much less frequently than reliance on the EV algorithm or the RDU algorithm did.*

Result 5b. *Reliance on the EV algorithm resulted in the selection of \$-Bets much less frequently than reliance on the RDU algorithm did.*

Consequently, we conclude that lottery type selections resulting from self-reliance significantly differ from those resulting from reliance on the RDU algorithm. One possible explanation for this could be that our RDU parameter estimation failed to capture the true risk preferences of our subjects. We discuss this possibility in Section 4 after examining the results of the RDU parameter estimation. Crucially, however, the frequent selection of P-Bets resulting from self-reliance is not just at odds with the lottery selections of the RDU algorithm but also with those of humans in pairwise choices (see e.g. Tversky et al., 1990; Cubitt et al., 2004) who do not exhibit a clear and consistent preference for P-Bets over \$-Bets.

It is therefore also possible that the RDU algorithm predicted what our subjects would have chosen in pairwise P-Bet versus \$-Bet choices but that, in the lottery selection task, the complexity of the decision environment pushed subjects to select P-Bets far more often than they would have in pairwise choices. In fact, Fig. 3.14 suggests that the proportion of P-Bet selections resulting from self-reliance was not sensitive to expected value differences between P-Bets and \$-Bets, which is indicative of the tendency to select P-Bets irrespective of individual risk preferences that the RDU algorithm was programmed to account for.

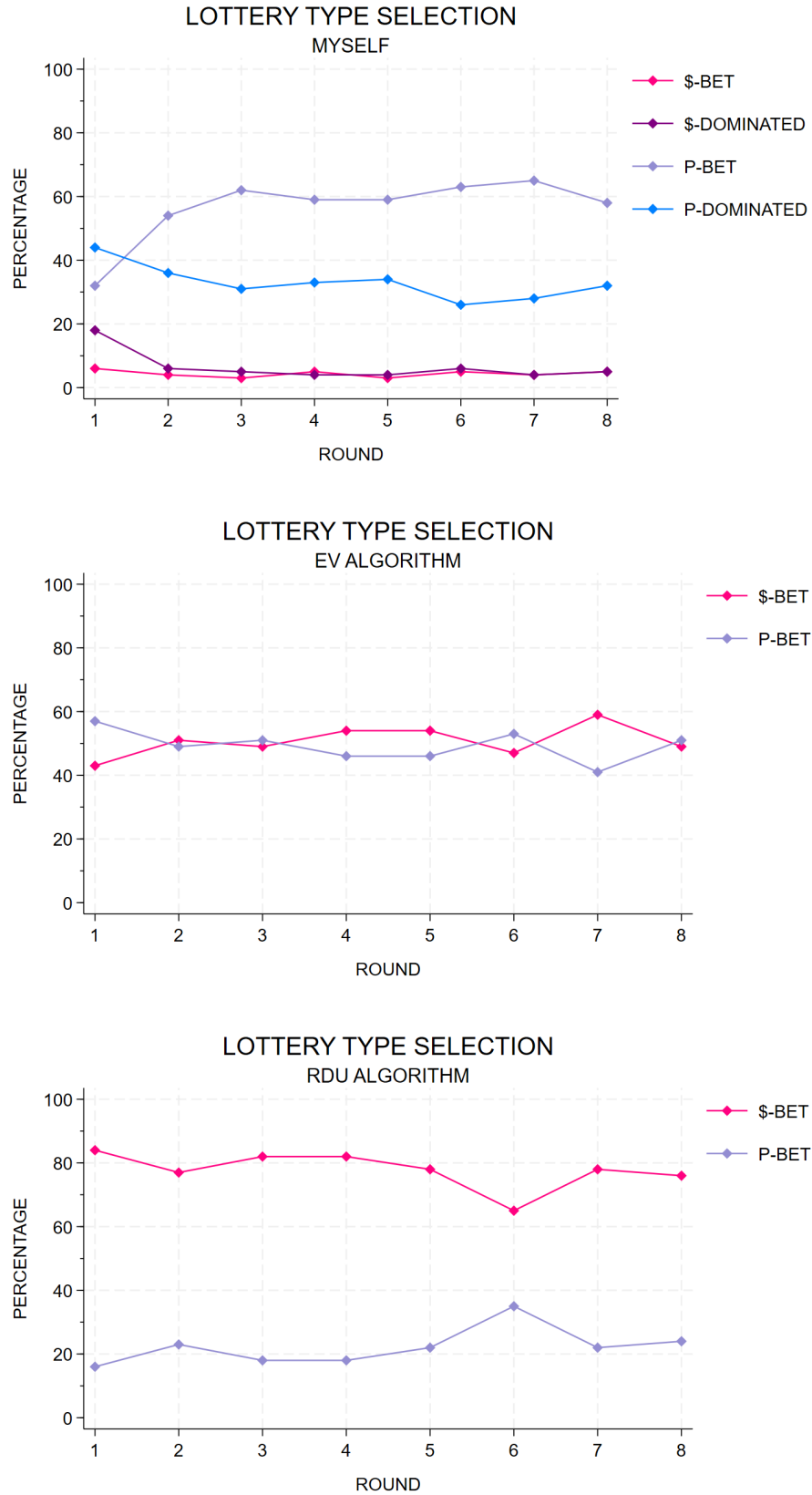
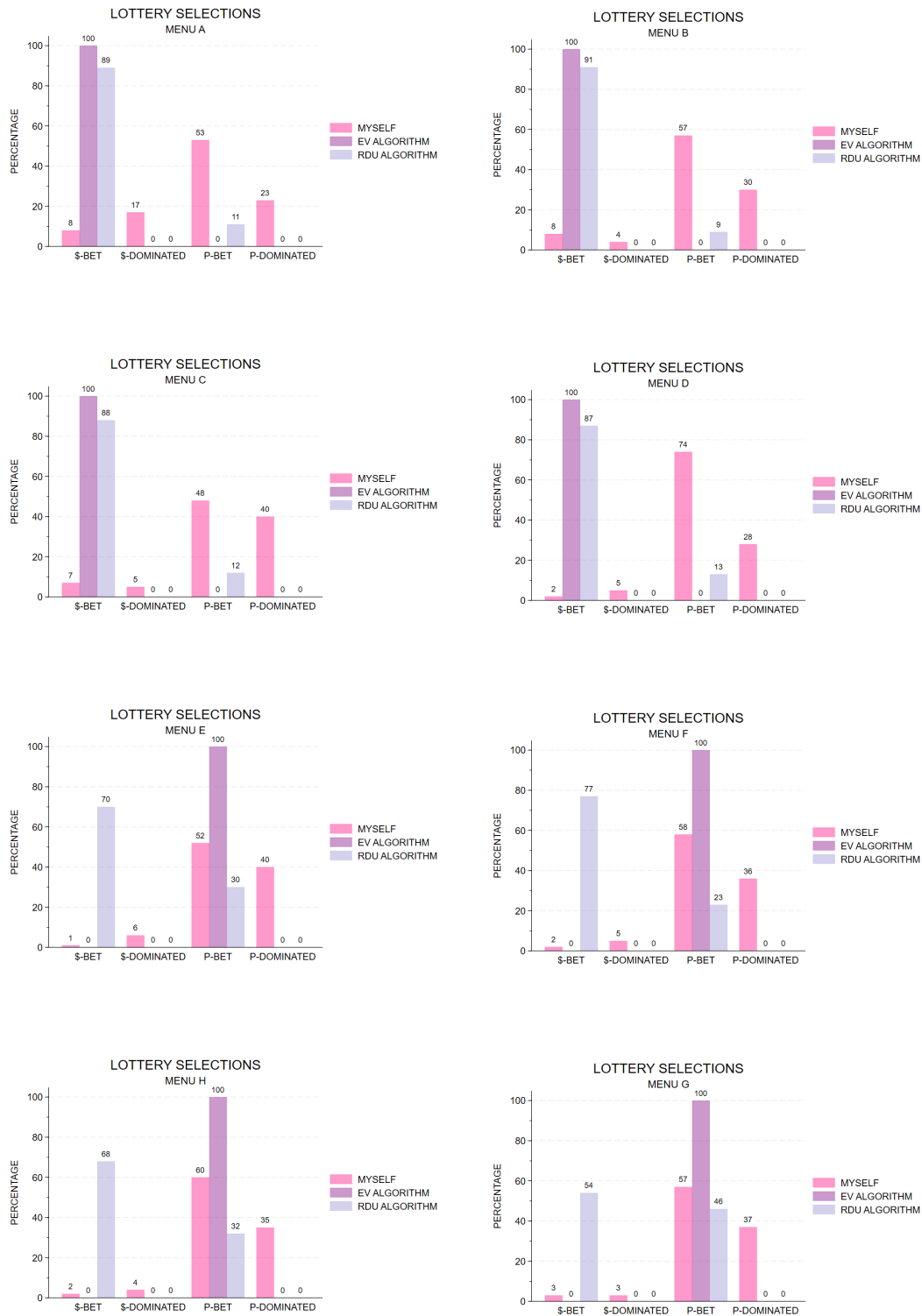


Figure 3.13: Lottery Type Selection Over Time



Note: Cf. Fig. 3.6: in MENUS A–D, the \$-Bet has a higher EV than the corresponding P-Bet; in MENUS E–H, the P-Bet has a higher EV than the corresponding \$-Bet.

Figure 3.14: Menu-by-Menu Lottery Type Selection

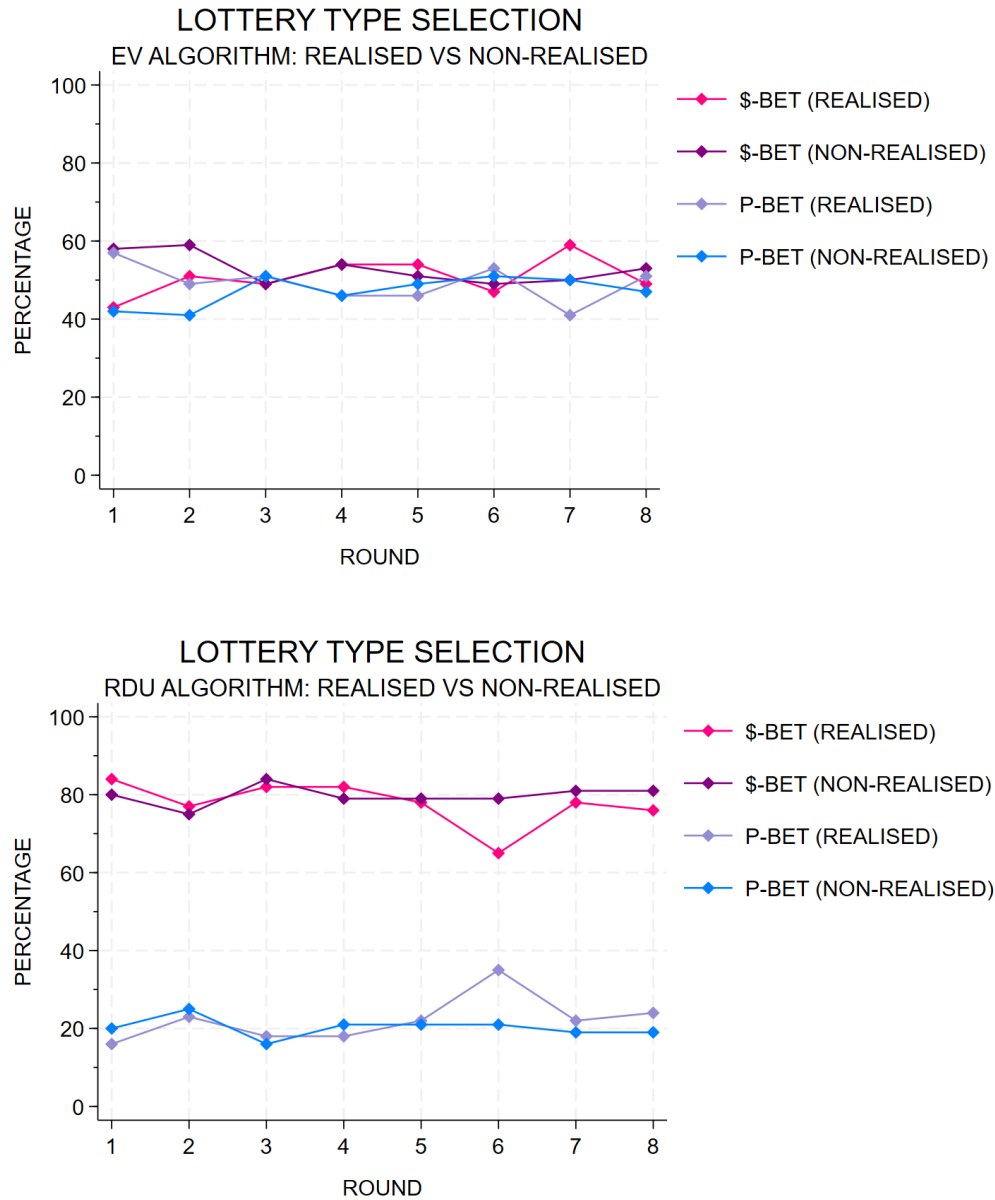


Figure 3.15: Lottery Type Selection (Realised vs. Non-Realised)

In either case, the consequences of reliance on the RDU algorithm are not congruent with the consequences of self-reliance with regard to lottery type selection, which we need to take into account when evaluating the significantly higher level of automation reliance among those who were offered the EV algorithm — especially considering that the frequent selection of \$-Bets by the RDU algorithm would also have led to more frequent outcomes of zero for subjects who received outcome feedback, which has been shown to be correlated with subsequent automation reliance.

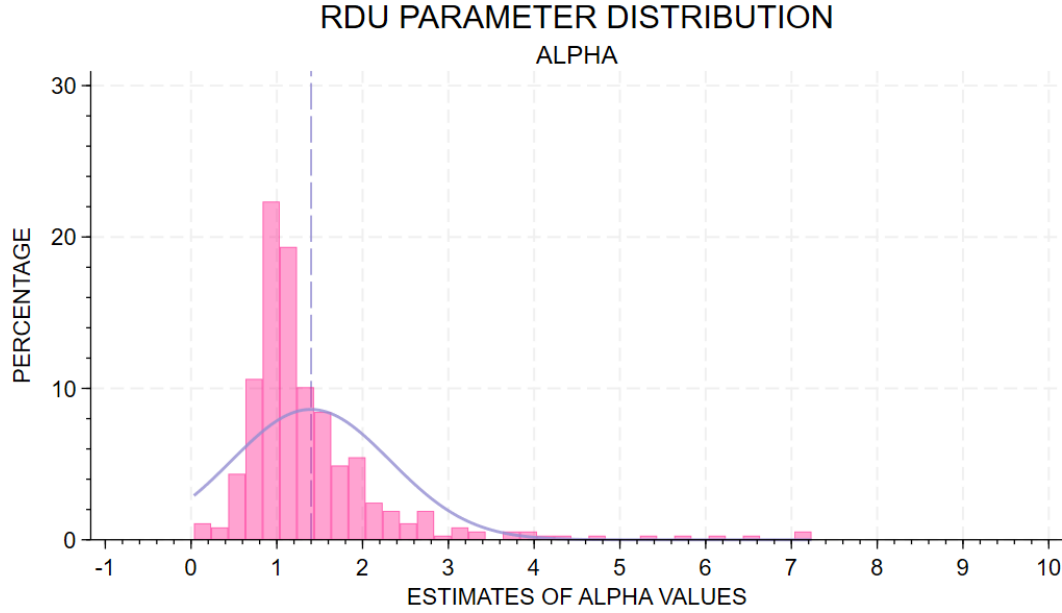
3.3 RDU Parameter Estimation

Finally, we therefore take a look at our RDU Parameter Estimation to investigate the underlying causes of the high proportion of \$-Bet selections by the RDU algorithm relative to both the EV algorithm and subjects who chose self-reliance. Fig. 3.16 – 3.18 illustrate the distributions of RDU parameter estimates α , β and γ , with the dashed vertical lines indicating their respective means¹¹. Firstly, we can reject the null hypotheses that $\alpha = 1$, $\beta = 1$, or $\gamma = 1$ ($p < 0.01$, Two-Sided One Sample t-Tests) and conclude that there is evidence of both non-linear probability weighting and non-linear valuations of consequences: α values are skewed towards the right, with a mean of 1.40 (std. err.: 0.05); the majority of β and γ estimates fall below 1, with mean values of 0.54 (std. err.: 0.03) and 0.55 (std. err.: 0.02), respectively.

Fig. 3.19 plots a probability weighting function based on the median estimates of β and γ ; this probability weighting function captures the central tendency of the distribution of individual probability weighting functions for subjects in all treatment groups. We see first of all that, on average, the RDU algorithm was programmed to overweight very small probabilities and underweight medium and large probabilities. Consequently, probabilities of 0.2 and 0.3 (the two probabilities with which the non-zero outcomes occurred in our \$-Bets) were, on average, slightly underweighted, but not as heavily as probabilities of 0.7 and 0.8 (the two probabilities with which the non-zero outcomes occurred in our P-Bets). In other words, while there is substantial heterogeneity in RDU parameter estimates, probability weighting functions likely often drove down the attractiveness of P-Bets relative to \$-Bets. In addition to this, the frequent convexity of the estimated utility functions as indicated by the α values in Fig. 3.16 led to an exponential valuation of consequences and, as such, drove up the attractiveness of \$-Bets relative to P-Bets. Combined with the fact that only half of the P-Bets had a higher expected value than the corresponding \$-Bets, this can explain the high number of \$-Bet choices by the RDU algorithm.

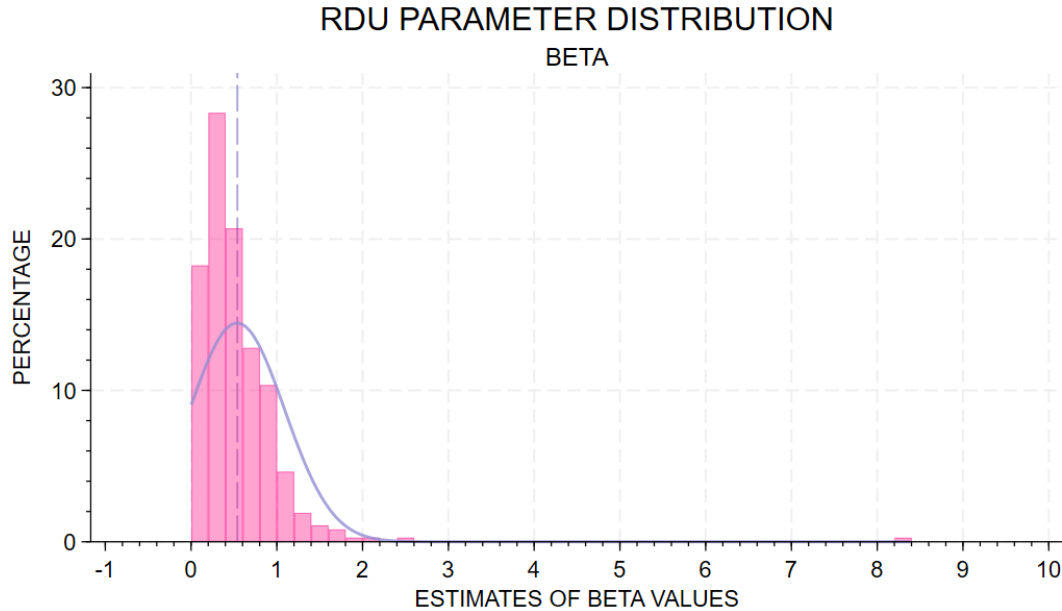
In conclusion, it appears that the discrepancy between lottery type selections resulting from self-reliance and those resulting from reliance on the RDU algorithm are the product of two effects: one, the complexity of our decision environment led subjects to disproportionately select P-Bets and their dominated variations, and two, the RDU parameter estimation yielded values of α , β and γ that resulted in our programming an RDU algorithm which often valued the \$-Bet higher than the P-Bet. In Section 4, we discuss the consequences of this for the interpretation of the observed treatment differences in automation reliance.

¹¹Our RDU parameter elicitation yielded no result for 2 subjects whose choices were indicative of extremely risk-loving behaviour. Here and previously, we have excluded those 2 observations as well as 8 outlier observations where at least one of the three estimates took on an extreme value of > 10 or < -10 ; this left us with 367 observations.



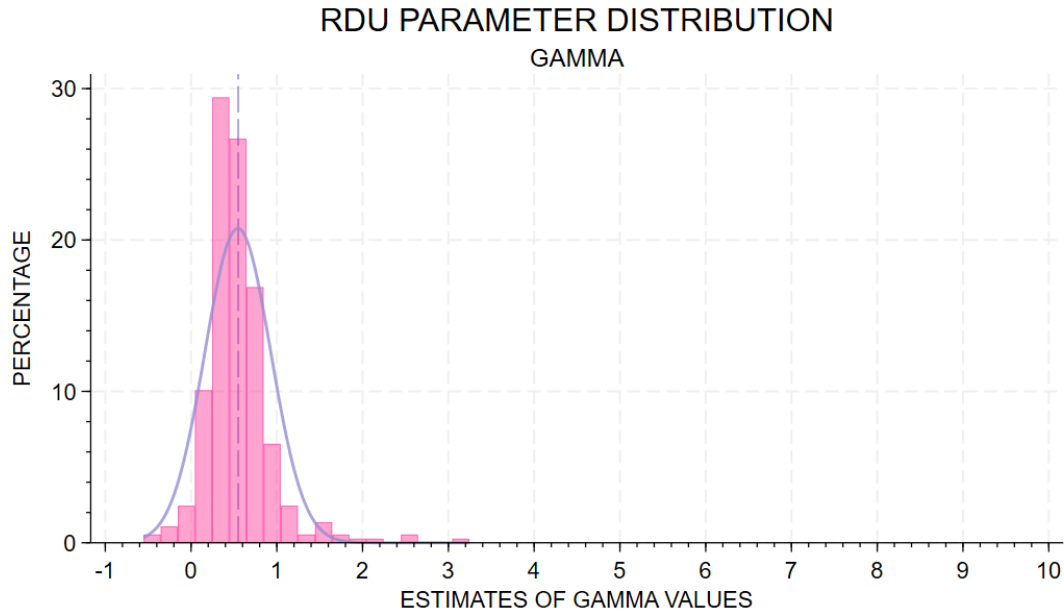
Note: Distribution of α estimates, dashed vertical line indicates the mean value of 1.40 (0.05). We excluded ten subjects for whom the RDU parameter elicitation yielded no result and 8 outliers where at least one estimate took on an extreme value of > 10 or < -10 .

Figure 3.16: RDU Parameter Estimate Distribution: α



Note: Distribution of β estimates, dashed vertical line indicates the mean value of 0.54 (0.03). We excluded two subjects for whom the RDU parameter elicitation yielded no result and 8 outliers where at least one estimate took on an extreme value of > 10 or < -10 .

Figure 3.17: RDU Parameter Estimate Distribution: β



Note: Distribution of γ estimates, dashed vertical line indicates the mean value of 0.55 (0.02). We excluded two subjects for whom the RDU parameter elicitation yielded no result and 8 outliers where at least one estimate took on an extreme value of > 10 or < -10 .

Figure 3.18: RDU Parameter Estimate Distribution: γ

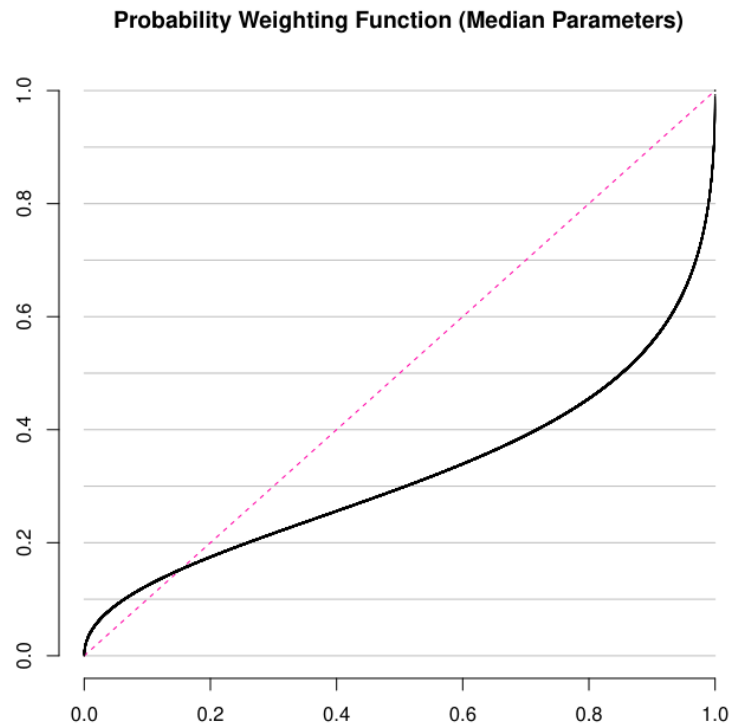


Figure 3.19: Probability Weighting Function

4 Discussion and Conclusion

Building on previous evidence of the sensitivity to algorithmic errors and a strong focus on environments in which such errors are objective, we conducted a two-part online experiment to investigate automation reliance in the preferential choice domain with and without outcome feedback. We measured individual risk preferences, which allowed us to program a preferential algorithm based on Rank Dependent Utility Theory as the counterpart to a normative algorithm based on Expected Value Theory; then, we asked subjects to complete 8 rounds of a lottery selection task on their own or by relying on a computer. Treatment groups varied (1) whether subjects were offered a computer using the normative algorithm or the preferential algorithm, and (2) whether or not they received outcome feedback regarding the selected lottery.

Our results indicate that, in our set-up, automation reliance is significantly and consistently lower for the RDU than for the EV algorithm. Secondly, the effect of outcome feedback is not as consistent nor as pronounced, but the rate of decline in treatment groups with outcome feedback is higher than that in treatment groups without outcome feedback. We furthermore see that satisfactory feedback reinforced a subject's previous decision to select the lottery themselves or have an algorithm select it for them, whereas unsatisfactory feedback led to changes from their previous decision — an effect which was significantly stronger for the preferential than for the normative algorithm. In addition to this, we find evidence that, regardless of the type of algorithm, automation reliance generally leads to lower satisfaction ratings, suggesting that algorithmic lottery selections are judged more negatively than those made by the subjects themselves.

For the evaluation and interpretation of the observed treatment differences, we first need to discuss the incongruence of human and algorithmic lottery type selections in our study. Self-reliance resulted in a significantly lower number of \$-Bet selections than automation reliance due to two different effects. Firstly, the complexity of the lottery selection task appears to have pushed subjects towards selecting P-Bets and their dominated variations much more frequently than \$-Bets and their dominated variations. We believe that this might have been the case because subjects tried to simplify the lottery selection task by focusing on winning *probabilities* rather than *amounts*, or because identifying dominated lotteries was easier for \$-Bets than it was for P-Bets due to the higher number of zero outcomes in \$-Bets. Secondly, RDU parameter estimates, on average, valued the \$-Bet more highly than the P-Bet — in many cases even when the P-Bet had a much higher expected value.

Crucially, this renders the discrepancy in the number of \$-Bet selections resulting from self-reliance versus automation reliance much more pronounced for the RDU algorithm than it is for the EV algorithm. As a consequence, the disproportionate selection of \$-Bets by the RDU algorithm could have triggered a faster decline of

automation reliance than the EV algorithm for two reasons: one, the higher likelihood of zero outcomes in \$-Bets which amplified the negative effect of unsatisfactory outcome feedback on subsequent automation reliance, and two, the incongruence with a subject's own lottery type selections which dissolved trust in the ability of the RDU algorithm to deliver on its promise of capturing a subject's risk preferences and predicting their lottery selections.

However, it is important to note that the comparative preference for the EV algorithm does not appear to have been *caused* by selection behaviour alone; in fact, the level of automation reliance is significantly higher for the EV algorithm even in the very first round of the lottery selection task. Consequently, subjects either anticipated the failure of the RDU algorithm to deliver on its promise of capturing their risk preferences and predicting their lottery selections, or they simply considered the EV algorithm a more useful tool than the RDU algorithm. We cannot say with certainty whether or how well our two types of algorithm aligned with the preferences of our human subjects, but a failure to deliver on its promise was more easily identifiable in case of the preferential algorithm: subjects were unable to contest the mathematical performance of the normative algorithm, but they would — in some cases perhaps mistakenly — have felt able to judge how well the selection behaviour of the preferential algorithm matched their own preferences. Similarly, subjects might have considered a normative algorithm more helpful than a preferential algorithm that tried to replicate their preferences if they were unaware of how a complex decision environment like the one in our study can affect their own ability to act in accordance with their true preferences.

In conclusion, in the interest of counteracting suboptimal levels automation reliance, the design of an algorithm — especially but not exclusively in the preferential choice domain — will likely benefit from careful considerations regarding ex-ante as well as ex-post beliefs about what a useful algorithm should promise, whether or not the algorithm can deliver on its promise, and how easily it is from a user perspective to identify its failure to do so.

References

- Abeliuk, A., Benjamin, D. M., Morstatter, F., & Galstyan, A. (2020). Quantifying Machine Influence Over Human Forecasters. *Scientific Reports*, 10(1), 1–14.
- Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C., & Sugden, R. (2009). *Experimental Economics: Rethinking the Rules*. Princeton University Press.
- Baron, J. & Hershey, J. C. (1988). Outcome Bias in Decision Evaluation. *Journal of Personality and Social Psychology*, 54(4), 569.
- Bartling, B., Fehr, E., & Herz, H. (2014). The Intrinsic Value of Decision Rights. *Econometrica*, 82(6), 2005–2039.
- Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The Social Dilemma of Autonomous Vehicles. *Science*, 352(6293), 1573–1576.
- Brownback, A. & Kuhn, M. A. (2019). Understanding Outcome Bias. *Games and Economic Behavior*, 117, 342–360.
- Bruhin, A., Fehr-Duda, H., & Epper, T. (2010). Risk and Rationality: Uncovering Heterogeneity in Probability Distortion. *Econometrica*, 78(4), 1375–1412.
- Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A Systematic Review of Algorithm Aversion in Augmented Decision Making. *Journal of Behavioral Decision Making*, 33(2), 220–239.
- Cubitt, R. P., Munro, A., & Starmer, C. (2004). Testing Explanations of Preference Reversal. *The Economic Journal*, 114(497), 709–726.
- Cubitt, R. P., Navarro-Martinez, D., & Starmer, C. (2015). On Preference Imprecision. *Journal of Risk and Uncertainty*, 50, 1–34.
- Cubitt, R. P., Starmer, C., & Sugden, R. (1998). On the Validity of the Random Lottery Incentive System. *Experimental Economics*, 1, 115–131.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General*, 144(1), 114.
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The Role of Trust in Automation Reliance. *International Journal of Human-Computer Studies*, 58(6), 697–718.
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors*, 44(1), 79–94.

- Epper, T., Fehr-Duda, H., & Bruhin, A. (2011). Viewing the Future Through a Warped Lens: Why Uncertainty Generates Hyperbolic Discounting. *Journal of Risk and Uncertainty*, 43, 169–203.
- Fehr, E., Herz, H., & Wilkening, T. (2013). The Lure of Authority: Motivation and Incentive Effects of Power. *American Economic Review*, 103(4), 1325–59.
- Fehr-Duda, H., Bruhin, A., Epper, T., & Schubert, R. (2010). Rationality on the Rise: Why Relative Risk Aversion Increases with Stake Size. *Journal of Risk and Uncertainty*, 40, 147–180.
- Fehr-Duda, H., De Gennaro, M., & Schubert, R. (2006). Gender, Financial Risk, and Probability Weights. *Theory and Decision*, 60, 283–313.
- Gillenkirch, R. M. & Velthuis, L. (2023). Delegated Risk-Taking, Accountability, and Outcome Bias. *Journal of Risk and Uncertainty*, 67(2), 137–161.
- Gogoll, J. & Uhl, M. (2018). Rage Against the Machine: Automation in the Moral Domain. *Journal of Behavioral and Experimental Economics*, 74, 97–103.
- Goldstein, W. M. & Einhorn, H. J. (1987). Expression Theory and the Preference Reversal Phenomena. *Psychological Review*, 94(2), 236.
- Grether, D. M. & Plott, C. R. (1979). Economic Theory of Choice and the Preference Reversal Phenomenon. *The American Economic Review*, 69(4), 623–638.
- Gurdal, M. Y., Miller, J. B., & Rustichini, A. (2013). Why Blame? *Journal of Political Economy*, 121(6), 1205–1247.
- Highhouse, S. (2008). Stubborn Reliance on Intuition and Subjectivity in Employee Selection. *Industrial and Organizational Psychology*, 1(3), 333–342.
- Hoff, K. A. & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors that Influence Trust. *Human Factors*, 57(3), 407–434.
- König-Kersting, C., Pollmann, M., Potters, J., & Trautmann, S. T. (2021). Good Decision vs. Good Results: Outcome Bias in the Evaluation of Financial Agents. *Theory and Decision*, 90, 31–61.
- Lichtenstein, S. & Slovic, P. (1971). Reversals of Preference Between Bids and Choices in Gambling Decisions. *Journal of Experimental Psychology*, 89(1), 46.
- Liehner, G. L., Brauner, P., Schaar, A. K., & Ziefle, M. (2021). Delegation of Moral Tasks to Automated Agents — The Impact of Risk and Context on Trusting a Machine to Perform a Task. *IEEE Transactions on Technology and Society*, 3(1), 46–57.
- Loomes, G. & Pogrebná, G. (2014). Measuring Individual Risk Attitudes When Preferences Are Imprecise. *The Economic Journal*, 124(576), 569–593.

- Loomes, G., Starmer, C., & Sugden, R. (1989). Preference reversal: information-processing effect or rational non-transitive choice? *The Economic Journal*, 99(395), 140–151.
- Madhavan, P. & Wiegmann, D. A. (2007). Similarities and Differences Between Human-Human and Human-Automation Trust: An Integrative Review. *Theoretical Issues in Ergonomics Science*, 8(4), 277–301.
- Murad, Z., Sefton, M., & Starmer, C. (2016). How Do Risk Attitudes Affect Measured Confidence? *Journal of Risk and Uncertainty*, 52, 21–46.
- Neri, C. & Rommeswinkel, H. (2017). Decision Rights: Freedom, Power, and Interference. Available at SSRN: <https://ssrn.com/abstract=2485107> or <http://dx.doi.org/10.2139/ssrn.2485107>.
- Owens, D., Grossman, Z., & Fackler, R. (2014). The Control Premium: A Preference for Payoff Autonomy. *American Economic Journal: Microeconomics*, 6(4), 138–61.
- Palan, S. & Schitter, C. (2018). Prolific.ac — A subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance*, 17, 22–27.
- Prahl, A. & Van Swol, L. (2017). Understanding Algorithm Aversion: When Is Advice from Automation Discounted? *Journal of Forecasting*, 36(6), 691–702.
- Schniter, E., Shields, T. W., & Sznycer, D. (2020). Trust in Humans and Robots: Economically Similar But Emotionally Different. *Journal of Economic Psychology*, 78, 102253.
- Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2017). Psychological Roadblocks to the Adoption of Self-Driving Vehicles. *Nature Human Behaviour*, 1(10), 694–696.
- Stott, H. P. (2006). Cumulative Prospect Theory’s Functional Menagerie. *Journal of Risk and Uncertainty*, 32, 101–130.
- Tversky, A., Slovic, P., & Kahneman, D. (1990). The Causes of Preference Reversal. *The American Economic Review*, (pp. 204–217).
- Wu, G., Zhang, J., & Gonzales, R. (2004). Decision Under Risk. In *The Blackwell Handbook of Judgment and Decision Making*, ed. by D. Koehler and N. Harvey. Oxford: Oxford University Press, (pp. 399–423).

Appendices

A Appendix for Chapter 1

A.1 Additional Analysis

A.1.1 Task Evaluation Survey

The Task Evaluation Survey was programmed using Qualtrics software and conducted via Prolific in November, 2020. We recruited a total of 200 subjects, who received a participation fee of £0.90 upon completion. Table A.1 summarises the demographic characteristics of the sample.

	TOTAL	PERCENTAGE
GENDER		
FEMALE	129	64.50
MALE	70	35.00
OTHER	1	0.50
AGE		
MIN	18	
MAX	75	
MEAN	34.70	
NATIONALITY		
UK	181	90.50
OTHER	19	9.50
HIGHEST LEVEL OF EDUCATION COMPLETED		
NO FORMAL QUALIFICATIONS	2	1.00
SECONDARY EDUCATION (E.G. GED/GCSE)	20	10.00
HIGH SCHOOL DIPLOMA/A-LEVELS	53	26.50
TECHNICAL/COMMUNITY COLLEGE	23	11.50
UNDERGRADUATE DEGREE (BA/BSc/OTHER)	68	34.00
GRADUATE DEGREE (MA/MSc/MPhil/OTHER)	15.50	19.05
DOCTORATE DEGREE (PhD/OTHER)	5	2.50
EMPLOYMENT STATUS		
EMPLOYED	119	59.50
SELF-EMPLOYED	24	12.00
UNEMPLOYED	19	9.50
HOMEMAKER	15	7.50
RETIRED	3	1.50
STUDENT	20	10.00

Table A.1: Demographic Sample Composition (Task Evaluation Survey)

To identify the HF Task and CF Task for the experiment, we asked respondents to evaluate five task candidates with respect to four statements¹² via 5-point Likert scales: a score of 1 corresponded to ‘strongly disagree’, a score of 5 corresponded to ‘strongly agree’. Overall, agreement to both COMPUTER and COMPUTER > HUMAN indicate the computer-friendliness of a task candidate, while agreement to HUMAN and disagreement to COMPUTER > HUMAN indicate its human-friendliness. Agreement to PROGRAMMING can also be viewed as an indication of computer-friendliness, albeit a less unambiguous one.

Table A.2 reports correlation coefficients between responses to the four statements for all five task candidates. In line with the above interpretations of the statements, there is a significant positive correlation between COMPUTER and COMPUTER > HUMAN, a significant negative correlation between HUMAN and COMPUTER > HUMAN, as well as a significant positive correlation between the difference of COMPUTER & HUMAN (C–H) and COMPUTER > HUMAN; this suggests that, on average, respondents understood the evaluation task and provided ‘sensible’ response patterns.

	COMP & C>H	HUM & C>H	C–H & C>H
MEMORY TASK	0.5651***	−0.2026***	0.3945***
PREDICTION TASK	0.6845***	−0.2675***	0.7124***
COORDINATION TASK	0.8312***	−0.1609**	0.7486***
CAPTCHA TASK	0.7810***	−0.4977***	0.7852***
ANAGRAM TASK	0.6460***	−0.2601***	0.6788***

Note: * p<0.1, ** p<0.05, *** p<0.01.

Table A.2: Spearman’s Correlations

Detailed summary statistics of the evaluation of all five task candidates can be found in Tables A.3 to A.7. The *Memory Task* stands out as having the (joint) highest mean and median scores for COMPUTER, COMPUTER > HUMAN and PROGRAMMING as well as the (joint) lowest mean and median scores for HUMAN, making it an ideal candidate for the CF Task. The *Coordination Task* has the (joint) lowest mean and median scores for COMPUTER, COMPUTER > HUMAN and PROGRAMMING, but it also has the (joint) lowest scores for HUMAN; nevertheless, if we define human-friendliness in this context as the absence of computer-friendliness, the *Coordination Task* appears to be the best contender for the HF Task.

¹²Hereafter, COMPUTER refers to “I would expect a computer to perform this task well.”; HUMAN refers to “I would expect a human to perform this task well.”; COMPUTER > HUMAN refers to “I would expect a computer to perform this task better than a human.”; PROGRAMMING refers to “I think a computer could be programmed to perform this task well.”

	TOTAL	PERCENTAGE
COMPUTER		
STRONGLY DISAGREE (1)	0	0.00
DISAGREE (2)	5	2.50
NEITHER AGREE NOR DISAGREE (3)	8	4.00
AGREE (4)	43	21.50
STRONGLY AGREE (5)	144	72.00
MEAN	4.63 (0.68)	
MEDIAN	5	
HUMAN		
STRONGLY DISAGREE (1)	12	6.00
DISAGREE (2)	73	36.50
NEITHER AGREE NOR DISAGREE (3)	43	21.50
AGREE (4)	62	31.00
STRONGLY AGREE (5)	10	5.00
MEAN	2.93 (1.06)	
MEDIAN	3	
COMPUTER > HUMAN		
STRONGLY DISAGREE (1)	2	1.00
DISAGREE (2)	5	2.50
NEITHER AGREE NOR DISAGREE (3)	14	7.00
AGREE (4)	53	26.50
STRONGLY AGREE (5)	126	63.00
MEAN	4.48 (0.81)	
MEDIAN	5	
PROGRAMMING		
STRONGLY DISAGREE (1)	2	1.00
DISAGREE (2)	2	1.00
NEITHER AGREE NOR DISAGREE (3)	5	2.50
AGREE (4)	56	28.00
STRONGLY AGREE (5)	135	67.50
MEAN	4.60 (0.69)	
MEDIAN	5	

Note: COMPUTER refers to “I would expect a computer to perform this task well.”; HUMAN refers to “I would expect a human to perform this task well.”; COMPUTER > HUMAN refers to “I would expect a computer to perform this task better than a human.”; PROGRAMMING refers to “I think a computer could be programmed to perform this task well.”. Standard errors of mean scores in parentheses.

Table A.3: Memory Task Evaluation

	TOTAL	PERCENTAGE
COMPUTER		
STRONGLY DISAGREE (1)	2	1.00
DISAGREE (2)	15	7.50
NEITHER AGREE NOR DISAGREE (3)	50	25.00
AGREE (4)	78	39.00
STRONGLY AGREE (5)	55	27.50
MEAN	3.85 (0.95)	
MEDIAN	4	
HUMAN		
STRONGLY DISAGREE (1)	1	0.50
DISAGREE (2)	25	12.50
NEITHER AGREE NOR DISAGREE (3)	67	33.50
AGREE (4)	91	45.50
STRONGLY AGREE (5)	16	8.00
MEAN	3.48 (0.83)	
MEDIAN	4	
COMPUTER > HUMAN		
STRONGLY DISAGREE (1)	7	3.50
DISAGREE (2)	42	21.00
NEITHER AGREE NOR DISAGREE (3)	49	24.50
AGREE (4)	56	28.00
STRONGLY AGREE (5)	46	23.00
MEAN	3.46 (1.16)	
MEDIAN	4	
PROGRAMMING		
STRONGLY DISAGREE (1)	4	2.00
DISAGREE (2)	20	10.00
NEITHER AGREE NOR DISAGREE (3)	37	18.50
AGREE (4)	85	42.50
STRONGLY AGREE (5)	54	27.00
MEAN	3.83 (1.00)	
MEDIAN	4	

Note: COMPUTER refers to “I would expect a computer to perform this task well.”; HUMAN refers to “I would expect a human to perform this task well.”; COMPUTER > HUMAN refers to “I would expect a computer to perform this task better than a human.”; PROGRAMMING refers to “I think a computer could be programmed to perform this task well.”. Standard errors of mean scores in parentheses.

Table A.4: Prediction Task Evaluation

	TOTAL	PERCENTAGE
COMPUTER		
STRONGLY DISAGREE (1)	12	6.00
DISAGREE (2)	37	18.50
NEITHER AGREE NOR DISAGREE (3)	56	28.00
AGREE (4)	59	29.50
STRONGLY AGREE (5)	36	18.00
MEAN	3.35 (1.15)	
MEDIAN	3	
HUMAN		
STRONGLY DISAGREE (1)	9	4.50
DISAGREE (2)	60	30.00
NEITHER AGREE NOR DISAGREE (3)	63	31.50
AGREE (4)	65	32.50
STRONGLY AGREE (5)	3	1.50
MEAN	2.97 (0.93)	
MEDIAN	3	
COMPUTER > HUMAN		
STRONGLY DISAGREE (1)	6	3.00
DISAGREE (2)	59	29.50
NEITHER AGREE NOR DISAGREE (3)	41	20.50
AGREE (4)	57	28.50
STRONGLY AGREE (5)	37	18.50
MEAN	3.30 (1.16)	
MEDIAN	3	
PROGRAMMING		
STRONGLY DISAGREE (1)	8	4.00
DISAGREE (2)	33	16.50
NEITHER AGREE NOR DISAGREE (3)	38	19.00
AGREE (4)	72	36.00
STRONGLY AGREE (5)	49	24.50
MEAN	3.61 (1.14)	
MEDIAN	4	

Note: COMPUTER refers to “I would expect a computer to perform this task well.”; HUMAN refers to “I would expect a human to perform this task well.”; COMPUTER > HUMAN refers to “I would expect a computer to perform this task better than a human.”; PROGRAMMING refers to “I think a computer could be programmed to perform this task well.”. Standard errors of mean scores in parentheses.

Table A.5: Coordination Task Evaluation

	TOTAL	PERCENTAGE
COMPUTER		
STRONGLY DISAGREE (1)	8	4.00
DISAGREE (2)	32	16.00
NEITHER AGREE NOR DISAGREE (3)	17	8.50
AGREE (4)	66	33.00
STRONGLY AGREE (5)	77	38.50
MEAN	3.86 (1.21)	
MEDIAN	4	
HUMAN		
STRONGLY DISAGREE (1)	16	8.00
DISAGREE (2)	41	20.50
NEITHER AGREE NOR DISAGREE (3)	26	13.00
AGREE (4)	97	48.50
STRONGLY AGREE (5)	20	10.00
MEAN	3.32 (1.15)	
MEDIAN	4	
COMPUTER > HUMAN		
STRONGLY DISAGREE (1)	14	7.00
DISAGREE (2)	45	22.50
NEITHER AGREE NOR DISAGREE (3)	17	8.50
AGREE (4)	42	21.00
STRONGLY AGREE (5)	82	41.00
MEAN	3.67 (1.39)	
MEDIAN	4	
PROGRAMMING		
STRONGLY DISAGREE (1)	3	1.50
DISAGREE (2)	18	9.00
NEITHER AGREE NOR DISAGREE (3)	22	11.00
AGREE (4)	72	36.00
STRONGLY AGREE (5)	85	42.50
MEAN	4.09 (1.01)	
MEDIAN	4	

Note: COMPUTER refers to “I would expect a computer to perform this task well.”; HUMAN refers to “I would expect a human to perform this task well.”; COMPUTER > HUMAN refers to “I would expect a computer to perform this task better than a human.”; PROGRAMMING refers to “I think a computer could be programmed to perform this task well.”. Standard errors of mean scores in parentheses.

Table A.6: CAPTCHA Task Evaluation

	TOTAL	PERCENTAGE
COMPUTER		
STRONGLY DISAGREE (1)	1	0.50
DISAGREE (2)	12	6.00
NEITHER AGREE NOR DISAGREE (3)	18	9.00
AGREE (4)	79	39.50
STRONGLY AGREE (5)	90	45.00
MEAN	4.23 (0.88)	
MEDIAN	4	
HUMAN		
STRONGLY DISAGREE (1)	2	1.00
DISAGREE (2)	18	9.00
NEITHER AGREE NOR DISAGREE (3)	24	12.00
AGREE (4)	128	64.00
STRONGLY AGREE (5)	28	14.00
MEAN	3.81 (0.82)	
MEDIAN	4	
COMPUTER > HUMAN		
STRONGLY DISAGREE (1)	0	0.00
DISAGREE (2)	31	15.50
NEITHER AGREE NOR DISAGREE (3)	40	20.00
AGREE (4)	64	32.00
STRONGLY AGREE (5)	65	32.50
MEAN	3.82 (1.06)	
MEDIAN	4	
PROGRAMMING		
STRONGLY DISAGREE (1)	1	0.50
DISAGREE (2)	10	5.00
NEITHER AGREE NOR DISAGREE (3)	8	4.00
AGREE (4)	84	42.00
STRONGLY AGREE (5)	97	48.50
MEAN	4.33 (0.82)	
MEDIAN	4	

Note: COMPUTER refers to “I would expect a computer to perform this task well.”; HUMAN refers to “I would expect a human to perform this task well.”; COMPUTER > HUMAN refers to “I would expect a computer to perform this task better than a human.”; PROGRAMMING refers to “I think a computer could be programmed to perform this task well.”. Standard errors of mean scores in parentheses.

Table A.7: Anagram Task Evaluation

Tables A.8 – A.11 show that COMPUTER, COMPUTER > HUMAN and PROGRAMMING scores for the *Memory* and *Coordination Task* significantly differ. While this is not the case for HUMAN, the *Coordination Task* still exhibits significantly lower scores for COMPUTER, COMPUTER > HUMAN and PROGRAMMING than all other task candidates.

	MEMORY	PREDICTION	COORDINATION	CAPTCHA	ANAGRAM
MEMORY	-	0.0000	0.0000	0.0000	0.0000
PREDICTION	0.0000	-	0.0000	0.7131	0.0000
COORDINATION	0.0000	0.0000	-	0.0000	0.0000
CAPTCHA	0.0000	0.7131	0.0000	-	0.0007
ANAGRAM	0.0000	0.0000	0.0000	0.0007	-

Note: Wilcoxon Signed-Rank p -values for tests of between-task equality of evaluation scores of the statement, “I would expect a computer to perform this task well.”

Table A.8: Wilcoxon Signed-Rank p -values for COMPUTER

	MEMORY	PREDICTION	COORDINATION	CAPTCHA	ANAGRAM
MEMORY	-	0.0000	0.5958	0.0002	0.0000
PREDICTION	0.0000	-	0.0000	0.2008	0.0000
COORDINATION	0.5958	0.0000	-	0.0007	0.0000
CAPTCHA	0.0002	0.2008	0.0007	-	0.0000
ANAGRAM	0.0000	0.0000	0.0000	0.0000	-

Note: Wilcoxon Signed-Rank p -values for tests of between-task equality of evaluation scores of the statement, “I would expect a human to perform this task well.”

Table A.9: Wilcoxon Signed-Rank p -values for HUMAN

	MEMORY	PREDICTION	COORDINATION	CAPTCHA	ANAGRAM
MEMORY	-	0.0000	0.0000	0.0000	0.0000
PREDICTION	0.0000	-	0.0942	0.0955	0.0029
COORDINATION	0.0000	0.0942	-	0.0012	0.0000
CAPTCHA	0.0000	0.0955	0.0012	-	0.3098
ANAGRAM	0.0000	0.0029	0.0000	0.3098	-

Note: Wilcoxon Signed-Rank p -values for tests of between-task equality of evaluation scores of the statement, “I would expect a computer to perform this task better than a human.”

Table A.10: Wilcoxon Signed-Rank p -values for COMPUTER > HUMAN

	MEMORY	PREDICTION	COORDINATION	CAPTCHA	ANAGRAM
MEMORY	-	0.0000	0.0000	0.0000	0.0000
PREDICTION	0.0000	-	0.0148	0.0092	0.0000
COORDINATION	0.0000	0.0148	-	0.0000	0.0000
CAPTCHA	0.0000	0.0092	0.0000	-	0.0028
ANAGRAM	0.0000	0.0000	0.0000	0.0028	-

Note: Wilcoxon Signed-Rank p -values for tests of between-task equality of evaluation scores of the statement, “I think a computer could be programmed to perform this task well.”

Table A.11: Wilcoxon Signed-Rank p -values for PROGRAMMING

A.1.2 Main Experiment

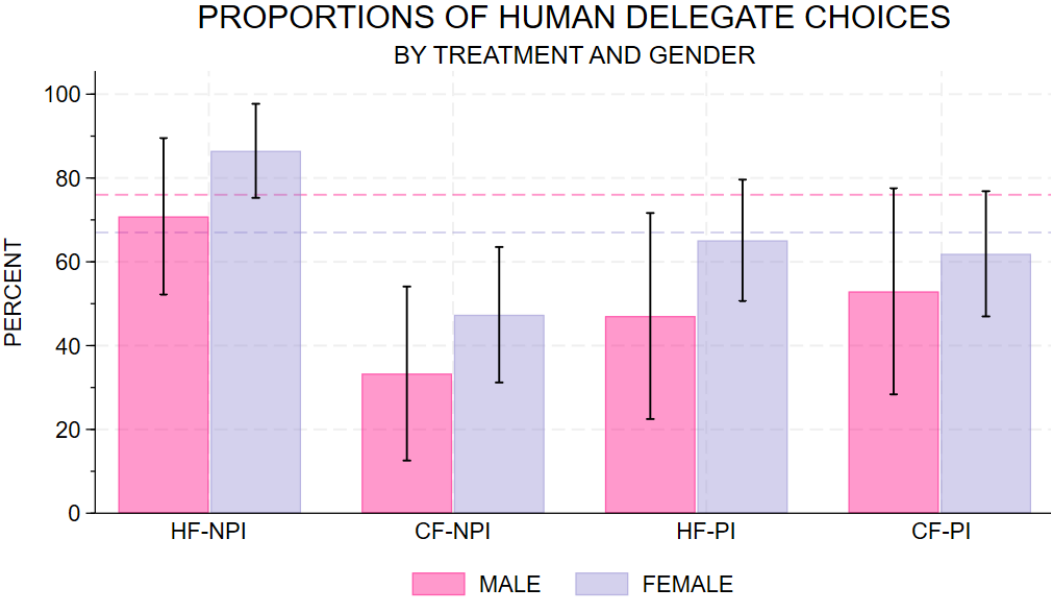
	TOTAL	PERCENTAGE
GENDER		
FEMALE	234	65.00
MALE	121	33.61
OTHER	3	0.83
PREFER NOT TO SAY	2	0.56
AGE		
MIN	18	
MAX	74	
MEAN	33.19	
NATIONALITY		
UK	339	94.17
OTHER	21	5.83
HIGHEST LEVEL OF EDUCATION COMPLETED		
NO FORMAL QUALIFICATIONS	2	0.56
SECONDARY EDUCATION (E.G. GED/GCSE)	29	8.06
HIGH SCHOOL DIPLOMA/A-LEVELS	87	24.17
TECHNICAL/COMMUNITY COLLEGE	28	7.78
UNDERGRADUATE DEGREE (BA/BSc/OTHER)	157	43.61
GRADUATE DEGREE (MA/MSc/MPhil/OTHER)	54	15.00
DOCTORATE DEGREE (PhD/OTHER)	3	0.83
EMPLOYMENT STATUS		
EMPLOYED	199	55.28
SELF-EMPLOYED	33	9.17
UNEMPLOYED	35	9.72
HOMEMAKER	15	4.17
RETIRED	10	2.78
STUDENT	68	18.89

Table A.12: Demographic Sample Composition (Main Experiment)

	BASELINE N=72	HF-NPI N=74	CF-NPI N=72	HF-PI N=70	CF-PI N=72
FEMALE SUBJECTS	72.22	56.76	63.89	67.14	65.28
MEAN AGE	31.28	34.73	31.67	35.63	32.67
UK NATIONALS	94.44	95.95	94.44	97.14	88.89
UNIVERSITY GRADUATES	63.89	54.05	56.94	64.29	58.33
STUDENTS	26.39	18.92	12.50	18.57	18.06

Note: Except for MEAN AGE, the reported values are within-treatment percentages.

Table A.13: Demographic Treatment Composition



Note: Whiskers indicate 95% confidence intervals; dashed lines indicate the BASELINE benchmarks.

Figure A.1: Gender Decomposition of Proportions of Human Delegate Choices

	μ_1	μ_2	p (H_0 : EQUAL PERF) <i>Mann-Whitney U Test</i>
COMPUTER DELEGATE — HUMAN DELEGATE			
HF TASK	2.02	1.88	0.4713
CF TASK	4.14	3.29	0.0000
COMPUTER DELEGATE — HUMAN NON-DELEGATE			
HF TASK	2.02	1.58	0.0000
CF TASK	4.14	3.38	0.0000
HUMAN DELEGATE — HUMAN NON-DELEGATE			
HF TASK	1.88	1.58	0.0448
CF TASK	3.29	3.38	0.6401

Note: We compare the performance of 300 computer delegates, 148 human delegates, and 152 human non-delegates; μ_1 denotes the mean number of correct answers for the first type of delegate in a given section, and μ_2 denotes the mean number of correct answers for the second type of delegate in a given section.

Table A.14: Comparisons of Performance

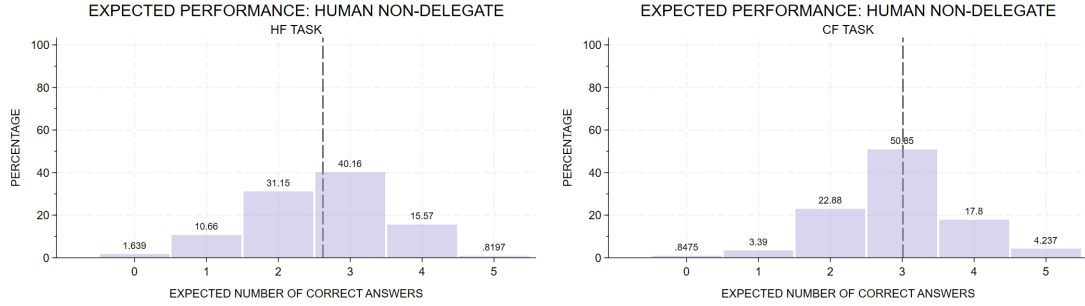
	μ_1	μ_2	p (H_0 : EQUAL EXP) <i>Wilcoxon Signed-Rank Test</i>
COMPUTER DELEGATE — HUMAN DELEGATE			
NO INFO (BASELINE)	3.73	2.92	0.0000
HF TASK	3.42	2.58	0.0000
HF-NPI	3.44	2.62	0.0000
HF-PI	3.40	2.53	0.0000
CF TASK	3.89	3.02	0.0000
CF-NPI	4.02	3.05	0.0000
CF-PI	3.77	2.98	0.0000
COMPUTER DELEGATE — HUMAN NON-DELEGATE			
NO INFO (BASELINE)	3.73	2.73	0.0000
HF TASK	3.42	2.62	0.0000
HF-NPI	3.44	2.66	0.0000
HF-PI	3.40	2.58	0.0000
CF TASK	3.89	3.01	0.0000
CF-NPI	4.02	3.03	0.0000
CF-PI	3.77	2.98	0.0000
HUMAN DELEGATE — HUMAN NON-DELEGATE			
NO INFO (BASELINE)	2.92	2.73	0.2265
HF TASK	2.58	2.62	0.6396
HF-NPI	2.62	2.66	0.9206
HF-PI	2.53	2.58	0.4370
CF TASK	3.02	3.01	0.8498
CF-NPI	3.05	3.03	0.9139
CF-PI	2.98	2.98	0.8722

Note: We have 300 performance expectations for each type of delegate; μ_1 denotes the mean number of expected correct answers for the first type of delegate in a given section, and μ_2 denotes the mean number of expected correct answers for the second type of delegate in a given section.

Table A.15: Comparisons of Performance Expectations

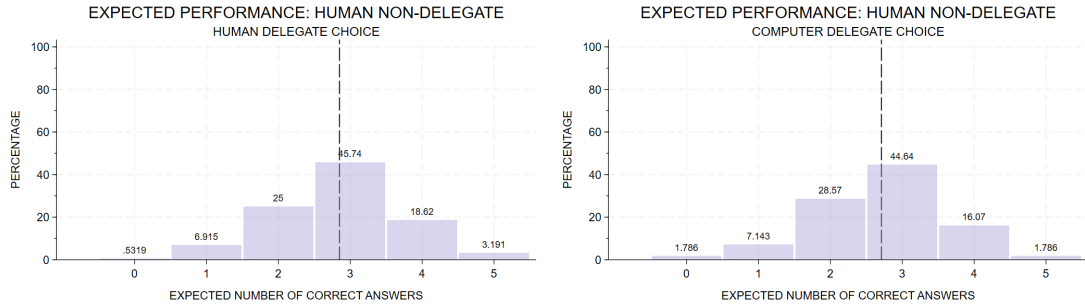
	μ_1	μ_2	p (H_0 : EQUAL EXP) <i>Mann-Whitney U Test</i>
HUMAN DELEGATE CHOICE — COMPUTER DELEGATE CHOICE			
COMPUTER — HUMAN DELEGATE	0.73	1.11	0.0030
COMPUTER — HUMAN NON-DELEGATE	0.73	1.05	0.0014
HUMAN DELEGATE — NON-DELEGATE	0.01	0.05	0.4937

Table A.16: Differences in Performance Expectations (By Delegate Choice)



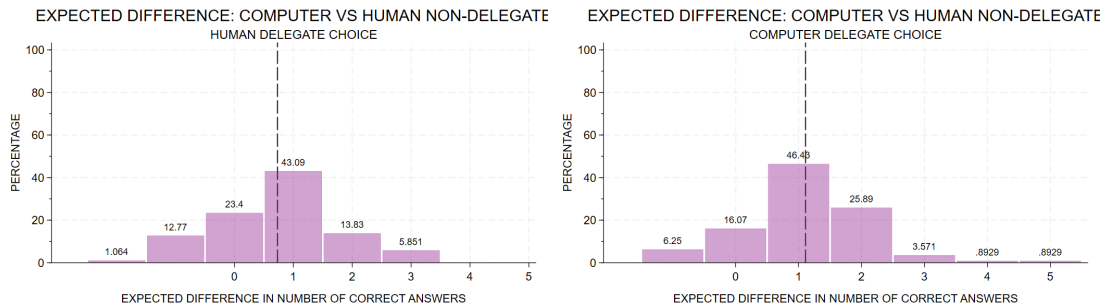
Note: Distribution of the number of expected correct answers for the human delegate, depending on the delegate's task. Dashed vertical lines indicate the mean number of expected correct answers.

Figure A.2: Expected Performance: Human Non-Delegate (By Delegate's Task)



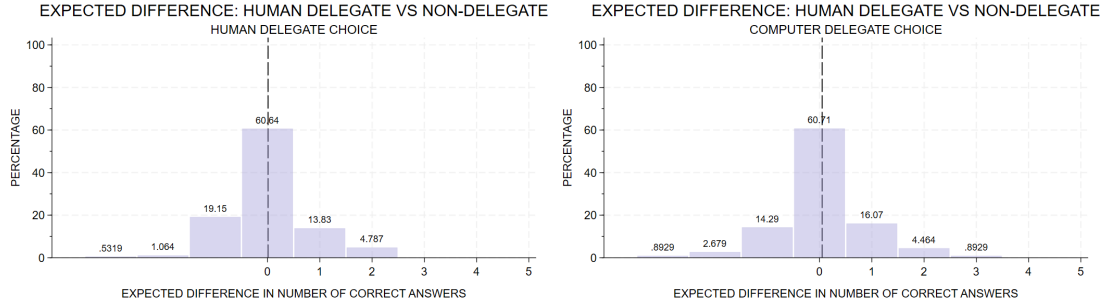
Note: Distribution of the number of expected correct answers for the human delegate, depending on delegate choice. Dashed vertical lines indicate the mean number of expected correct answers.

Figure A.3: Expected Performance: Human Non-Delegate (By Delegate Choice)



Note: Distribution of individual differences in the number of expected correct answers between the computer delegate and the human non-delegate, depending on delegate choice; calculated by subtracting human non-delegate performance expectations from computer delegate performance expectations. Dashed vertical lines indicate the mean difference in the number of expected correct answers.

Figure A.4: Expected Difference: Computer Delegate vs Human Non-Delegate



Note: Distribution of individual differences in the number of expected correct answers between the human delegate and the human non-delegate, depending on delegate choice; calculated by subtracting human non-delegate performance expectations from human delegate performance expectations. Dashed vertical lines indicate the mean difference in the number of expected correct answers.

Figure A.5: Expected Difference: Human Delegate vs Human Non-Delegate

	μ_1	μ_2	p (H_0 : EQUAL EXP) <i>Mann-Whitney U Test</i>
EXPECTATION: HUMAN DELEGATE			
HF-NPI – CF-NPI	2.62	3.05	0.0251
HF-PI – CF-PI	2.53	2.98	0.0010
HF-NPI – HF-PI	2.62	2.53	0.5534
CF-NPI – CF-PI	3.05	2.98	0.8876
EXPECTATION: HUMAN NON-DELEGATE			
HF-NPI – CF-NPI	2.66	3.03	0.0281
HF-PI – CF-PI	2.58	2.98	0.0235
HF-NPI – HF-PI	2.66	2.58	0.6492
CF-NPI – CF-PI	3.03	2.98	0.7357
EXPECTATION: COMPUTER DELEGATE			
HF-NPI – CF-NPI	3.44	4.02	0.0041
HF-PI – CF-PI	3.40	3.77	0.0118
HF-NPI – HF-PI	3.44	3.40	0.5072
CF-NPI – CF-PI	4.02	3.77	0.0541

Note: We have 300 performance expectations for each type of delegate; μ_1 denotes the mean number of expected correct answers for the indicated type of delegate in the first treatment group in a given row, and μ_2 denotes the mean number of expected correct answers for the indicated type of delegate in the second treatment group in a given row.

Table A.17: Treatment Comparisons of Performance Expectations

A.2 Instructions

A.2.1 Task Evaluation Survey

Consider the following task, which might be given to a computer or a human, and then answer the questions below.

Memory Task

A list of 25 words needs to be memorised. The order in which the words are presented is randomised. The list will be visible for a time-limit of 5 seconds before it disappears. After the list disappears, 5 words will be shown: 4 of these words were on the list, 1 of them was not.

The goal of the task is to select the word that was not on the list within a time-limit of 10 seconds. Note that success in this task is determined only by whether the goal is met within the time-limit, not by how fast this is accomplished.

Prediction Task

A game of Rock-Paper-Scissors needs to be played against a virtual agent. The virtual agent is programmed to choose each of the three words with a fixed probability: this means that the probability with which it chooses Rock, Paper or Scissors is the same every time the virtual agent makes a choice. A list of the virtual agent’s previous 25 choices will be visible for a time-limit of 5 seconds before it disappears. After the list disappears, there will be three words to choose from: Rock, Paper or Scissors. Rock beats Scissors, Scissors beats Paper, and Paper beats Rock. When both parties choose the same word, they play again until one party wins.

The goal of the task is to select a word that beats the virtual agent’s next choice within a time-limit of 10 seconds. Note that success in this task is determined only by whether the goal is met within the time-limit, not by how fast this is accomplished.

Coordination Task

One of 5 English words needs to be selected. In an experiment conducted in 2010, 44 human participants were presented with these 5 words and incentivised to select the same word as another participant who was chosen at random and incentivised in the same way. The words were arranged in a row from left to right, but the participants were told that the order of the words would be different for different participants. The participants did not know each other and could not communicate with each other.

The goal of the task is to select the same one of the 5 words as a randomly chosen participant from the 2010 experiment within a time-limit of 10 seconds. Note that success in this task is determined only by whether the goal is met within the time-limit, not by how fast this is accomplished.

CAPTCHA Task

A string of 6 lower-case letters needs to be deciphered. The letters are randomly generated, such that each of the 6 letters is equally likely to be any of the 26 letters in the English alphabet. The letters are distorted and overlaid with dots and a horizontal line.

The goal of the task is to type out all 6 lower-case letters in the correct order within a time-limit of 10 seconds. Note that success in this task is determined only by whether the goal is met within the time-limit, not by how fast this is accomplished.

Anagram Task

A 4-letter verb needs to be identified. The letters of the verb are scrambled, which means that their order was changed. There may be more than one way to rearrange the scrambled letters into a word of the English language, but there is only one solution that is a verb.

The goal of the task is to correctly type out the verb within a time-limit of 10 seconds. Note that success in this task is determined only by whether the goal is met within the time-limit, not by how fast this is accomplished.

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
I would expect a computer to perform this task well.	☐	☐	☐	☐	☐
I would expect a human to perform this task well.	☐	☐	☐	☐	☐
I would expect a computer to perform this task better than a human.	☐	☐	☐	☐	☐
I think a computer could be programmed to perform this task well.	☐	☐	☐	☐	☐

Automation Proficiency and Perceptions

How would you describe your **overall level of proficiency** when it comes to computers?

- ☐ Basic
- ☐ Intermediate
- ☐ Advanced
- ☐ Expert

On average, how many **hours per week** do you use a computer?

- ☐ 0 – 15 hours
- ☐ 15 – 30 hours
- ☐ 30 – 45 hours
- ☐ 45 – 60 hours
- ☐ 60+ hours

Are you familiar with any **programming languages**?

- ☐ Yes
- ☐ No

Which **programming language(s)** are you familiar with?

In our everyday lives, we often have the opportunity to have computers **guide** or **make decisions for us**. Consider for example using your **phone as a navigation device**, receiving **personalised ads** on the internet, or writing an electronic message with **auto-correct tools** on.

Thinking about the various ways in which computers can **guide** or **make** decisions for us... How would you rate the **overall influence** that computers have had on your life?

- ☐ Extremely negative
- ☐ Somewhat negative
- ☐ Neither positive nor negative
- ☐ Somewhat positive
- ☐ Extremely positive

Thinking about the various ways in which computers can **guide** or **make** decisions for us... How do you feel about it when computers **guide decisions** for you?

- ☐ Extremely bad
- ☐ Somewhat bad
- ☐ Neither good nor bad
- ☐ Somewhat good
- ☐ Extremely good

Thinking about the various ways in which computers can **guide** or **make** decisions for us... How do you **feel** about it when computers **make decisions** for you?

- o Extremely bad
- o Somewhat bad
- o Neither good nor bad
- o Somewhat good
- o Extremely good

Thinking about the various ways in which computers can **guide** or **make** decisions for us... When deciding whether or not to rely on a computer to guide or make decisions for you, how important is it to you to have **information about what the computer is programmed to do?**

- o Not at all important
- o Moderately important
- o Very important
- o Extremely important

Do you think each of the following things will or will not happen **in the next 20 years?**

Driverless vehicles, which can operate on their own without a human driver, are a widespread means of transportation.

- o Will definitely not happen
- o Will probably not happen
- o Will probably happen
- o Will definitely happen

Firms use computer programs that evaluate job candidates based on an automated review of their resume, application, and standardised test scores.

- o Will definitely not happen
- o Will probably not happen
- o Will probably happen
- o Will definitely happen

Computers and robots are able to do most of the jobs that are currently done by humans.

- o Will definitely not happen
- o Will probably not happen
- o Will probably happen
- o Will definitely happen

How **enthusiastic** are you, if at all, about the possibility of **driverless vehicles**?

- ☐ Not at all enthusiastic
- ☐ A little enthusiastic
- ☐ Quite enthusiastic
- ☐ Extremely enthusiastic

How **worried** are you, if at all, about the possibility of **driverless vehicles**?

- ☐ Not at all worried
- ☐ A little worried
- ☐ Quite worried
- ☐ Extremely worried

How **enthusiastic** are you, if at all, about the possibility of **computer programs evaluating job candidates**?

- ☐ Not at all enthusiastic
- ☐ A little enthusiastic
- ☐ Quite enthusiastic
- ☐ Extremely enthusiastic

How **worried** are you, if at all, about the possibility of **computer programs evaluating job candidates**?

- ☐ Not at all worried
- ☐ A little worried
- ☐ Quite worried
- ☐ Extremely worried

How **enthusiastic** are you, if at all, about the possibility **that computers and robots could do most of the jobs currently done by humans**?

- ☐ Not at all enthusiastic
- ☐ A little enthusiastic
- ☐ Quite enthusiastic
- ☐ Extremely enthusiastic

How **worried** are you, if at all, about the possibility **that computers and robots could do most of the jobs currently done by humans**?

- ☐ Not at all worried
- ☐ A little worried
- ☐ Quite worried
- ☐ Extremely worried

Demographic Questions

Gender

Age

Employment Status / Student Status

Employment Industry / Field of Study

How **likely**, if at all, do you think it is that your current job or profession / the job or profession associated with your current field of study will be **mostly done by computers or robots** in your lifetime?

- ☐ Extremely unlikely
- ☐ Somewhat unlikely
- ☐ Neither likely nor unlikely
- ☐ Somewhat likely
- ☐ Extremely likely

Comments

Thank you for your participation in this study. Please let us know if you have any comments.

A.2.2 Main Experiment*Overview*¹³

In this experiment, we want to investigate aspects of decision-making involving humans and computers.

There will be **three stages**. In each stage, you will make decisions for which you can earn money. At the end of the experiment, **one of the three stages will be randomly selected for payment**. You receive a **participation fee of £1.50** and you can earn a **bonus of up to £4.00** depending on which stage is selected for payment and what happens in the experiment.

After the experiment, there will be a short questionnaire. By clicking the “Continue” button below, you will begin **Stage 1** of the experiment.

¹³The instructions presented here are the ones displayed to the subjects in Cohorts 2 through 5, who participated in all three stages; adjustments were made for Cohorts 1 and 6 where necessary.

Stage 1

This is **Stage 1**.

In this stage, you have to **choose a delegate** to perform **a sequence of 5 tasks** for you, each of which has a unique correct solution. For every task that the delegate solves correctly, you will earn **£0.40**.

You can either choose a HUMAN or a COMPUTER as your delegate.

The human you can choose as your delegate is a **participant who will complete the experiment after you**. If you choose the human as your delegate, this person will be **informed that they are acting as a delegate** for someone who completed the experiment before them and who earns £0.40 for every task that they solve correctly. They will also receive £0.40 **themselves** for every task that they solve correctly. Regardless of whether you choose the human or the computer as your delegate, **the other delegate will still perform the task** and, if human, receive their own payment for it.

To make sure that you have **read and understood** the above information, please answer the following questions.

What will happen if you choose the **human as your delegate**?

- o They will be informed that they are acting as a delegate
- o They will not be informed that they are acting as a delegate

If you choose the **human as your delegate**...

- o Only the human will perform the sequence of 5 tasks
- o Only the computer will perform the sequence of 5 tasks
- o The human and the computer will both perform the sequence of 5 tasks but only the human acts as my delegate

If you choose the **computer as your delegate**...

- o Only the human will perform the sequence of 5 tasks
- o Only the computer will perform the sequence of 5 tasks
- o The human and the computer will both perform the sequence of 5 tasks but only the computer acts as my delegate

Before you choose one of the two delegates to perform the sequence of 5 tasks for you, note that:

- (1) Each task in the sequence has a **similar form**.
- (2) The human is incentivised to solve the tasks and the computer is programmed to do so, but **we expect neither delegate to perform perfectly**.

HF-NPI
HF Task Description

CF-NPI
CF Task Description

HF-PI
HF Task Description + Programming Information

CF-PI
CF Task Description + Programming Information¹⁴

Which delegate do you want to choose?

- o Human Delegate o Computer Delegate

Recall that you will earn **£0.40** for every task that your chosen delegate solves correctly and that there are 5 tasks in the sequence. This means that you can earn **between £0.00 and £2.00** based on their performance.

At the moment, your chosen delegate is the human/computer.

You now have the option to **swap** your chosen delegate for the other delegate **in exchange for an additional payment**. If you swap your chosen delegate, you will earn £0.40 for every task that **the computer/human** solves correctly and receive the additional payment. If you do not swap your chosen delegate, you will earn £0.40 for every task that **the human/computer** solves correctly and not receive the additional payment.

Please indicate whether you would want to **keep** or **swap** your chosen delegate in exchange for each of the payments on the list below. We will then **randomly select one of the payments on the list** and implement your decision for this payment.

For example, if the randomly selected payment is £0.80 and you selected ‘keep’ for this payment, **you will not make a swap nor be paid for swapping and the human/computer will remain your delegate**. If you selected ‘swap’ for this payment, **the computer/human will become your delegate and you receive an additional £0.80 for swapping**.

¹⁴Exact task descriptions are provided at the end of this section.

	Keep	Swap
£0.00	o	o
£0.20	o	o
...	o	o
£2.00	o	o

This concludes Stage 1 of the experiment. By clicking the “Continue” button below, you will move on to **Stage 2** of the experiment.

Stage 2

This is **Stage 2**.

In this stage, you have to **solve a sequence of 5 tasks**, each of which has a unique correct solution. For every task that you solve correctly, you earn **£0.40**.

If chosen as ‘Human Delegate’: You have also been chosen to **act as a delegate for a participant who completed the experiment before you**. This means that for every task that you solve correctly, this person also earns **£0.40**. Note that it is **not possible that you are the delegate of the person who will be your delegate**. Even if your delegate is a human, it will be a **participant who completes the experiment after you**. In contrast, you are the delegate for a participant who completed the experiment before you.

On the following page, you will be given some information about the sequence of 5 tasks.

B1, HF-NPI & HF-PI
HF Task Description

B2, CF-NPI & CF-PI
CF Task Description

On the following page, you will be shown **an example** to familiarise yourself with the task. Your performance in this example is **not relevant for your payment**.

Example Task

This concludes the Example Task. By clicking on the “Continue” button below, **you will begin the sequence of 5 tasks**.

B1, HF-NPI & HF-PI
HF Task

B2, CF-NPI & CF-PI
CF Task

This concludes Stage 2 of the experiment. By clicking the “Continue” button below, you will move on to **Stage 3** of the experiment.

Stage 3

This is **Stage 3**.

Recall that in Stage 1, you were asked to choose either a human delegate or a computer delegate perform a sequence of **5 tasks** for you.

If the human is your delegate, this person will be **informed that they are acting as a delegate** for someone who completed the experiment before them and who earns £0.40 for every task that they solve correctly. Regardless of whether your delegate is the human or the computer, **the other delegate will still perform the task**.

The **nature** of the sequence of 5 tasks that you *delegated* in Stage 1 is **different** from the **nature** of the sequence of 5 tasks that you *performed* in Stage 2. In this stage, you should focus **only on the sequence of 5 tasks that you delegated in Stage 1**.

On the following page, you will be reminded of the information you were given about the sequence of 5 tasks that you delegated in Stage 1.

- (1) Each task in the sequence has a **similar form**.
- (2) The human is incentivised to solve the tasks and the computer is programmed to do so, but **we expect neither delegate to perform perfectly**.

HF-NPI
HF Task Description

CF-NPI
CF Task Description

HF-PI

HF Task Description + Programming Information

CF-PI

CF Task Description + Programming Information

In this stage, we would like you to **guess how many of the 5 tasks** in the sequence **the human will solve correctly if they are acting as a delegate**.

If the human **is acting as a delegate** and your guess **matches** the number of tasks that they solve correctly, you will earn **£1.00**. Otherwise, you will not earn anything for this task:

We would also like you to **guess how many of the 5 tasks in the sequence the human will solve correctly if they are not acting as a delegate**.

If the human **is not acting as a delegate** and your guess **matches** the number of tasks that they solve correctly, you will earn **£1.00**. Otherwise, you will not earn anything for this task:

Now consider **the computer**. We would like you to **guess how many of the 5 tasks in the sequence the computer will solve correctly**.

If your guess **matches** the number of tasks that the computer solves correctly, you will earn **£1.00**. Otherwise, you will not earn anything for this task:

This concludes Stage 3 of the experiment. By clicking the “Continue” button below, you will move on to **the questionnaire**.

Automation Proficiency and Perceptions

In the following questionnaire, please answer some questions about **computers**. Your answers to these questions will **not** affect your payment, but you need to answer them in order to be paid.

How would you describe your **overall level of proficiency** when it comes to computers?

- ☐ Basic
- ☐ Intermediate
- ☐ Advanced
- ☐ Expert

Are you familiar with any **programming languages**?

- ☐ Yes
- ☐ No

Which **programming language(s)** are you familiar with?

In our everyday lives, we often have the opportunity to have computers **guide** or **make decisions for us**. Consider for example using your **phone as a navigation device**, receiving **personalised ads** on the internet, or writing an electronic message with **auto-correct tools** on.

Thinking about the various ways in which computers can **guide** or **make** decisions for us... How would you rate the **overall influence** that computers have had on your life?

- ☐ Extremely negative
- ☐ Somewhat negative
- ☐ Neither positive nor negative
- ☐ Somewhat positive
- ☐ Extremely positive

Thinking about the various ways in which computers can **guide** or **make** decisions for us... How do you feel about it when computers **guide or make decisions** for you?

- ☐ Extremely bad
- ☐ Somewhat bad
- ☐ Neither good nor bad
- ☐ Somewhat good
- ☐ Extremely good

Thinking about the various ways in which computers can **guide** or **make** decisions for us... When deciding whether or not to rely on a computer to guide or make decisions for you, how important is it to you to have **information about what the computer is programmed to do**?

- ☐ Not at all important
- ☐ Moderately important
- ☐ Very important
- ☐ Extremely important

Demographic Questions

Gender

Age

Employment Status / Student Status

Employment Industry / Field of Study

How **likely**, if at all, do you think it is that your current job or profession / the job or profession associated with your current field of study will be **mostly done by computers or robots** in your lifetime?

- o Extremely unlikely
- o Somewhat unlikely
- o Neither likely nor unlikely
- o Somewhat likely
- o Extremely likely

Comments

Thank you for your participation in this study. Please let us know if you have any comments.

A.2.3 Task Descriptions

HF Task

Each task in the sequence has the following form:

One of 5 English words needs to be selected. In an experiment reported in 2010, 44 human participants who did not know each other, could not communicate with each other, and had no access to the internet were presented with these five words. The words were arranged in a row from left to right, but the participants were informed that the order of the words would be different for different participants. Each participant was told that they would be rewarded if they selected the same word as another participant who was chosen at random and rewarded in the same way.

The goal of the task is to select the same one of the 5 words as a randomly chosen participant from the 2010 experiment within a time-limit of 10 seconds. Note that success in this task is determined only by whether the goal is met within the time-limit, not by how fast this is accomplished. Note that each of the 5 tasks in the sequence features a **different set of words**.

*The computer accesses **Google search data from 2010 via Google Trends**, a website that analyses the popularity of Google search queries. It compares the number of 2010 Google search queries for the 5 words, and **selects the word with the highest number of Google search queries** in that year.*

EXAMPLE¹⁵: **football** – chess – squash – tennis – volleyball

TASK 1: **Earth** – Mars – Mercury – Saturn – Venus

TASK 2: **walking** – jogging – *sunbathing* – swimming – sitting

TASK 3: **red** – *blue* – green – orange – purple

TASK 4: **Florida** – Colorado – Louisiana – Nevada – Ontario

TASK 5: **John** – David – Michael – Robert – Steven

CF Task

Each task in the sequence has the following form:

A list of 25 English words needs to be memorised. The order in which the words are presented is randomised. The list will be visible for 5 seconds before it disappears. After the list disappears, 5 words will be shown: 4 of these words were on the list, 1 of them was not.

The goal of the task is to select the word that was not on the list within a time-limit of 10 seconds. Note that success in this task is determined only by whether the goal is met within the time-limit, not by how fast this is accomplished. Note that each of the 5 tasks in the sequence features a **different list of words**.

*The computer records **the first 20 words** on the list. Among the 5 words that are shown after the list disappears, **it identifies the words that it did not record**. If there is only one, it will select this word. If there is more than one, it will **randomly** select one of them.*

EXAMPLE¹⁶: bed, rest, awake, tired, dream, wake, snooze, blanket, doze, slumber, snore, nap, peace, yawn, drowsy, nightmare, alarm, pillow, sheets, relax, late, soft, cloud, forest, dinner

plant – pillow – dinner – snooze – awake

TASK 1: door, glass, pane, shade, ledge, sill, house, open, curtain, frame, view, breeze, sash, screen, shutter, room, closed, light, floor, roof, clean, wall, pencil, car, phone

towel – room – glass – screen – phone

¹⁵Computer choices based on Google Trends are marked in **bold**; the most popular choices of the 44 human subjects in Bardsley et al. (2010) are marked in *italics*.

¹⁶The correct solution for each list is underlined; the first 15 words on each list are taken from Deese (1959), the remaining 10 words were added for the purpose of this experiment.

TASK 2: sour, candy, sugar, bitter, good, taste, tooth, nice, honey, soda, chocolate,
heart, cake, tart, pie, fruit, tongue, mouth, syrup, biscuits, pudding, snack, lamp,
earring, letter

scissors – cake – syrup – letter – nice

TASK 3: hot, snow, warm, winter, ice, wet, frigid, chilly, heat, weather, freeze, air,
shiver, Arctic, frost, water, goosebumps, coat, scarf, humidity, gloves, hail, flag,
mirror, notebook

backpack – water – shiver – mirror – hail

TASK 4: nurse, sick, lawyer, medicine, health, hospital, dentist, physician, ill, patient,
office, stethoscope, surgeon, clinic, cure, disease, scrubs, vaccination, pill, recovery,
white, scalpel, vase, parcel, radiator

soap – health – lawyer – vase – disease

TASK 5: note, sound, piano, sing, radio, band, melody, horn, concert, instrument,
symphony, jazz, orchestra, art, rhythm, violin, flute, percussion, guitar, harmony,
loud, record, table, book, wool

wallet – band – art – guitar – table

B Appendix for Chapter 2

B.1 Additional Analysis

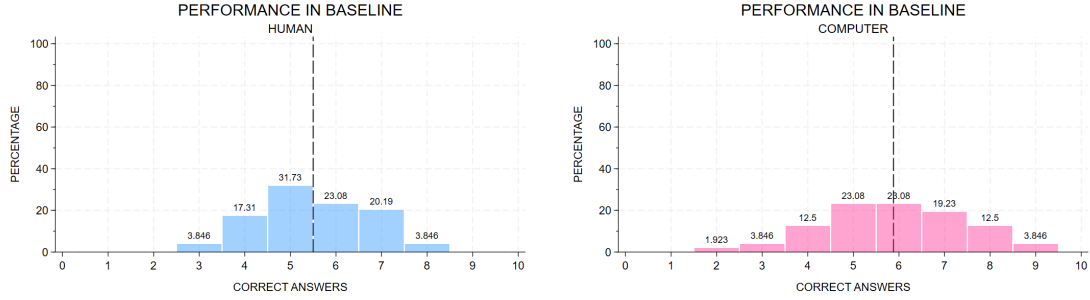
	TOTAL	PERCENTAGE
GENDER		
FEMALE	306	60.71
MALE	192	38.10
OTHER	6	1.19
AGE		
MIN	18	
MAX	75	
MEAN	40.63	
NATIONALITY		
UK	484	96.03
OTHER	20	3.97
HIGHEST LEVEL OF EDUCATION COMPLETED		
NO FORMAL QUALIFICATIONS	0	0.00
SECONDARY EDUCATION (E.G. GED/GCSE)	52	10.32
HIGH SCHOOL DIPLOMA/A-LEVELS	73	14.48
TECHNICAL/COMMUNITY COLLEGE	66	13.10
UNDERGRADUATE DEGREE (BA/BSc/OTHER)	205	40.67
GRADUATE DEGREE (MA/MSc/MPhil/OTHER)	96	19.05
DOCTORATE DEGREE (PhD/OTHER)	12	2.38

Table B.1: Demographic Sample Composition

	BASILINE N=104	BO-CC N=101	GO-CC N=101	BO-EC N=99	GO-EC N=99
FEMALE SUBJECTS	60.58	63.37	61.39	56.57	61.62
MEAN AGE	42.81	39.73	40.73	39.72	40.04
MEAN RISK SCORE	0.47	0.46	0.50	0.50	0.41
MEAN AMBIGUITY SCORE	0.46	0.43	0.45	0.45	0.44

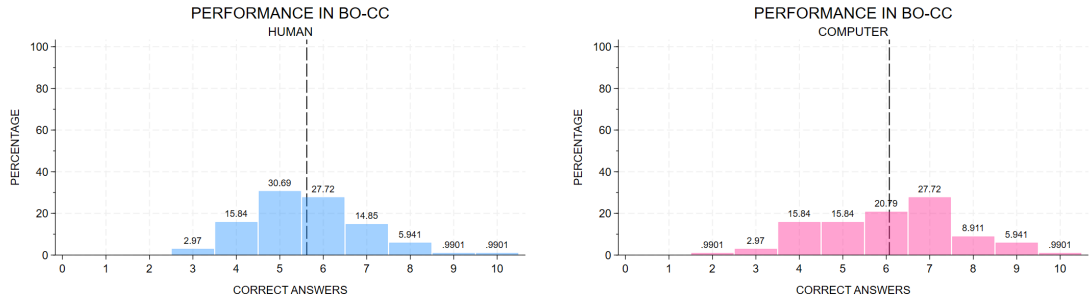
Note: RISK SCORE is a score from 0 to 1, where 0 corresponds to the most risk-averse and 1 to the most risk-seeking behaviour. AMBIGUITY SCORE is a probability where 0 corresponds to the most ambiguity-averse and 1 to the most ambiguity-seeking behaviour. FEMALE SUBJECTS represents within-treatment percentages of female subjects.

Table B.2: Demographic Treatment Composition



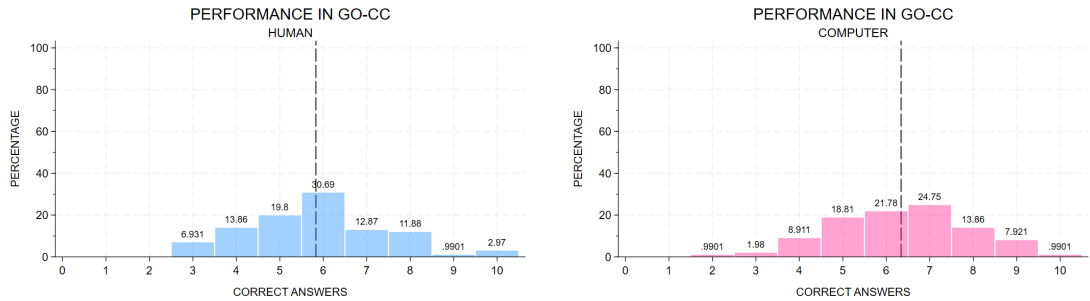
Note: Distribution of the number of correct answers in the ten rounds of the Dot Task in the BASELINE. Dashed vertical lines indicate the mean numbers of correct answers.

Figure B.1: Performance in BASELINE



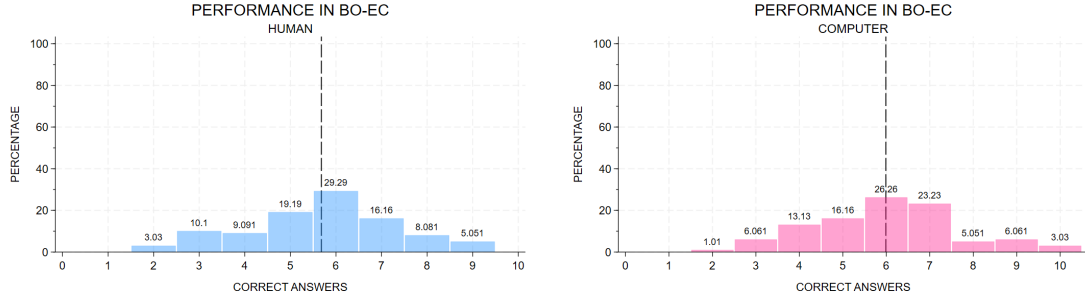
Note: Distribution of the number of correct answers in the ten rounds of the Dot Task in BO-CC. Dashed vertical lines indicate the mean numbers of correct answers.

Figure B.2: Performance in BO-CC



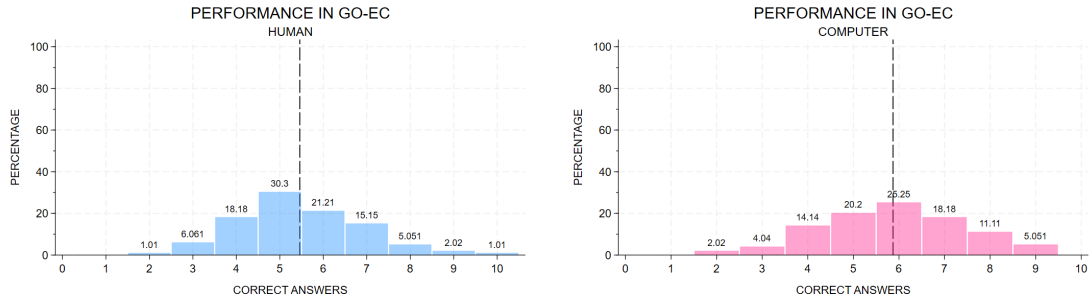
Note: Distribution of the number of correct answers in the ten rounds of the Dot Task in GO-CC. Dashed vertical lines indicate the mean numbers of correct answers.

Figure B.3: Performance in GO-CC



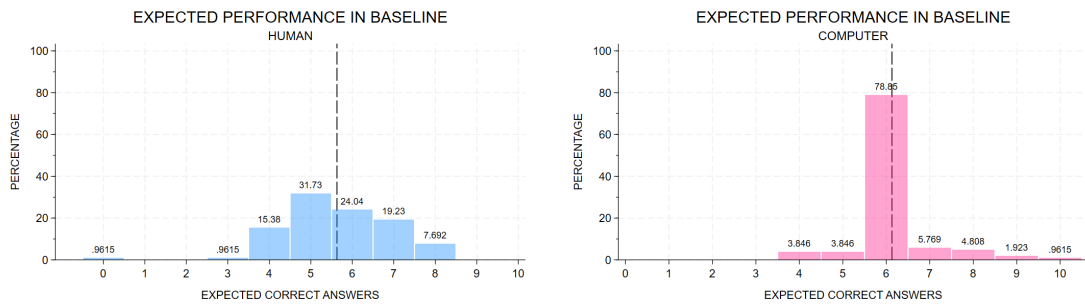
Note: Distribution of the number of correct answers in the ten rounds of the Dot Task in BO-EC. Dashed vertical lines indicate the mean numbers of correct answers.

Figure B.4: Performance in BO-EC



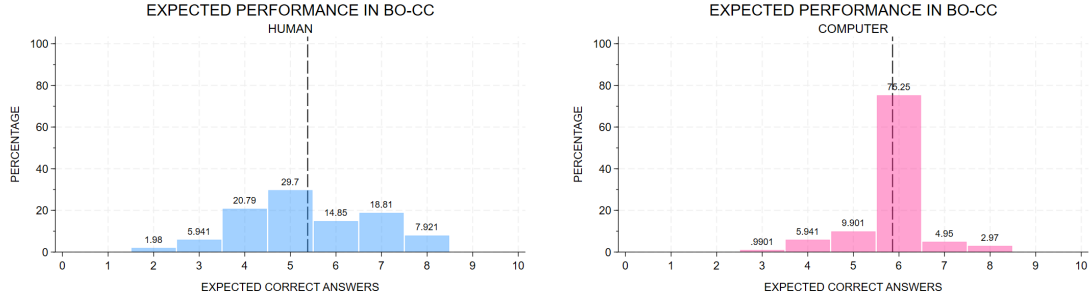
Note: Distribution of the number of correct answers in the ten rounds of the Dot Task in GO-EC. Dashed vertical lines indicate the mean numbers of correct answers.

Figure B.5: Performance in GO-EC



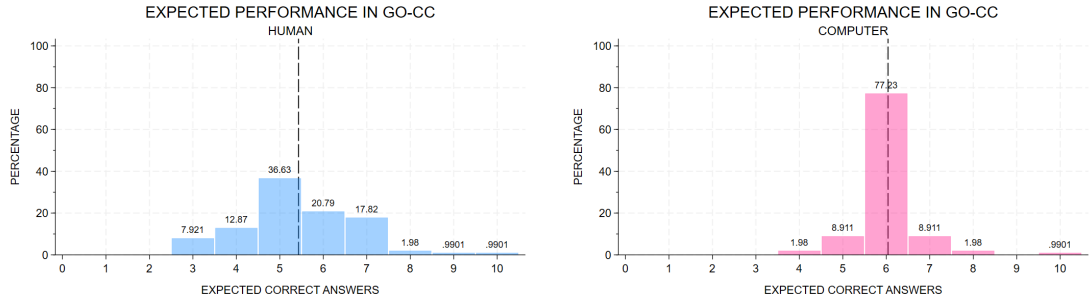
Note: Distribution of the number of expected correct answers in the ten rounds of the Dot Task in the BASELINE. Dashed vertical lines indicate the mean numbers of expected correct answers.

Figure B.6: Expected Performance in BASELINE



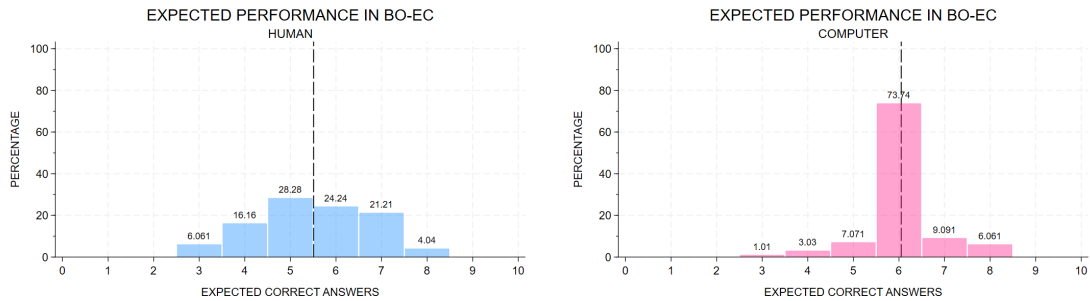
Note: Distribution of the number of expected correct answers in the ten rounds of the Dot Task in BO-CC. Dashed vertical lines indicate the mean numbers of expected correct answers.

Figure B.7: Expected Performance in BO-CC



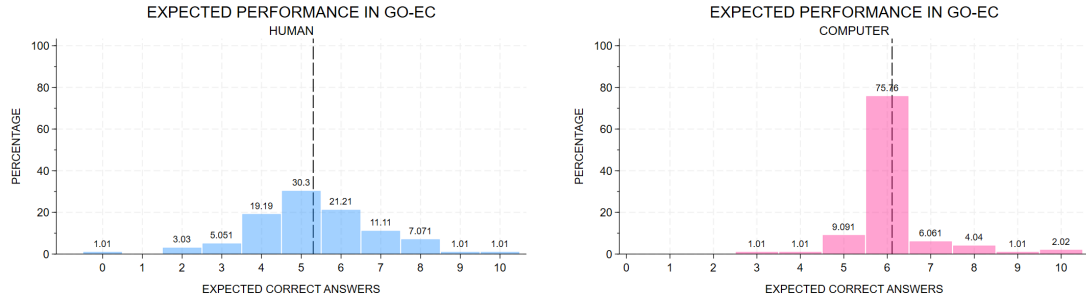
Note: Distribution of the number of expected correct answers in the ten rounds of the Dot Task in GO-CC. Dashed vertical lines indicate the mean numbers of expected correct answers.

Figure B.8: Expected Performance in GO-CC



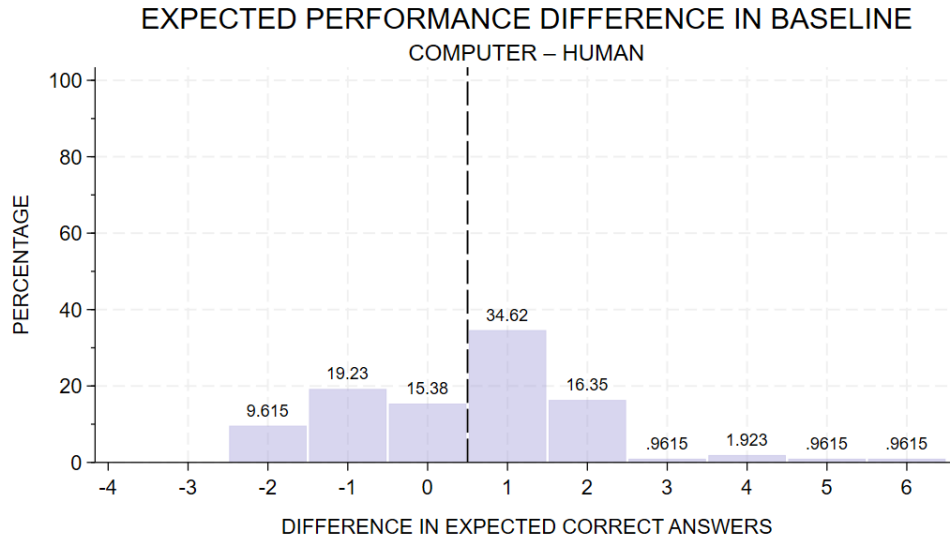
Note: Distribution of the number of expected correct answers in the ten rounds of the Dot Task in BO-EC. Dashed vertical lines indicate the mean numbers of expected correct answers.

Figure B.9: Expected Performance in BO-EC



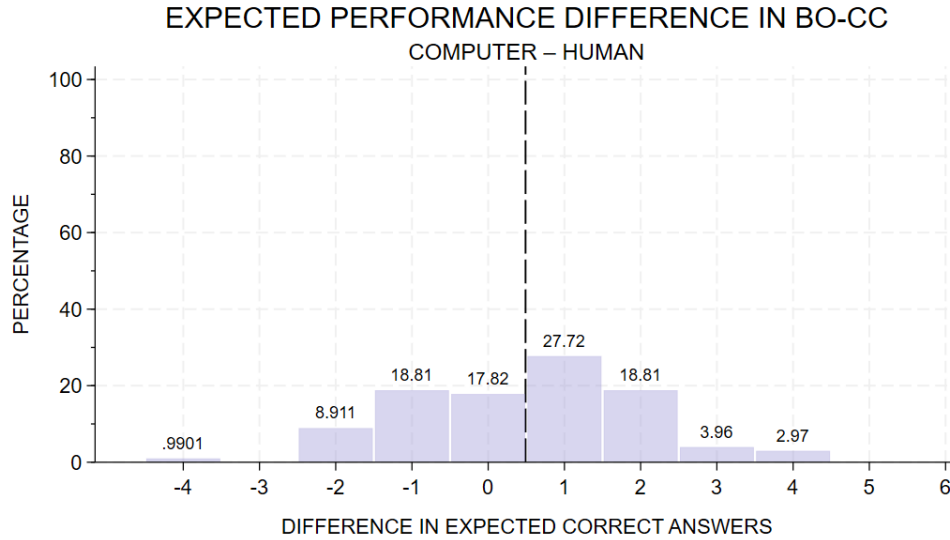
Note: Distribution of the number of expected correct answers in the ten rounds of the Dot Task in GO-EC. Dashed vertical lines indicate the mean numbers of expected correct answers.

Figure B.10: Expected Performance in GO-EC



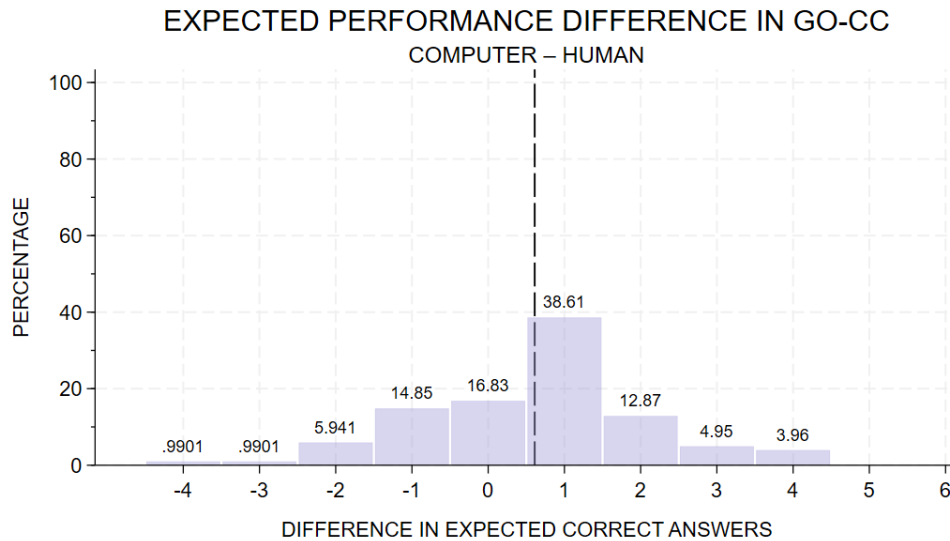
Note: Distribution of individual differences in the number of expected correct answers in the BASELINE; calculated by subtracting own performance expectations from computer performance expectations. Positive (negative) values indicate higher (lower) computer performance expectations, zero indicates equal performance expectations. Dashed vertical line indicates the mean difference in the number of expected correct answers.

Figure B.11: Performance Expectation Differences in BASELINE



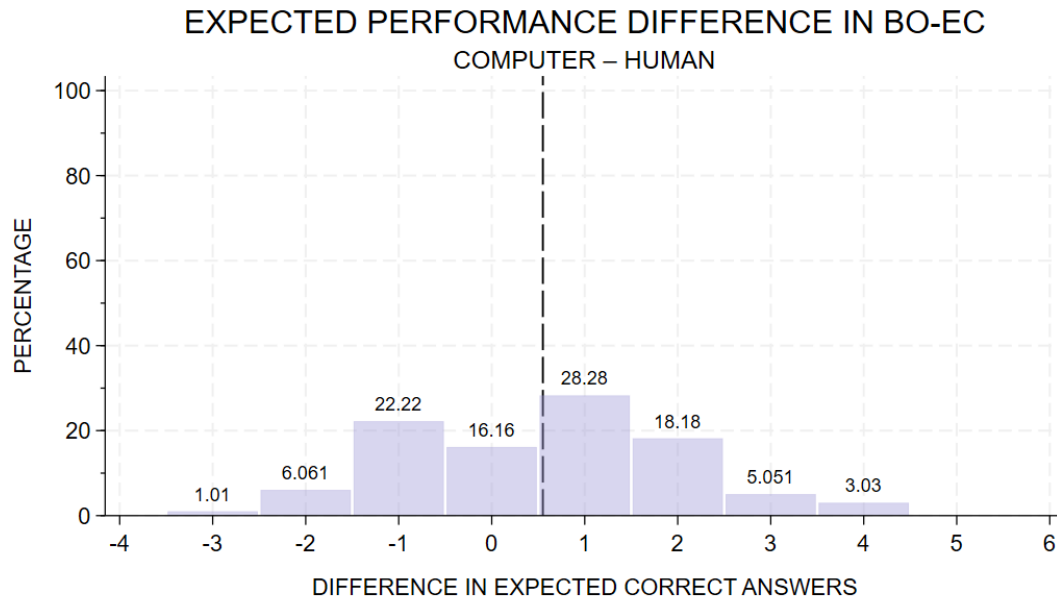
Note: Distribution of individual differences in the number of expected correct answers in BO-CC; calculated by subtracting own performance expectations from computer performance expectations. Positive (negative) values indicate higher (lower) computer performance expectations, zero indicates equal performance expectations. Dashed vertical line indicates the mean difference in the number of expected correct answers.

Figure B.12: Performance Expectation Differences in BO-CC



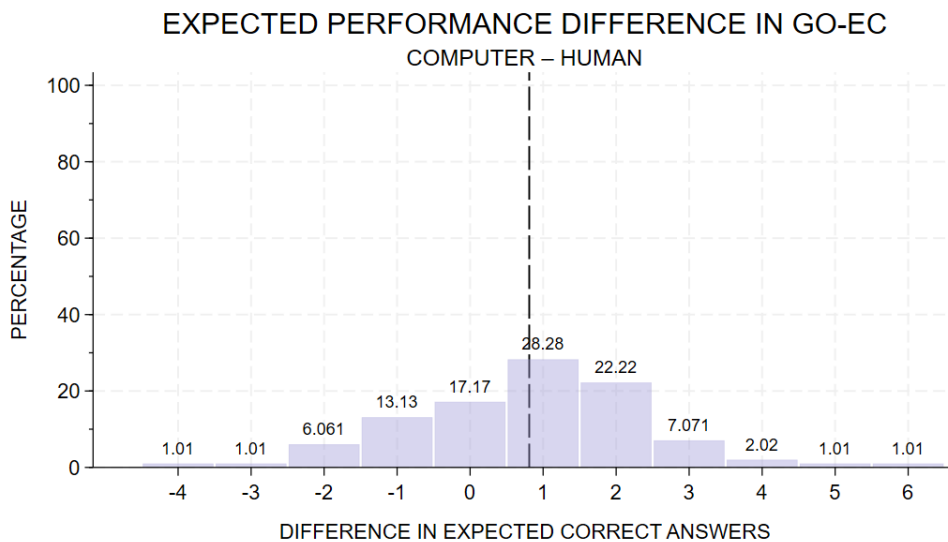
Note: Distribution of individual differences in the number of expected correct answers in GO-CC; calculated by subtracting own performance expectations from computer performance expectations. Positive (negative) values indicate higher (lower) computer performance expectations, zero indicates equal performance expectations. Dashed vertical line indicates the mean difference in the number of expected correct answers.

Figure B.13: Performance Expectation Differences in GO-CC



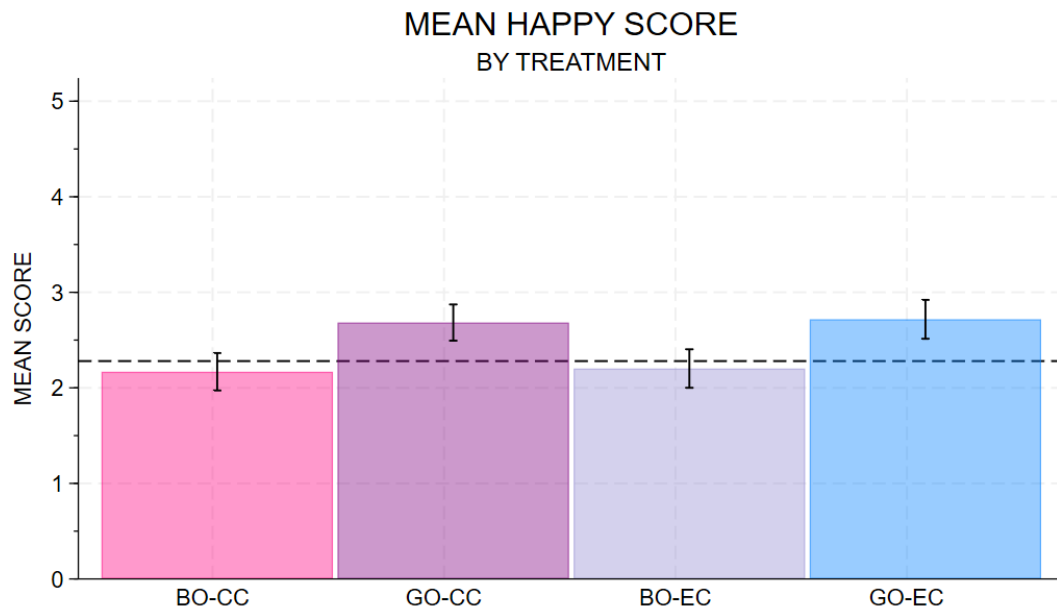
Note: Distribution of individual differences in the number of expected correct answers in BO-EC; calculated by subtracting own performance expectations from computer performance expectations. Positive (negative) values indicate higher (lower) computer performance expectations, zero indicates equal performance expectations. Dashed vertical line indicates the mean difference in the number of expected correct answers.

Figure B.14: Performance Expectation Differences in BO-EC



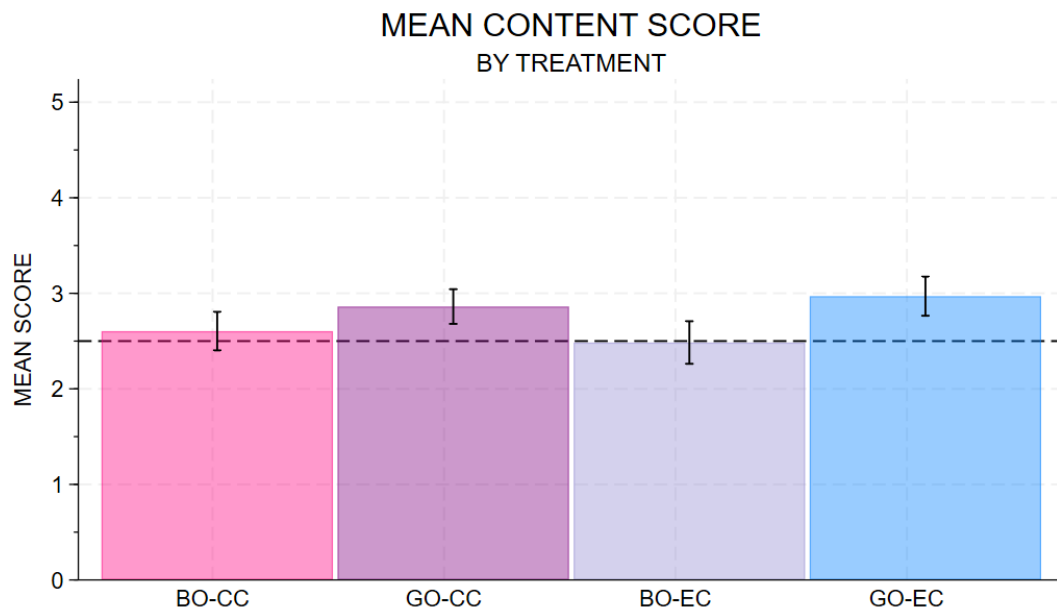
Note: Distribution of individual differences in the number of expected correct answers in GO-EC; calculated by subtracting own performance expectations from computer performance expectations. Positive (negative) values indicate higher (lower) computer performance expectations, zero indicates equal performance expectations. Dashed vertical line indicates the mean difference in the number of expected correct answers.

Figure B.15: Performance Expectation Differences in GO-EC



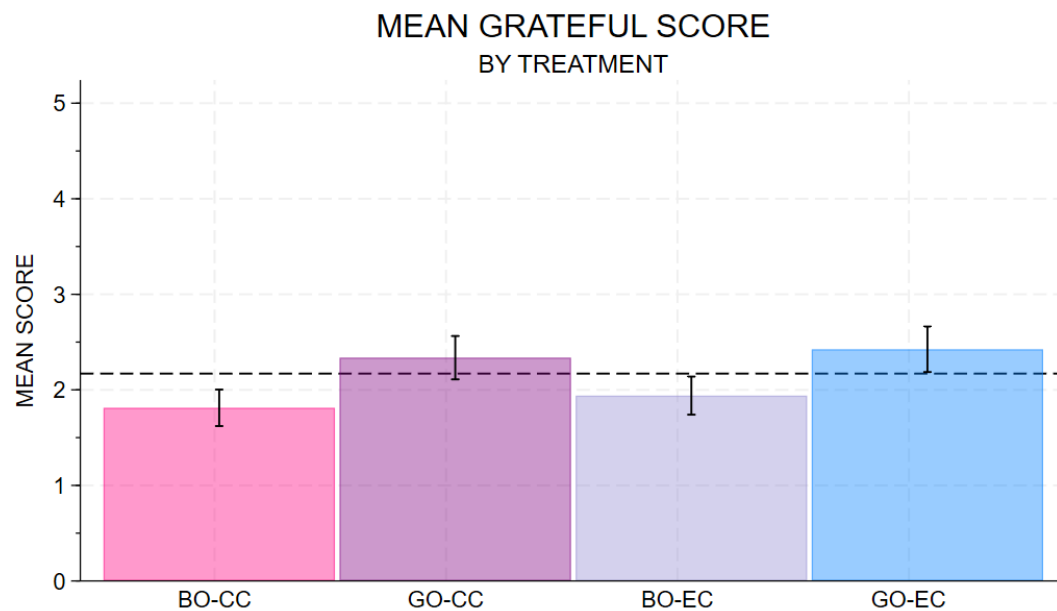
Note: Score ranges from 1 for minimum to 5 for maximum happiness. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.16: Happy Score



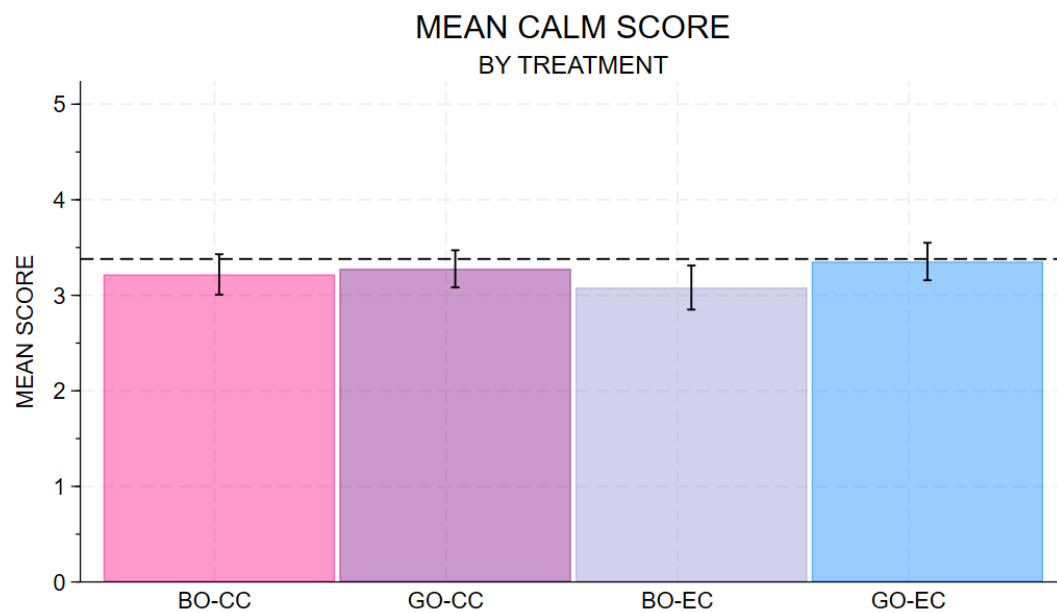
Note: Score ranges from 1 for minimum to 5 for maximum contentment. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.17: Content Score



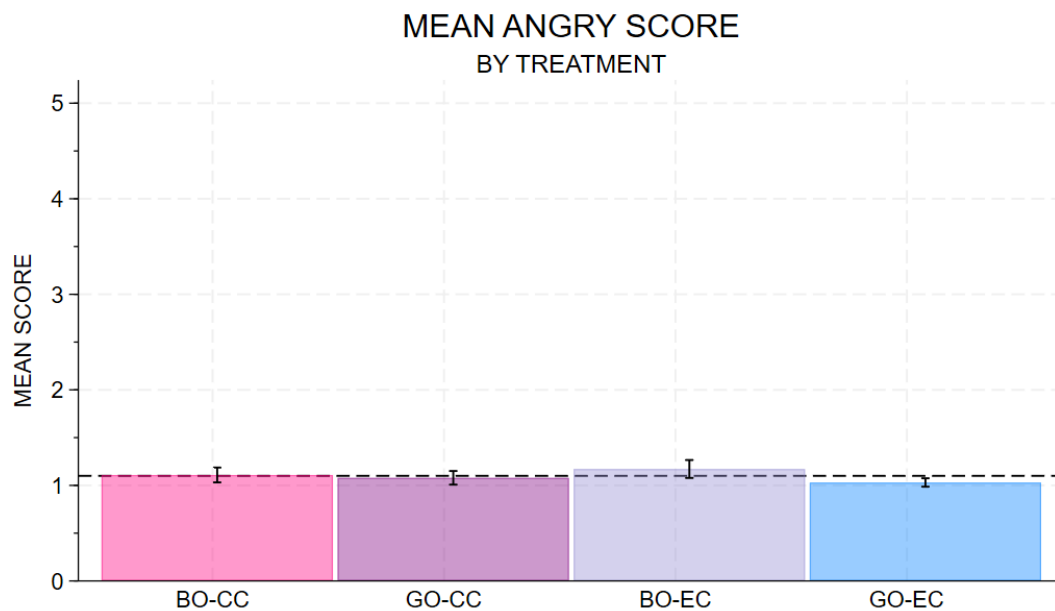
Note: Score ranges from 1 for minimum to 5 for maximum gratefulness. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.18: Grateful Score



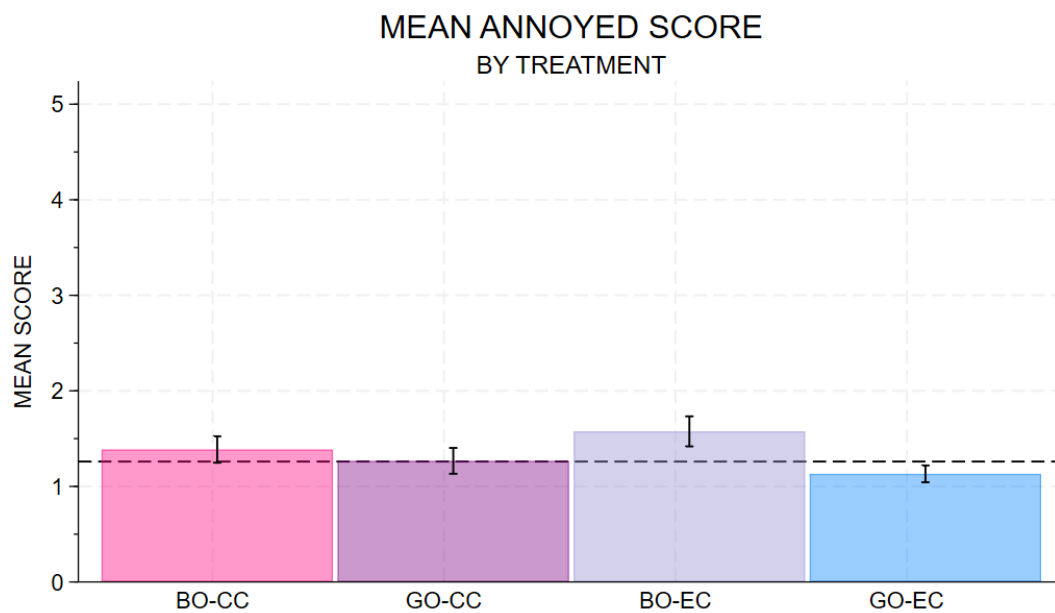
Note: Score ranges from 1 for minimum to 5 for maximum calmness. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.19: Calm Score



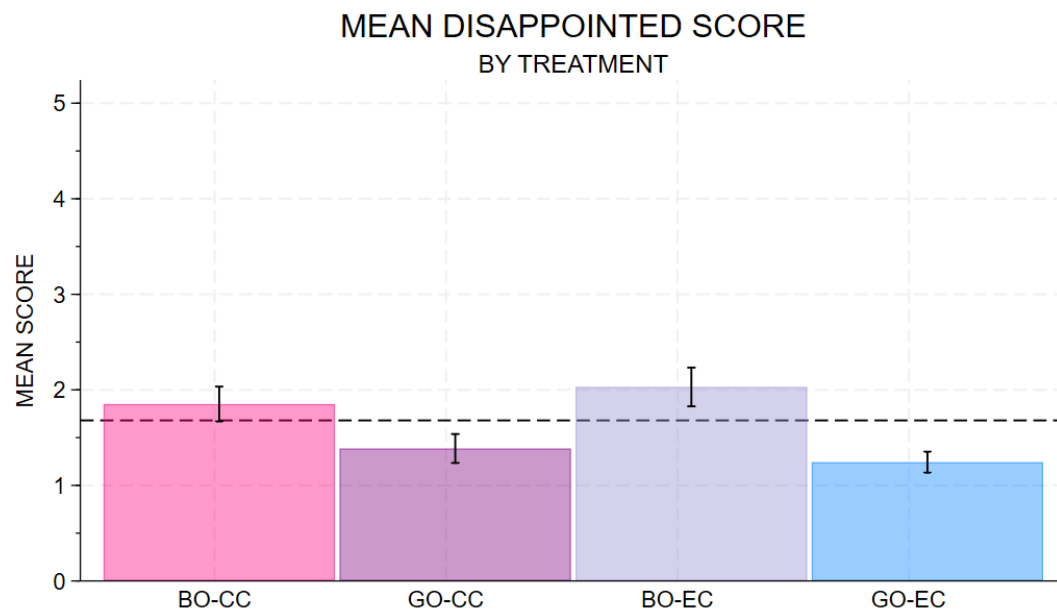
Note: Score ranges from 1 for minimum to 5 for maximum anger. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.20: Angry Score



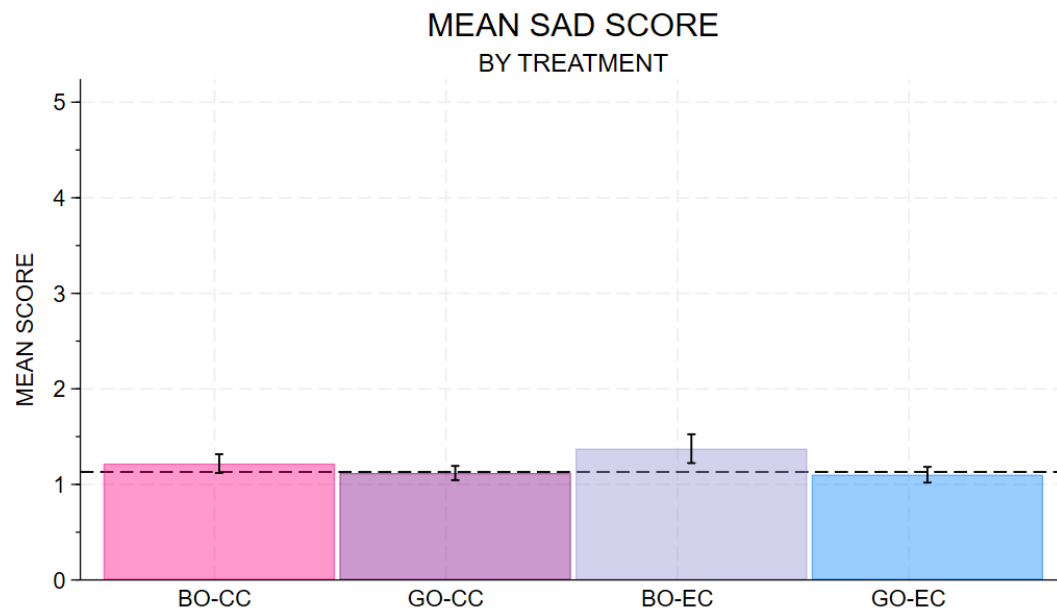
Note: Score ranges from 1 for minimum to 5 for maximum annoyance. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.21: Annoyed Score



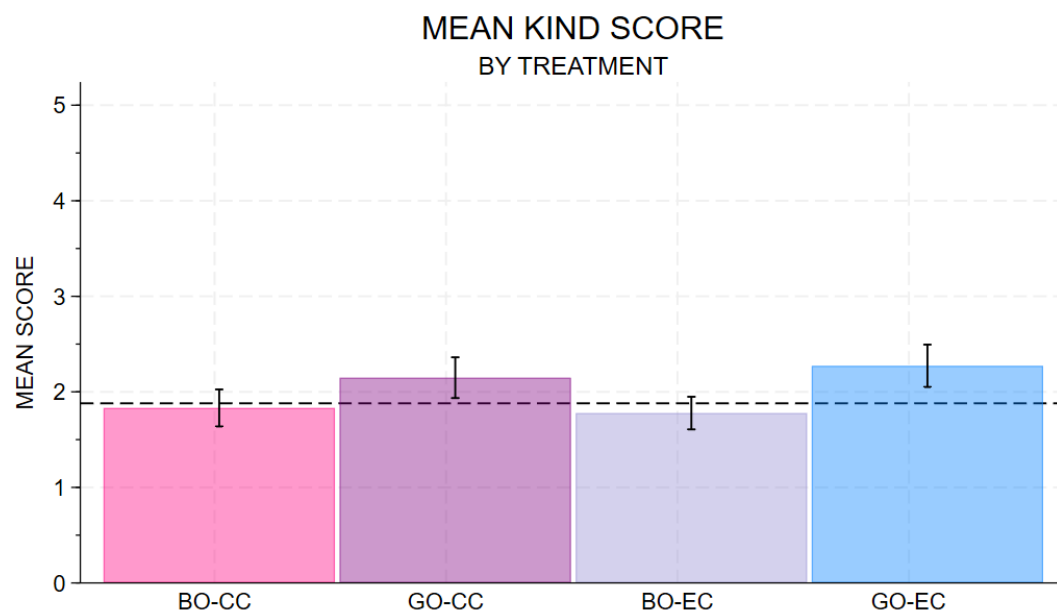
Note: Score ranges from 1 for minimum to 5 for maximum disappointment. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.22: Disappointed Score



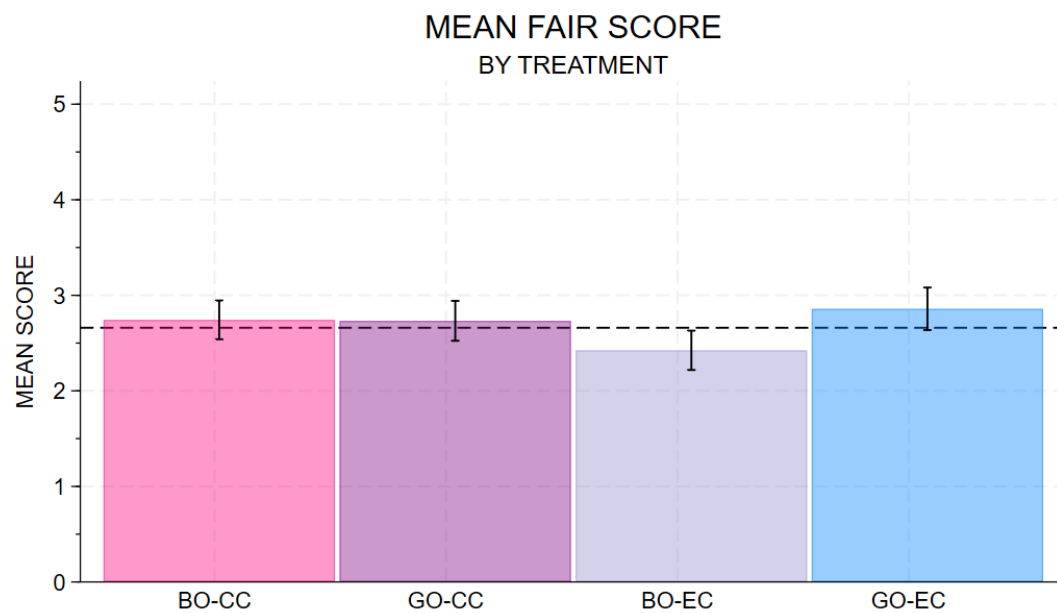
Note: Score ranges from 1 for minimum to 5 for maximum sadness. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.23: Sad Score



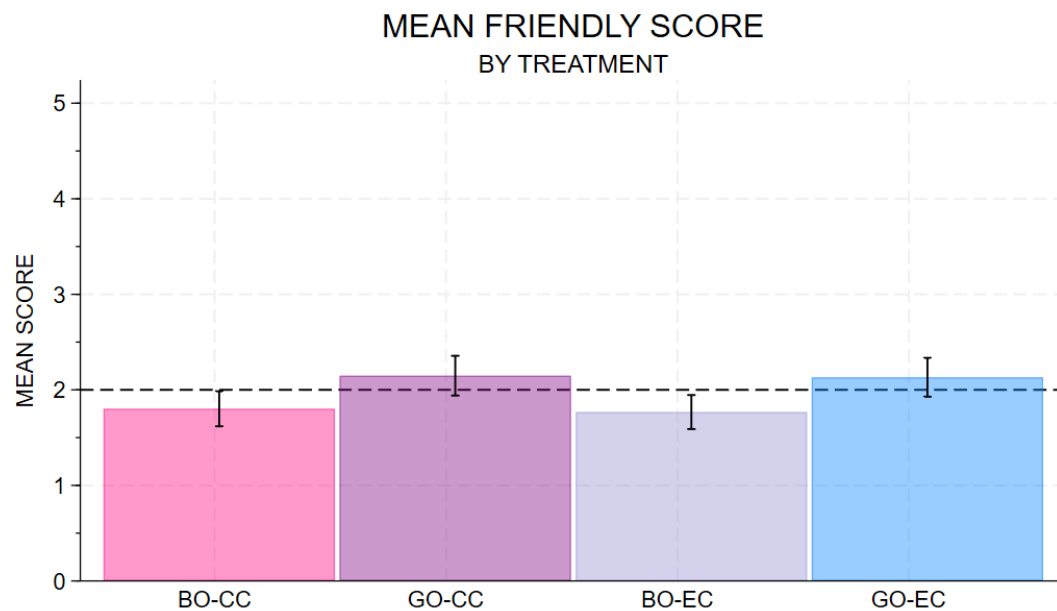
Note: Score ranges from 1 for minimum to 5 for maximum kindness. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.24: Kind Score



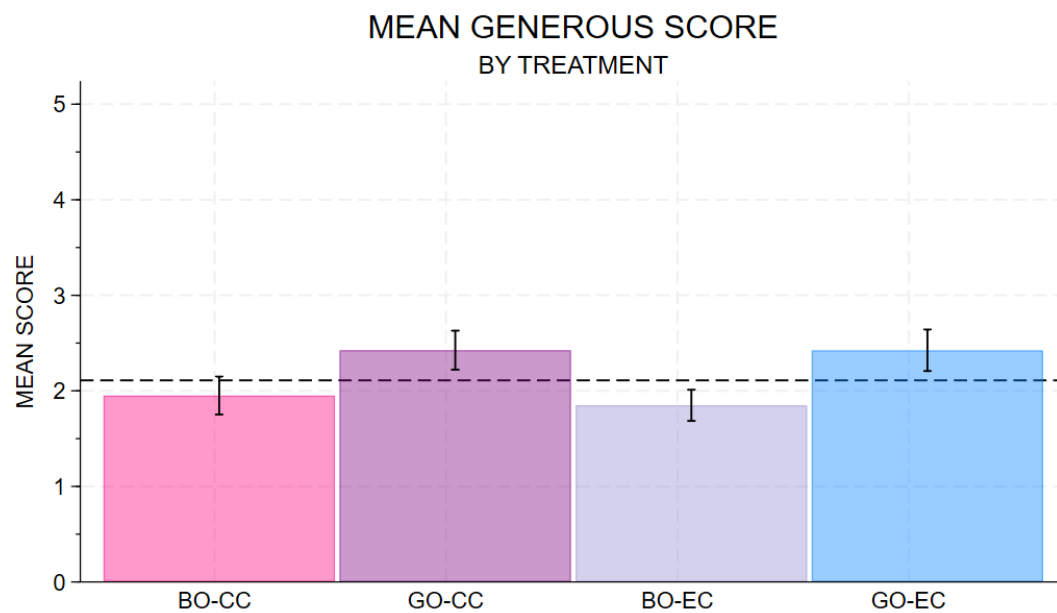
Note: Score ranges from 1 for minimum to 5 for maximum fairness. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.25: Fair Score



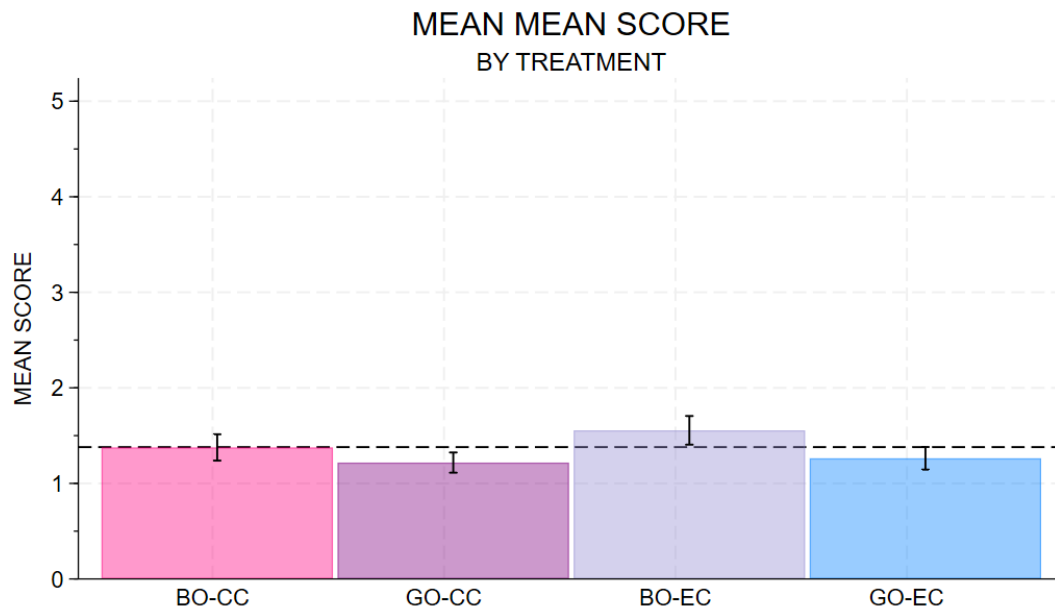
Note: Score ranges from 1 for minimum to 5 for maximum friendliness. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.26: Friendly Score



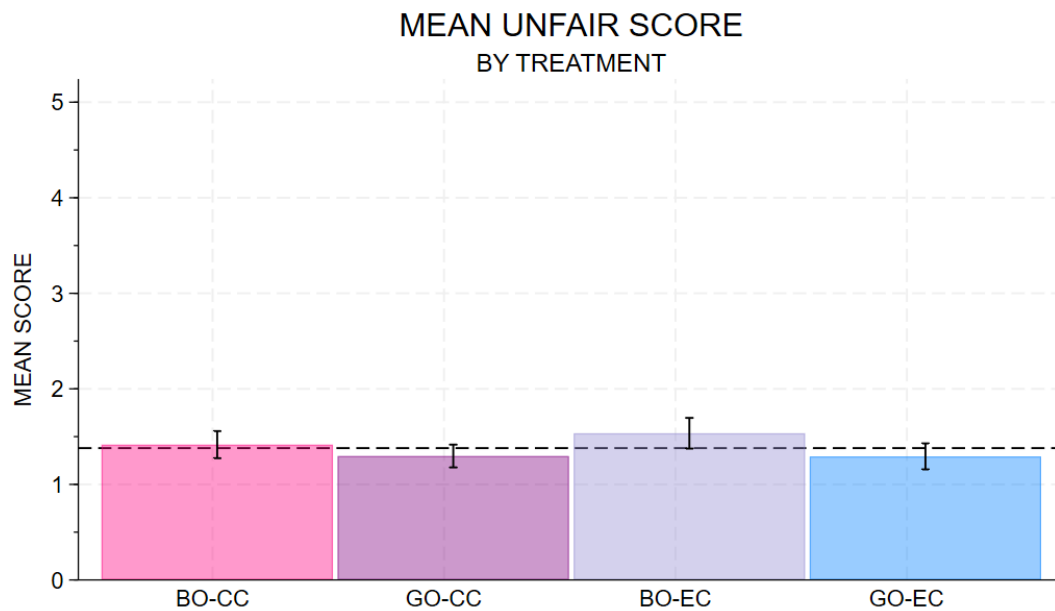
Note: Score ranges from 1 for minimum to 5 for maximum generosity. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.27: Generous Score



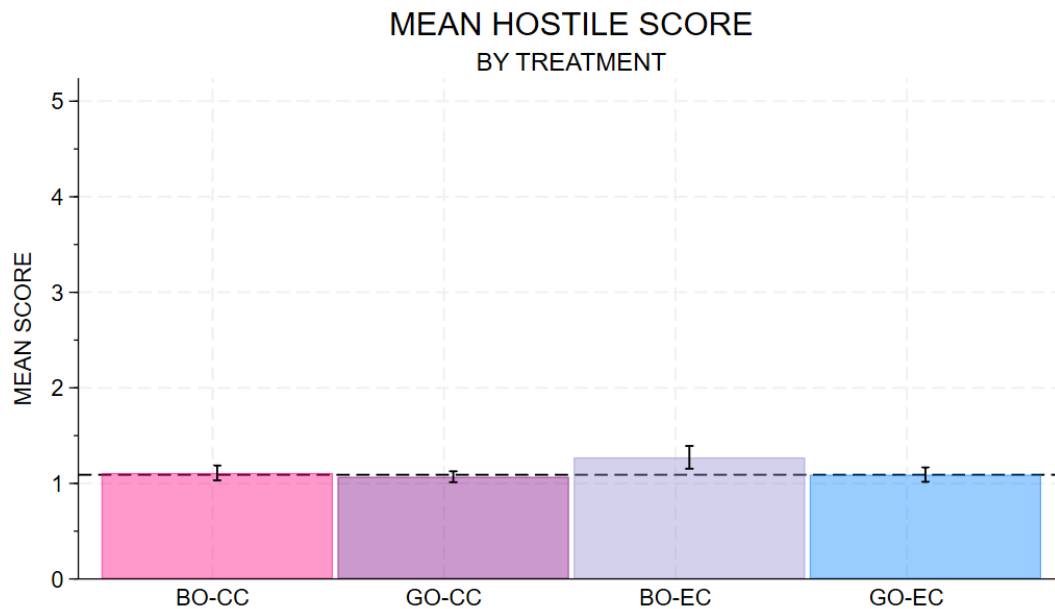
Note: Score ranges from 1 for minimum to 5 for maximum meanness. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.28: Mean Score



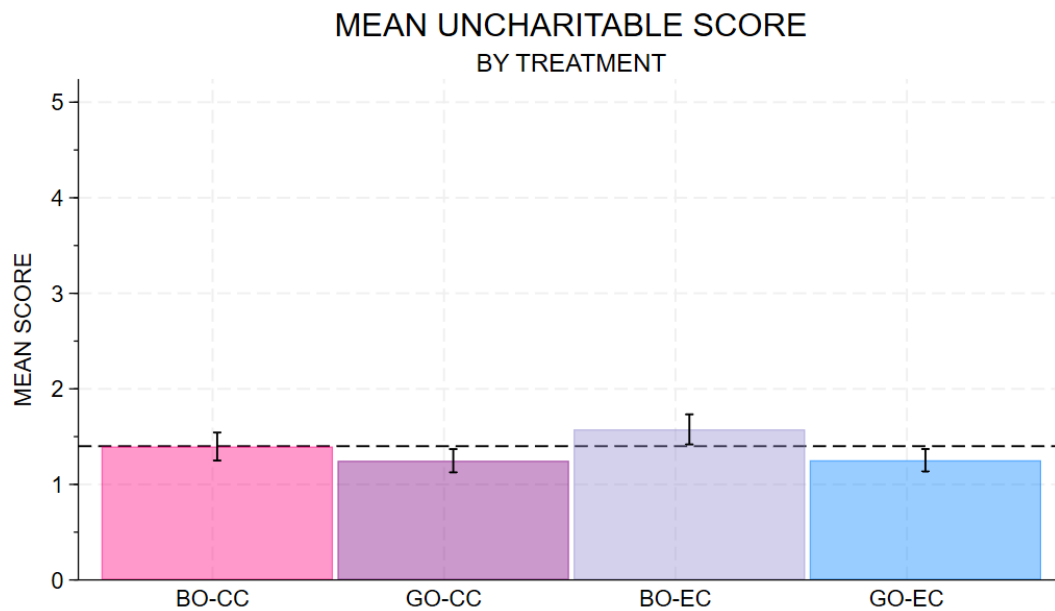
Note: Score ranges from 1 for minimum to 5 for maximum unfairness. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.29: Unfair Score



Note: Score ranges from 1 for minimum to 5 for maximum hostility. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.30: Hostile Score



Note: Score ranges from 1 for minimum to 5 for maximum uncharitableness. Whiskers indicate 95% confidence intervals; dashed line indicates the BASELINE benchmark.

Figure B.31: Uncharitable Score

	BASELINE	BO-CC	GO-CC	BO-EC	GO-EC
HAPPY	2	2	3	2	3
CONTENT	3	3	3	2	3
GRATEFUL	2	1	2	2	2
CALM	4	3	3	3	3
ANGRY	1	1	1	1	1
ANNOYED	1	1	1	1	1
DISAPPOINTED	1	2	1	2	1
SAD	1	1	1	1	1

Table B.3: Median Emotion Scores

	BASELINE	BO-CC	GO-CC	BO-EC	GO-EC
KIND	2	1	2	2	2
FAIR	3	3	3	2	3
FRIENDLY	2	2	2	1	2
GENEROUS	2	2	3	2	2
MEAN	1	1	1	1	1
UNFAIR	1	1	1	1	1
HOSTILE	1	1	1	1	1
UNCHARITABLE	1	1	1	1	1

Table B.4: Median Kindness Scores

	HAPPY	CONTENT	GRATEFUL	CALM
HAPPY	1.0000***			
CONTENT	0.6315***	1.0000***		
GRATEFUL	0.6330***	0.5019***	1.0000***	
CALM	0.4350***	0.5357***	0.2829***	1.0000***

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B.5: Intercorrelations: Positive Emotional Reactions

	ANGRY	ANNOYED	DISAPPOINTED	SAD
ANGRY	1.0000***			
ANNOYED	0.4351***	1.0000***		
DISAPPOINTED	0.3244***	0.6002***	1.0000***	
SAD	0.5791***	0.4637***	0.4664***	1.0000***

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B.6: Intercorrelations: Negative Emotional Reactions

	KIND	FAIR	FRIENDLY	GENEROUS
KIND	1.0000***			
FAIR	0.4671***	1.0000***		
FRIENDLY	0.7832***	0.4260***	1.0000***	
GENEROUS	0.6600***	0.5600***	0.6351***	1.0000***

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B.7: Intercorrelations: Positive Kindness Perceptions

	MEAN	UNFAIR	HOSTILE	UNCHARITABLE
MEAN	1.0000***			
UNFAIR	0.5219***	1.0000***		
HOSTILE	0.4454***	0.4066***	1.0000***	
UNCHARITABLE	0.5513***	0.5273***	0.4070***	1.0000***

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B.8: Intercorrelations: Negative Kindness Perceptions

	POS-EMOTIONS	NEG-EMOTIONS	POS-KINDNESS	NEG-KINDNESS
CROHNBACH'S α	0.8155	0.8079	0.8601	0.7732

Note: POS-EMOTIONS comprises HAPPY, CONTENT, GRATEFUL, and CALM scores; NEG-EMOTIONS comprises ANGRY, ANNOYED, DISAPPOINTED, and SAD scores. POS-KINDNESS comprises KIND, FAIR, FRIENDLY, and GENEROUS scores. NEG-KINDNESS comprises MEAN, UNFAIR, HOSTILE, and UNCHARITABLE scores.

Table B.9: Internal Consistency Reliabilities: Emotions & Kindness Perceptions

	UNDERRELiance		UNDERRELiance
HAPPY	-0.1468***	KIND	-0.1162***
CONTENT	-0.1474***	FAIR	-0.0933**
GRATEFUL	-0.2466***	FRIENDLY	-0.1376***
CALM	-0.0873*	GENEROUS	-0.1138**
ANGRY	0.0220	MEAN	-0.0491
ANNOYED	0.0355	UNFAIR	0.0300
DISAPPOINTED	0.1108**	HOSTILE	-0.0052
SAD	-0.0283	UNCHARITABLE	0.0230

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B.10: Correlations: Underreliance & Emotional Reactions

Table B.11: Correlations: Underreliance & Kindness Perceptions

GENERAL RISK QUESTION	FREQUENCY	PERCENTAGE
0.0000	10	1.98
0.1000	33	6.55
0.2000	56	11.11
0.3000	78	15.48
0.4000	53	10.52
0.5000	62	12.30
0.6000	76	15.08
0.7000	80	15.87
0.8000	43	8.53
0.9000	10	1.98
1.0000	3	0.60

Note: Respondents indicated their risk attitude on a scale of 1 – 10, where 0 meant ‘not at all willing to take risks’ and 10 meant ‘very willing to take risks’; we then normalised this score to range between 0 and 1.

Table B.12: Risk Preferences (Dohmen et al., 2011)

MATCHING PROBABILITY m	FREQUENCY	PERCENTAGE
0.0250	13	2.58
0.0750	1	0.20
0.1250	3	0.60
0.1750	3	0.60
0.2250	11	2.18
0.2750	10	1.98
0.3250	20	3.97
0.3750	108	21.43
0.4250	54	10.71
0.4750	107	21.23
0.5250	99	19.64
0.5750	49	9.72
0.6250	6	1.19
0.6750	14	2.78
0.7750	2	0.40
0.8500	2	0.40
0.9500	2	0.40

Note: Matching probabilities below 0.5 indicate ambiguity-averse attitudes; conversely, matching probabilities above 0.5 indicate ambiguity-seeking attitudes.

Table B.13: Ambiguity Preferences (Cavatorta & Schröder, 2019)

B.2 Instructions

Overview

In this study, we want to investigate aspects of decision-making involving humans and computers.

There will be **2 Stages**. In Stage 1, you will practise a task that you will then perform in Stage 2. In Stage 2, in addition to performing the task, you will also be asked to answer some questions. You need to complete both Stages, but only your decisions in Stage 2 will affect your payment. You will receive a **participation fee of £3.00** for completing the experiment, and you can earn a **bonus of up to £2.00** depending on what happens in Stage 2.

You will also need to complete a short questionnaire at the end of the study.

Stage 1

This is **Stage 1**.

In this stage, you will perform **ten practice rounds** of the **Dot Task** so that you can familiarise yourself with the task. You have to perform these ten practice rounds, but you will receive **no feedback on your performance** and **your performance does not affect your payment**.

Here is how the **Dot Task** works:

In each round, we will show you **two circles**, Circle A and Circle B. Once you begin the task, dots will appear in both circles. Circle A and Circle B will look different in each round but, in each round, there will be **one circle that contains a higher number of dots** than the other. From now on, we will refer to this circle as **the winning circle** for that round.

In each round, you will be able to see the dots for **5 seconds** before they disappear. After the dots disappear, **your goal is to select the winning circle** for that round.

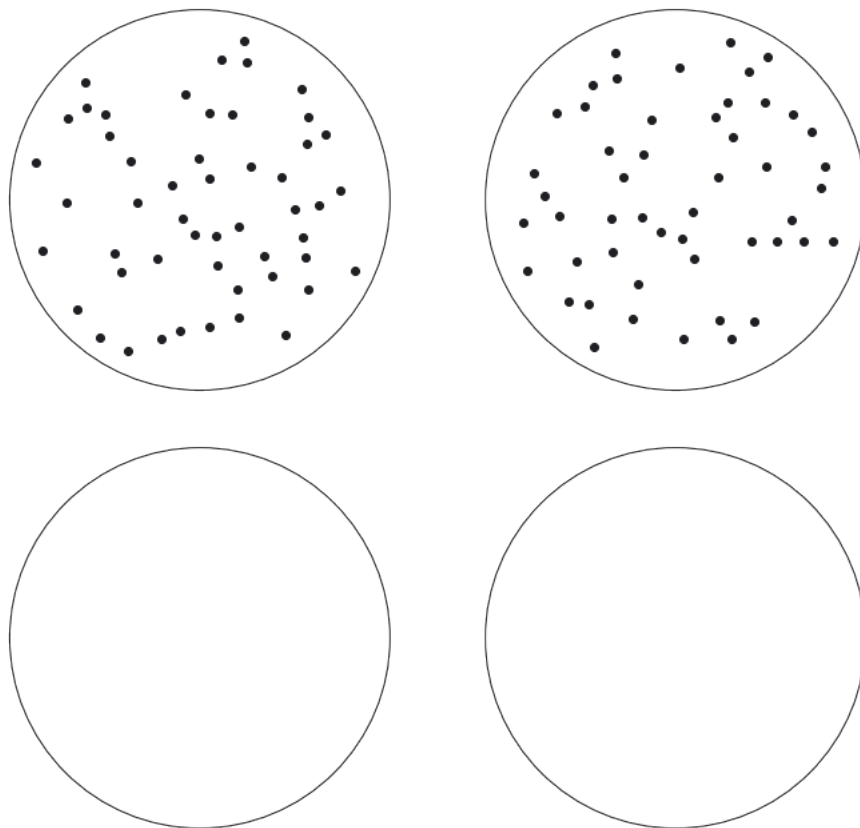
Note that the winning circle contains the same number of dots in each round and the difference between the two circles, in numbers of dots, is always the same. This means that each round of the Dot Task is of **very similar difficulty**.

To make sure that you have **read and understood** the above instructions, please answer the following question.

What is your **goal** after the dots disappear?

- o To select the circle with the lower number of dots.
- o To select the circle with the higher number of dots.

Practise Rounds



Which of the two circles is **the winning circle**?

- o Circle A
- o Circle B

Stage 2

This is **Stage 2**.

You will now perform another **ten rounds** of the **Dot Task**. As before, in each round, you will be able to see the dots for 5 seconds before they disappear and **your goal is to select the winning circle** — i.e. the one with the higher number of dots — for that round.

Recall that each round of the Dot Task is of **very similar difficulty**.

In this Stage, **your performance may affect your payment** as will be explained on the next page.

You will now perform the **Dot Task** at the same time as a **computer**. After Stage 2, **one of the ten rounds will be randomly chosen** for payment. **Both you and the computer will perform the task**, but you can decide whether you want to **rely on your own performance or that of the computer**.

If you decide to **rely on your own performance** and **you** selected the winning circle in the randomly chosen round, you will **receive a £1 bonus** for this task. If **you** did not select the winning circle, you will **receive no bonus** for this task.

If you decide to **rely on the computer's performance** and **the computer** selected the winning circle in the randomly chosen round, you will **receive a £1 bonus** for this task. If **the computer** did not select the winning circle, you will **receive no bonus** for this task.

Before we give you some information about the computer, please answer the following questions to make sure that you have **read and understood** the above instructions.

What happens if you decide to **rely on the computer's performance**?

- o Both I and the computer perform the task.
- o Only I perform the task.
- o Only the computer performs the task.

Imagine that you decide to **rely on the computer's performance**. How much would you earn as a bonus if, in the randomly chosen round, **you selected the winning circle**?

- o £0
- o £1
- o It depends on whether the computer selected the winning circle or not.

Imagine that you decide to **rely on your own performance**. How much would you earn as a bonus if, in the randomly chosen round, **you selected the winning circle**?

- o £0
- o £1
- o It depends on whether the computer selected the winning circle or not.

Treatment Groups

BASELINE

Here is some information about the computer:

The **computer uses an algorithm** in the Dot Task. The algorithm determines the **chance of success** with which the computer will, in each separate round, **select the winning circle**.

Please click on the **REVEAL** button below to reveal **the algorithm**.

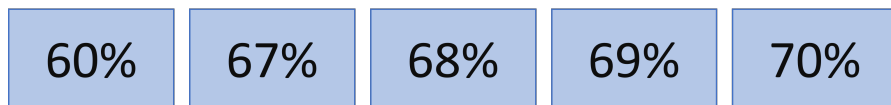


The **computer uses an algorithm** that will, in each separate round, **select the winning circle** with a **60% chance of success**.

BO-CC

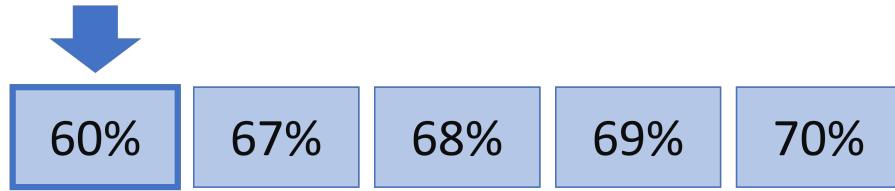
Here is some information about the computer:

The **computer controls the algorithm that it uses** in the Dot Task. Which algorithm the computer uses determines the **chance of success** with which it will, in each separate round, **select the winning circle**. There are **five algorithms for the computer to choose from**, each with a different **chance of success** as shown below:



It is possible for every one of the five algorithms to be chosen, but not all algorithms are equally likely to be chosen.

Please click on the **REVEAL** button below to reveal **the computer's choice**.

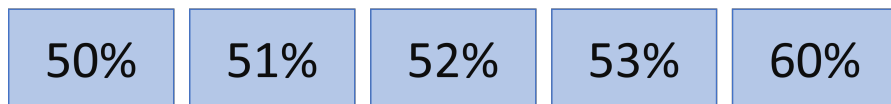


The **computer has chosen the algorithm** that will, in each separate round, **select the winning circle** with a **60% chance of success**.

GO-CC

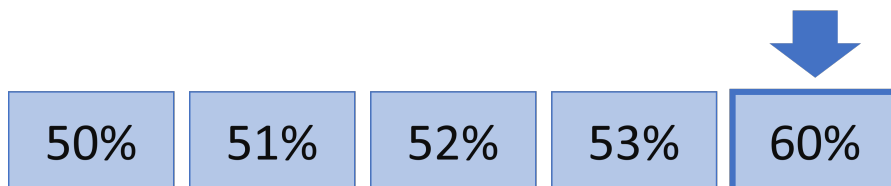
Here is some information about the computer:

The computer **controls the algorithm that it uses** in the Dot Task. Which algorithm the computer uses determines the **chance of success** with which it will, in each separate round, **select the winning circle**. There are **five algorithms** for **the computer to choose from**, each with a different **chance of success** as shown below:



It is possible for every one of the five algorithms to be chosen, but not all algorithms are equally likely to be chosen.

Please click on the **REVEAL** button below to reveal **the computer's choice**.

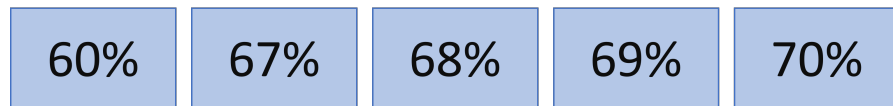


The **computer has chosen the algorithm** that will, in each separate round, **select the winning circle** with a **60% chance of success**.

BO-EC

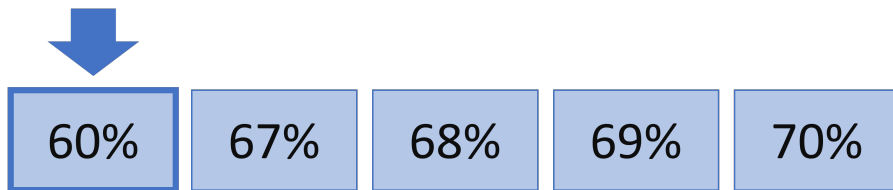
Here is some information about the computer:

The **experimenter controls the algorithm that the computer uses** in the Dot Task. Which algorithm the computer uses determines the **chance of success** with which it will, in each separate round, select **the winning circle**. There are **five algorithms for the experimenter to choose from**, each with a different **chance of success** as shown below:



It is possible for every one of the five algorithms to be chosen, but not all algorithms are equally likely to be chosen.

Please click on the **REVEAL** button below to reveal **the experimenter's choice**.

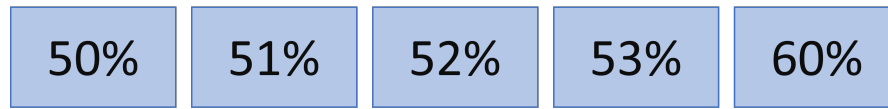


The **experimenter has chosen the algorithm** that will, in each separate round, **select the winning circle** with a **60% chance of success**.

GO-EC

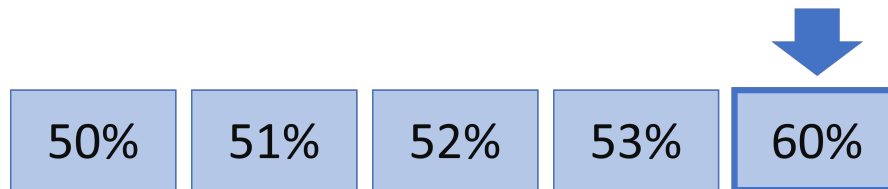
Here is some information about the computer:

The **experimenter controls the algorithm that the computer uses** in the Dot Task. Which algorithm the computer uses determines the **chance of success** with which it will, in each separate round, select **the winning circle**. There are **five algorithms for the experimenter to choose from**, each with a different **chance of success** as shown below:



It is possible for every one of the five algorithms to be chosen, but not all algorithms are equally likely to be chosen.

Please click on the **REVEAL** button below to reveal **the experimenter's choice**.



The **experimenter has chosen the algorithm** that will, in each separate round, **select the winning circle** with a **60% chance of success**.

Automation Reliance Task

Recall that:

1. **Both you and the computer** will perform the task and **one of the ten rounds will be randomly chosen** for payment.
2. If you **rely on your own performance** and **you** select the winning circle in the randomly chosen round, you will **receive a £1 bonus** for this task; if **you** do not select the winning circle is, you will **receive no bonus** for this task.
3. If you **rely on the computer's performance** and **the computer** selects the winning circle in the randomly chosen round, you will **receive a £1 bonus** for this task; if **the computer** does not select the winning circle, you will **receive no bonus** for this task.

Now please select whether you want to **rely on...**

☐ your own performance

☐ the computer's performance

Additional Measurements

In how many of the ten rounds of the **Dot Task** do you think **you** will select **the winning circle**?

Please type in your best guess as a **whole number between 0 and 10**.

Regardless of whether you have decided to rely on your own performance or that of the computer, if your guess matches the actual number of rounds in which you select the winning circle, you will **receive a £0.50 bonus** for this question. If your guess does not match the actual number of rounds in which you select the winning circle, you will **receive no bonus** for this question.

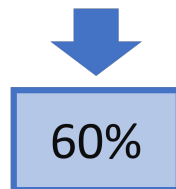
In how many of the ten rounds of the **Dot Task** do you think **the computer** will select **the winning circle**?

Please type in your best guess as a **whole number between 0 and 10**.

Regardless of whether you have decided to rely on your own performance or that of the computer, if your guess matches the actual number of rounds in which the computer selects the winning circle, you will **receive a £0.50 bonus** for this question. If your guess does not match the actual number of rounds in which the computer selects the winning circle, you will **receive no bonus** for this question.

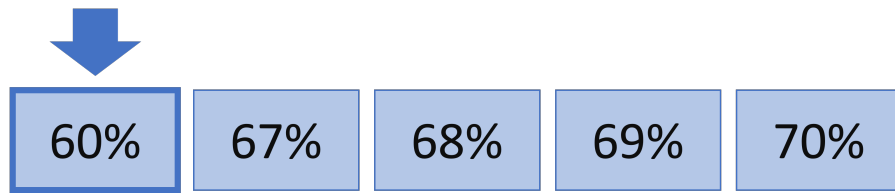
Please think back to when **the computer's algorithm** was revealed to you with the following message:

BASELINE



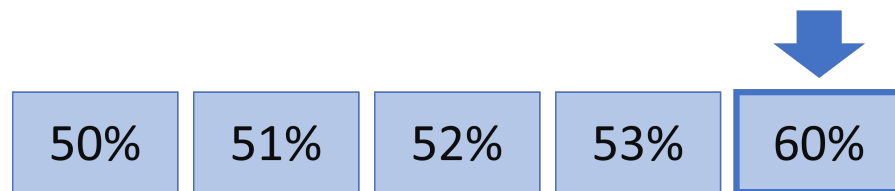
The **computer uses an algorithm** that will, in each separate round, **select the winning circle** with a **60% chance of success**.

BO-CC



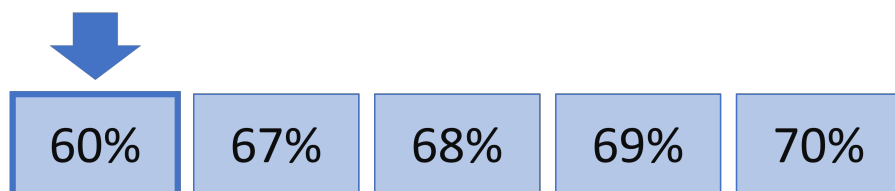
The **computer** has chosen the **algorithm** that will, in each separate round, **select the winning circle** with a **60% chance of success**.

GO-CC



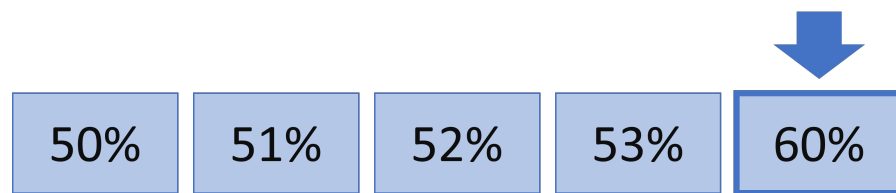
The **computer** has chosen the **algorithm** that will, in each separate round, **select the winning circle** with a **60% chance of success**.

BO-EC



The **experimenter** has chosen the **algorithm** that will, in each separate round, **select the winning circle** with a **60% chance of success**.

GO-EC



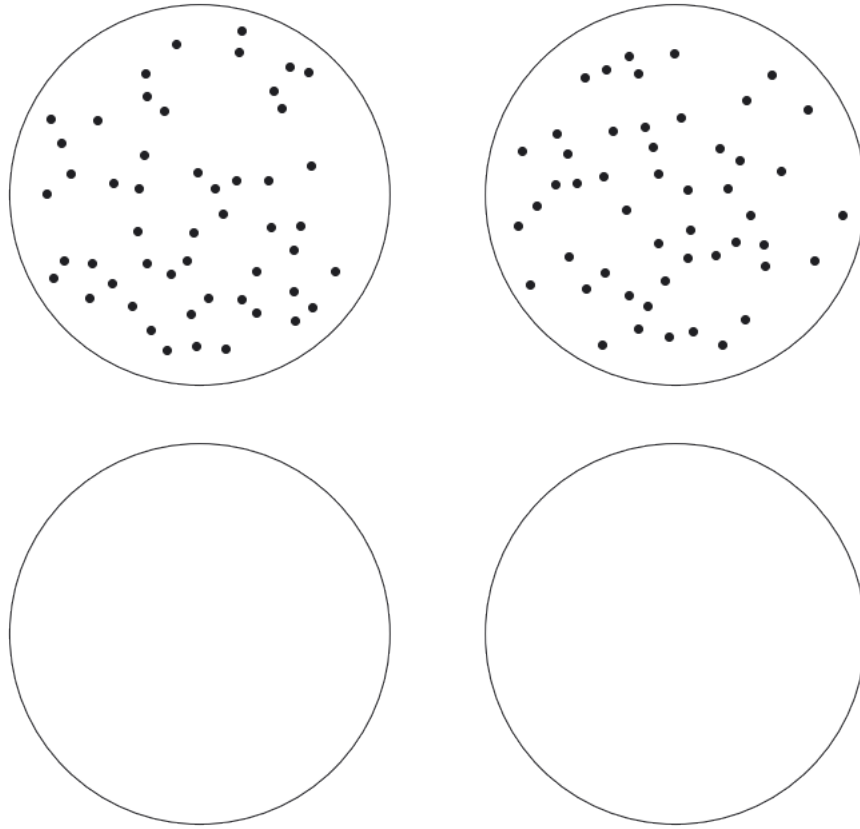
The **experimenter has chosen the algorithm** that will, in each separate round, **select the winning circle** with a **60% chance of success**.

When reading this message, to **what extent** did you experience the following **emotions**? Your answers to these questions will **not** affect your payment, but you need to answer them in order to be paid.

	not at all	a little bit	moderately	very	extremely
angry	☐	☐	☐	☐	☐
annoyed	☐	☐	☐	☐	☐
disappointed	☐	☐	☐	☐	☐
sad	☐	☐	☐	☐	☐
happy	☐	☐	☐	☐	☐
content	☐	☐	☐	☐	☐
grateful	☐	☐	☐	☐	☐
calm	☐	☐	☐	☐	☐

To **what extent** would you say that the following **descriptions** apply to the algorithm? Your answers to these questions will **not** affect your payment, but you need to answer them in order to be paid.

	not at all	a little bit	moderately	very	extremely
kind	☐	☐	☐	☐	☐
fair	☐	☐	☐	☐	☐
friendly	☐	☐	☐	☐	☐
generous	☐	☐	☐	☐	☐
mean	☐	☐	☐	☐	☐
unfair	☐	☐	☐	☐	☐
hostile	☐	☐	☐	☐	☐
uncharitable	☐	☐	☐	☐	☐

Dot Task

Which of the two circles is **the winning circle**?

☐ Circle A

☐ Circle B

Ambiguity Preferences

Ambiguity Preference Survey Module (Cavatorta and Schroeder, 2019):

Please imagine the following situation: You can choose between drawing a ball from two different urns, urn A and urn B. Urn A contains 100 balls, some are white and some are black. However, you don't know how many balls are white and how many balls are black. Any combination is possible. There might be from 0 to 100 white balls, with the remaining balls being black. Put differently, you do not know the probability of drawing a white or a black ball. Urn B contains 100 balls as well, but you know the exact number of white and black balls in this urn. In other words, you know the exact probability of drawing a white or a black ball. Now choose a color, either white or black. Suppose you win \$100 if you draw a ball of the color you have selected. If the ball is of the other color, you win nothing.

Please choose the color of the ball that provides you \$100:

- ☐ white
- ☐ black

We present you now several situations.³ The composition of urn A is identical in each situation¹⁷. The composition of urn B is different in each situation.

1. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 45 white balls and 55 black balls, i.e., there is a 45% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

- (a) Urn A → go to situation 2
- (b) Urn B → go to situation 3

2. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 65 white balls and 35 black balls, i.e., there is a 65% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

- (a) Urn A → go to situation 4
- (b) Urn B → go to situation 5

3. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 25 white balls and 75 black balls, i.e., there is a 25% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

- (a) Urn A → go to situation 6
- (b) Urn B → go to situation 7

4. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 75 white balls and 25 black balls, i.e., there is a 75% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

- (a) Urn A → go to situation 8
- (b) Urn B → go to situation 9

¹⁷The actual situations presented to the subjects depend on the color chosen. Here, we assume that the selected color is white. If the selected color is black, the word white has to be replaced with black, and vice versa.

5. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 55 white balls and 45 black balls, i.e., there is a 55% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

(a) Urn A $\rightarrow$ go to situation 10

(b) Urn B $\rightarrow$ go to situation 11

6. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 35 white balls and 65 black balls, i.e., there is a 35% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

(a) Urn A $\rightarrow$ go to situation 12

(b) Urn B $\rightarrow$ go to situation 13

7. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 15 white balls and 85 black balls, i.e., there is a 15% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

(a) Urn A $\rightarrow$ go to situation 14

(b) Urn B $\rightarrow$ go to situation 15

8. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 80 white balls and 20 black balls, i.e., there is a 80% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

(a) Urn A $\rightarrow$ go to situation 16

(b) Urn B $\rightarrow$ stop

9. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 70 white balls and 30 black balls, i.e., there is a 70% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

(a) Urn A $\rightarrow$ stop

(b) Urn B $\rightarrow$ stop

10. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 60 white balls and 40 black balls, i.e., there is a 60% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

(a) Urn A $\rightarrow$ stop

(b) Urn B $\rightarrow$ stop

11. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 50 white balls and 50 black balls, i.e., there is a 50% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

(a) Urn A $\rightarrow$ stop

(b) Urn B $\rightarrow$ stop

12. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 40 white balls and 60 black balls, i.e., there is a 40% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

(a) Urn A $\rightarrow$ stop

(b) Urn B $\rightarrow$ stop

13. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 30 white balls and 70 black balls, i.e., there is a 30% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

(a) Urn A $\rightarrow$ stop

(b) Urn B $\rightarrow$ stop

14. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 20 white balls and 80 black balls, i.e., there is a 20% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

(a) Urn A $\rightarrow$ stop

(b) Urn B $\rightarrow$ stop

15. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 10 white balls and 90 black balls, i.e., there is a 10% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

(a) Urn A $\rightarrow$ stop

(b) Urn B $\rightarrow$ go to situation 17

16. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 90 white balls and 10 black balls, i.e., there is a 90% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

(a) Urn A $\rightarrow$ stop

(b) Urn B $\rightarrow$ stop

17. What would you prefer? Drawing a ball from urn A with unknown composition, i.e., you do not know the probability of drawing a white ball or drawing a ball from urn B in which there are 5 white balls and 95 black balls, i.e., there is a 5% chance of drawing a white ball? Remember, you win \$100 if you draw a white ball, and nothing otherwise.

(a) Urn A $\rightarrow$ stop

(b) Urn B $\rightarrow$ stop

Risk Preferences

General Risk Question (Dohmen et al. 2011):

How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?

Please tick a box on the scale, where the value 0 means: ‘not at all willing to take risks’ and the value 10 means: ‘very willing to take risks’.

o o o o o o o o o o o

0

10

Hypothetical Willingness To Pay

In Stage 2, we asked you whether you wanted to **rely on your own performance or that of the computer**. You chose to rely on **the computer’s performance**.

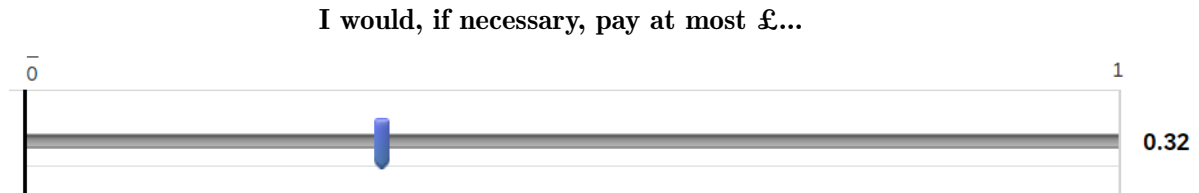
Imagine facing the same choice again, except now you have to **pay to rely on the computer’s performance** rather than your own. Recall that if you were to **rely on the computer’s performance** and **the computer** selected the winning circle in the randomly chosen round, you would **receive a £1 bonus**.

Using the slider below, please indicate the **maximum amount of money** that, if necessary, you would **pay to rely on the computer’s performance**. Note that this is a **hypothetical task**.

In Stage 2, we asked you whether you wanted to **rely on your own performance or that of the computer**. You chose to rely on **your own performance**.

Imagine facing the same choice again, except now you have to **pay to rely on your own performance** rather than that of the computer. Recall that if you were to **rely on your own performance** and **you** selected the winning circle in the randomly chosen round, you would **receive a £1 bonus**.

Using the slider below, please indicate the **maximum amount of money** that, if necessary, you would **pay to rely on your own performance**. Note that this is a **hypothetical task**.



Finally, we are interested to know **why you chose to rely on the computer’s performance** rather than your own. Please explain your reasoning in as much detail as possible.

Finally, we are interested to know **why you chose to rely on your own performance** rather than that of the computer. Please explain your reasoning in as much detail as possible.

Demographic Questions

Gender

Age

Feedback and Comments

Thank you for your participation in this study.

If you’re interested, here is a summary of **your performance in Stage 2**:
You solved x out of 10 rounds correctly.

Please let us know if you have any **comments**.

C Appendix for Chapter 3

C.1 Additional Analysis

	TOTAL	PERCENTAGE
GENDER		
FEMALE	184	48.81
MALE	192	50.93
RATHER NOT SAY	1	0.27
AGE		
MIN	20	
MAX	77	
MEAN	43.57	
NATIONALITY		
UK	377	100.00
HIGHEST LEVEL OF EDUCATION COMPLETED		
NO FORMAL QUALIFICATIONS	2	0.53
SECONDARY EDUCATION (E.G. GED/GCSE)	34	9.02
HIGH SCHOOL DIPLOMA/A-LEVELS	60	15.92
TECHNICAL/COMMUNITY COLLEGE	41	10.88
UNDERGRADUATE DEGREE (BA/BSc/OTHER)	157	41.64
GRADUATE DEGREE (MA/MSc/MPhil/OTHER)	68	18.04
DOCTORATE DEGREE (PhD/OTHER)	15	3.98
PROGRAMMING SKILLS		
YES	97	25.73
No	280	74.27

Table C.1: Demographic Sample Composition

	EV-F N=93	EV-NF N=94	RDU-NF N=95	RDU-F N=95
FEMALE SUBJECTS	45.61	51.06	52.63	46.32
MEAN AGE	43.72	43.29	43.27	44.01
PROGRAMMING SKILLS	22.58	26.60	22.11	31.58
ALPHA	1.43	1.38	1.38	1.40
BETA	0.62	0.50	0.50	0.53
GAMMA	0.58	0.54	0.48	0.58

Note: FEMALE SUBJECTS and PROGRAMMING SKILLS represent within-treatment percentages, all other values are treatment averages.

Table C.2: Demographic Treatment Composition

	AR ₁	AR ₂	p (H ₀ : AR ₁ = AR ₂) <i>Two-Sample Test of Proportion</i>
MENU A	0.43		
EV-NF – EV-F	0.63	0.46	0.151
RDU-NF – RDU-F	0.33	0.31	0.7549
EV-NF – RDU-NF	0.63	0.33	0.0000
EV-F – RDU-F	0.46	0.31	0.0312
MENU B	0.46		
EV-NF – EV-F	0.61	0.52	0.2061
RDU-NF – RDU-F	0.40	0.31	0.1718
EV-NF – RDU-NF	0.61	0.40	0.0035
EV-F – RDU-F	0.52	0.31	0.0026
MENU C	0.44		
EV-NF – EV-F	0.56	0.51	0.5061
RDU-NF – RDU-F	0.32	0.37	0.4445
EV-NF – RDU-NF	0.56	0.32	0.0008
EV-F – RDU-F	0.51	0.37	0.0489
MENU D	0.46		
EV-NF – EV-F	0.65	0.46	0.0099
RDU-NF – RDU-F	0.41	0.35	0.3696
EV-NF – RDU-NF	0.65	0.41	0.0013
EV-F – RDU-F	0.46	0.35	0.1228
MENU E	0.45		
EV-NF – EV-F	0.58	0.50	0.2686
RDU-NF – RDU-F	0.41	0.32	0.1746
EV-NF – RDU-NF	0.58	0.41	0.0197
EV-F – RDU-F	0.50	0.32	0.0100
MENU F	0.46		
EV-NF – EV-F	0.62	0.45	0.0153
RDU-NF – RDU-F	0.43	0.36	0.2988
EV-NF – RDU-NF	0.62	0.43	0.0084
EV-F – RDU-F	0.45	0.36	0.2126
MENU G	0.42		
EV-NF – EV-F	0.54	0.53	0.9375
RDU-NF – RDU-F	0.36	0.24	0.0816
EV-NF – RDU-NF	0.54	0.36	0.0132
EV-F – RDU-F	0.53	0.24	0.0000
MENU H	0.46		
EV-NF – EV-F	0.63	0.47	0.0222
RDU-NF – RDU-F	0.42	0.34	0.2316
EV-NF – RDU-NF	0.63	0.42	0.0034
EV-F – RDU-F	0.47	0.34	0.0658

Table C.3: Menu-by-Menu Treatment Comparisons of Automation Reliance

	AUTOMATION RELIANCE		
	(1)	(2)	(3)
RDU ALGORITHM	−0.24*** (0.07)	−0.21*** (0.07)	−0.20*** (0.07)
FEEDBACK	−0.06 (0.07)	−0.04 (0.07)	−0.04 (0.07)
RDU ALGORITHM × FEEDBACK	−0.03 (0.09)	−0.09 (0.09)	−0.08 (0.10)
FINAL ROUND SATISFACTION	−0.00	−0.00 (0.02)	0.00 (0.02)
ALPHA		−0.00 (0.03)	−0.01 (0.03)
BETA		0.04 (0.05)	0.05 (0.07)
GAMMA		0.18*** (0.06)	0.20*** (0.07)
CONSTANT	0.47*** (0.05)	0.36*** (0.10)	0.12 (0.17)
Demographic Controls	No	No	Yes
Observations	377	367	340
Prob > F	0.0000	0.0000	0.0001
R-squared	0.0810	0.01112	0.01128
Root MSE	0.4479	0.4421	0.4460

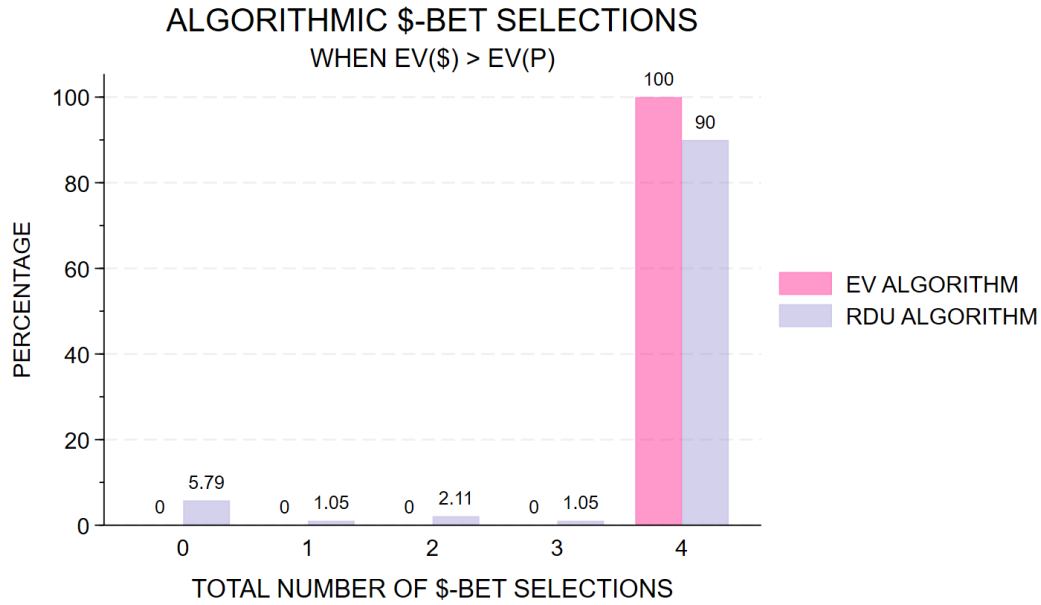
Note: Dependent variable AUTOMATION RELIANCE is a dummy that takes the value of 1 for subjects who chose to have the computer algorithm select the lottery for them in the hypothetical continuation of the lottery selection task. Demographic controls include gender, age, education, income, and programming skills. Standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.4: OLS Regressions: Hypothetical Automation Reliance

	MYSELF	COMPUTER
(1) DON'T TRUST COMPUTER/DON'T TRUST MYSELF	0.19	0.04
EV-NF	0.10	0.00
EV-F	0.16	0.05
RDU-NF	0.23	0.05
RDU-F	0.23	0.14
(2) CAN DO A BETTER JOB THAN COMPUTER/ COMPUTER CAN DO A BETTER JOB THAN ME	0.57	0.61
EV-NF	0.43	0.68
EV-F	0.38	0.67
RDU-NF	0.64	0.50
RDU-F	0.70	0.43
(3) DON'T LIKE RELYING ON ANYONE BUT MYSELF/ DON'T LIKE RELYING ON MYSELF	0.22	0.05
EV-NF	0.22	0.02
EV-F	0.24	0.05
RDU-NF	0.21	0.09
RDU-F	0.22	0.07
(4) WANT TO SEE HOW WELL I CAN DO/ WANT TO SEE HOW WELL COMPUTER CAN DO	0.47	0.55
EV-NF	0.53	0.50
EV-F	0.58	0.56
RDU-NF	0.36	0.55
RDU-F	0.47	0.64
(5) DID NOT UNDERSTAND PART 1	0.00	-
(6) OTHER	0.12	0.15

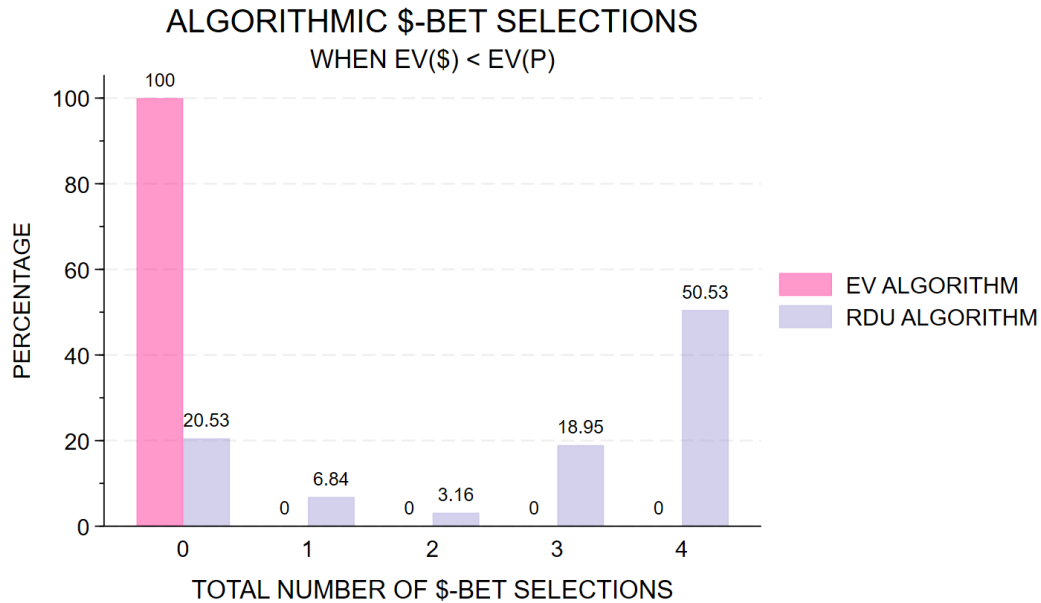
Note: Subjects are divided into those who expressed a preference for self-reliance (N=258, 68.44%) and those who expressed a preference for automation reliance (N=119, 31.56%), and subsequently saw differing versions of reasons (1) – (2). Reason (5) was shown only to subjects in RDU-NF and RDU-F who expressed a preference for self-reliance as it was not applicable otherwise. Each subject had to select at least one reason, but could select multiple reasons.

Table C.5: Reasons for Decision in Hypothetical Continuation



Note: Distribution of the number of \$-Bet selections by the EV algorithm and the RDU algorithm in the four lottery menus where the expected value of the \$-Bet was higher than that of the P-Bet.

Figure C.1: Number of Algorithmic \$-Bet Selections: $EV(\$-Bet) > EV(P-Bet)$



Note: Distribution of the number of \$-Bet selections by the EV algorithm and the RDU algorithm in the four lottery menus where the expected value of the \$-Bet was lower than that of the P-Bet.

Figure C.2: Number of Algorithmic \$-Bet Selections: $EV(\$-Bet) < EV(P-Bet)$

C.2 Instructions

C.2.1 Part 1: RDU Estimation

Overview

Thank you very much for participating in PART 1 of our decision-making study.

It should take no more than **20 minutes**, for which you will receive a **participation fee of £2.00**. In addition to this, you can earn a **bonus of up to £6.00** depending on your choices and on chance. Note, however, that this study consists of **two separate parts**: if you complete this part – **Part 1** – today, you will be invited **three days from today** to complete **Part 2**. Both parts are similar in terms of duration and possible earnings, but you will have to complete Part 2 within 24 hours of receiving your invitation and **any bonus you earn in Part 1 will be paid out only upon completion of Part 2**.

In **Part 1** of the study, you will be asked to complete **25 tasks**.

You will receive a **participation fee of £2.00** for completing **all 25 tasks**, and you can earn a **bonus of up to £6.00** depending on your choices in these tasks and on chance. Remember that any bonus you earn in Part 1 will be paid out only upon completion of Part 2, to which you will be invited three days from today.

Please note that your choices in Part 1 might affect what happens in Part 2. It is therefore in your best interest to **pay attention and think carefully** about your choices.

Instructions

In each of the 25 tasks, you will be **presented with a table**. Each table consists of **20 rows**, each of which describes a choice between **two options A or B** that can earn you **Points**. You will receive a **participation fee of £2.00** for completing all 25 tasks, but you can earn a **bonus payment of up to £6.00** depending on the amount of **Points** you earn. Points you earn will be converted to money payment using an exchange rate of **50 Points = £1.00**.

Option A is always a two-outcome lottery and it is **always the same for every row** within a table. **Option B** is always a **guaranteed amount of money** which is **always different for every row** within a table. All 25 tables follow this pattern, but the lottery and the guaranteed amounts will be different for each table, so it is important that you pay attention to each individual table.

For every row of the table, you must **make a choice between Option A and Option B**. You should consider which of these two options you really would prefer for each row, then mark your choice by selecting the corresponding button. All the tables are completely independent of each other, and a choice you make in one table does not affect any of the other tables.

Below, you can see the first few rows of an example table. Here, Row 1 is a choice between **Option A which gives a 90% chance of getting 150 Points and a 10% chance of getting 50 Points** or **Option B which is a guaranteed amount of 150 Points**. Since the guaranteed payoff for Option B in Row 1 is as good as the large prize of the lottery, we think that you will want to choose Option B for this row.

However, since Option B gets progressively worse as you move down the rows while Option A stays exactly the same, there may come a row where the guaranteed payoff of Option B is sufficiently small that you prefer the lottery. **If you find such a row, you should then choose Option A for that row and the rows below it since Option B continues to get worse all the way down the table.** This means that, in any given table, **you should switch at most once and only from Option B to Option A** but not the other way around.

	A	B	
	LOTTERY	GUARANTEED AMOUNT	
90% CHANCE OF 150 & 10% CHANCE OF 50	☐	☐	150
90% CHANCE OF 150 & 10% CHANCE OF 50	☐	☐	145
90% CHANCE OF 150 & 10% CHANCE OF 50	☐	☐	140
90% CHANCE OF 150 & 10% CHANCE OF 50	☐	☐	135
90% CHANCE OF 150 & 10% CHANCE OF 50	☐	☐	130
. . .	. . .	. . .	. . .

Payment

Your **Points** will be determined based on **one of the tables** you have completed, which will be **randomly chosen** at the end of the study. Once the table has been selected, one row of that table will also be randomly chosen.

If you have chosen **Option B** on the selected row, you will be awarded the **guaranteed amount** of Points for that row. If you have chosen **Option A**, then the lottery will be played out by **randomly drawing a whole number between 1 and 100**. For example, if the chance of winning of the selected lottery is 90%, then you will get the large prize if a number from 1 to 90 is drawn, and get the small prize if a number from 91 to 100 is drawn.

This payment structure has been designed so that you have **an incentive to think carefully about each and every choice** you have to make. Each time you have to choose between two options, **the best thing for you to do is to choose the option you prefer**.

Comprehension

To make sure that you have **read and understood** these instructions, please answer the following **three comprehension questions**.

Which of the following statements is **true for Option A** within a table?

- o Option A is always a lottery and it is always the same for every row
- o Option A is always a lottery and it is always different for every row
- o Option A is always a guaranteed amount of points which is always different for every row

If you find a row where the guaranteed payoff of **Option B** is **sufficiently small that you prefer the lottery of Option A**, you should...

- o Choose Option B for that row and the rows below it
- o Choose Option A for that row and go back to Option B for the rows below it
- o Choose Option A for that row and the rows below it

After you complete all 25 tables, your **Points** will be determined based on...

- o One randomly chosen row of one randomly chosen table
- o All 25 tables
- o Every row of one randomly chosen table

RDU Elicitation

25 Tables

Demographic Questions

Gender

Age

Comments

Thank you for completing **Part 1** of our study.

As explained earlier, it is **very important** for our research that you also complete **Part 2**, to which you will be invited **three days from today** via Prolific. Any **bonus** that you may have earned in this part of the study will also be paid out **only upon completion of Part 2**.

In the meantime, please let us know if you have any **comments or thoughts** that you would like to share with us.

C.2.2 Part 2: Main Experiment*Overview*

Thank you very much for participating in PART 2 of our decision-making study.

It should take no more than **10 minutes**, for which you will receive a **participation fee of £1.00**. In addition to this, you can earn a **bonus of up to £4.50** depending on your choices and on chance.

In **Part 2** of the study, you will be asked to complete **8 tasks**.

You will receive a **participation fee of £1.00** for completing all **8 tasks**, and you can earn a **bonus of up to £4.50** depending on your choices in these tasks and on chance.

Instructions

In each task, you will be **presented with a list of lotteries**. Each list consists of **8 lotteries** and you can earn **Points** depending on which lottery is selected and on chance. As in Part 1, Points you earn will be converted to money payment using an exchange rate of **50 Points = £1.00**.

Every one of the **8 lotteries labelled A – H** is comprised of **10 outcomes**. Which outcome you get for any given lottery will depend on a **fair 10-sided die** that is rolled after every task. In other words, all outcomes are **equally likely no matter which lottery is selected**. All lists follow this pattern, but the 8 lotteries will be different for each list.

Below, you can see the first few lotteries of an example list. Here, if Lottery A were selected, you would win **0 Points if the die lands on 1, 175 Points if the die lands on 2** and so on. Alternatively, if Lottery B were selected, you would win **65 Points if the die lands on 1, 60 Points if the die lands on 2** and so on.

Your **Points** will be determined based on **one of the tasks** you have completed, which will be **randomly chosen** at the end of the study. To illustrate this, imagine that the example below were the task that was randomly chosen for payment. If, in this task, Lottery C were selected and the die landed on 5, you would earn 0 Points. If, on the other hand, Lottery D were selected and the die landed on 6, you would earn 180 Points and so on.

	DIE ROLL	DIE ROLL	DIE ROLL	DIE ROLL	DIE ROLL	DIE ROLL	DIE ROLL	DIE ROLL	DIE ROLL	DIE ROLL
	1	2	3	4	5	6	7	8	9	10
LOTTERY A	0	175	0	0	0	185	0	180	0	0
LOTTERY B	65	60	0	60	0	70	70	0	60	70
LOTTERY C	65	65	0	65	0	65	65	0	65	65
LOTTERY D	0	180	0	0	0	180	0	180	0	0
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.

Before you are shown the list of lotteries in each task, you are given **two options for how the lottery will be selected**:

- (1) select the lottery yourself
- (2) have a computer select the lottery for you

If you pick (1), the list of lotteries will be **visible for 10 seconds** before it disappears. After that, **you have 10 seconds to select a lottery**.

If you pick (2), the list of lotteries will be **visible for 10 seconds** before it disappears. After that, **the computer takes 10 seconds to select a lottery**.

In other words, **the only difference between the two options is whether you select the lottery or whether the computer selects it for you.** You will decide whether it is you or the computer that makes the selection before every task, and your decision in one task will not affect any other task.

Treatment Groups

EV-NF

If, for a given task, you decide to have the computer select a lottery for you, its choice will be determined by an algorithm: **this algorithm uses mathematics to work out how much each lottery would yield on average, and selects the lottery for which this is highest.**

After completing a task, regardless of whether you decided to select the lottery yourself or have the computer select it for you, **you will be shown the selected lottery.**

Lastly, you will be asked after every task to rate your satisfaction with the lottery selection for that task on a scale of 1 to 5 stars, where **1 star means "not satisfied at all" and 5 stars means "completely satisfied".**

To make sure that you have **read and understood** these instructions, please answer the following **three comprehension questions.**

If you decide to **have the computer select a lottery** for you in a given task, what does the computer algorithm do?

- o It selects a lottery at random
- o It uses mathematics to work out how much each lottery would yield on average, and selects the lottery for which this is highest

After completing a task...

- o I will be shown the selected lottery
- o I will not be given any information

After you complete all 8 tasks, your **Points** will be determined based on the outcome(s) of...

- o All 8 tasks
- o One randomly chosen task

EV-F

If, for a given task, you decide to have the computer select a lottery for you, its choice will be determined by an algorithm: **this algorithm uses mathematics to work out how much each lottery would yield on average, and selects the lottery for which this is highest.**

After completing a task, regardless of whether you decided to select the lottery yourself or have the computer select it for you, **you will be shown the selected lottery, the outcome of the die roll, and the Points you will earn if this task is chosen for payment.**

Lastly, you will be asked after every task to rate your satisfaction with the lottery selection for that task on a scale of 1 to 5 stars, where **1 star means "not satisfied at all" and 5 stars means "completely satisfied".**

To make sure that you have **read and understood** these instructions, please answer the following **three comprehension questions.**

If you decide to **have the computer select a lottery** for you in a given task, what does the computer algorithm do?

- o It selects a lottery at random
- o It uses mathematics to work out how much each lottery would yield on average, and selects the lottery for which this is highest

After completing a task...

- o I will be shown the selected lottery, the outcome of the die roll, and the points I will earn if this task is chosen for payment
- o I will not be given any information

After you complete all 8 tasks, your **Points** will be determined based on the outcome(s) of...

- o All 8 tasks
- o One randomly chosen task

RDU-NF

If, for a given task, you decide to have the computer select a lottery for you, its choice will be determined by an algorithm: **this algorithm uses modern decision science and your choices in Part 1 of the study to model your behaviour, and selects the lottery that it predicts you would select.**

After completing a task, regardless of whether you decided to select the lottery yourself or have the computer select it for you, **you will be shown the selected lottery.**

Lastly, you will be asked after every task to rate your satisfaction with the lottery selection for that task on a scale of 1 to 5 stars, where **1 star means "not satisfied at all" and 5 stars means "completely satisfied".**

To make sure that you have **read and understood** these instructions, please answer the following **three comprehension questions.**

If you decide to **have the computer select a lottery** for you in a given task, what does the computer algorithm do?

- o It selects a lottery at random
- o It uses modern decision science and my choices in Part 1 to model my behaviour, and selects the lottery that it predicts I would select

After completing a task...

- o I will be shown the selected lottery
- o I will not be given any information

After you complete all 8 tasks, your **Points** will be determined based on the outcome(s) of...

- o All 8 tasks
- o One randomly chosen task

RDU-F

If, for a given task, you decide to have the computer select a lottery for you, its choice will be determined by an algorithm: **this algorithm uses modern decision science and your choices in Part 1 of the study to model your behaviour, and selects the lottery that it predicts you would select.**

After completing a task, regardless of whether you decided to select the lottery yourself or have the computer select it for you, **you will be shown the selected lottery, the outcome of the die roll, and the Points you will earn if this task is chosen for payment.**

Lastly, you will be asked after every task to rate your satisfaction with the lottery selection for that task on a scale of 1 to 5 stars, where **1 star means "not satisfied at all" and 5 stars means "completely satisfied".**

To make sure that you have **read and understood** these instructions, please answer the following **three comprehension questions.**

If you decide to **have the computer select a lottery** for you in a given task, what does the computer algorithm do?

- o It selects a lottery at random
- o It uses modern decision science and my choices in Part 1 to model my behaviour, and selects the lottery that it predicts I would select

After completing a task...

- o I will be shown the selected lottery, the outcome of the die roll, and the points I will earn if this task is chosen for payment
- o I will not be given any information

After you complete all 8 tasks, your **Points** will be determined based on the outcome(s) of...

- o All 8 tasks
- o One randomly chosen task

Automation Reliance Task

EV-NF

Remember that:

- (1) *The computer algorithm uses mathematics to work out how much each lottery would yield on average, and selects the lottery for which this is highest.*
- (2) *After completing a task, regardless of whether you decided to select the lottery yourself or have the computer select it for you, you will be shown the selected lottery.*

In this task, I want the lottery to be selected by...

- o myself
- o the computer

Lottery Selection + No Feedback

EV-F

Remember that:

- (1) *The computer algorithm uses mathematics to work out how much each lottery would yield on average, and selects the lottery for which this is highest.*
- (2) *After completing a task, regardless of whether you decided to select the lottery yourself or have the computer select it for you, you will be shown the selected lottery, the outcome of the die roll, and the Points you will earn if this task is chosen for payment.*

In this task, I want the lottery to be selected by...

☐ myself ☐ the computer

Lottery Selection + Feedback

RDU-NF

Remember that:

(1) The computer algorithm uses modern decision science and your choices in Part 1 of the study to model your behaviour, and selects the lottery that it predicts you would select.

(2) After completing a task, regardless of whether you decided to select the lottery yourself or have the computer select it for you, you will be shown the selected lottery.

In this task, I want the lottery to be selected by...

☐ myself ☐ the computer

Lottery Selection + No Feedback

RDU-F

Remember that:

(1) The computer algorithm uses modern decision science and your choices in Part 1 of the study to model your behaviour, and selects the lottery that it predicts you would select.

(2) After completing a task, regardless of whether you decided to select the lottery yourself or have the computer select it for you, you will be shown the selected lottery, the outcome of the die roll, and the Points you will earn if this task is chosen for payment.

In this task, I want the lottery to be selected by...

☐ myself ☐ the computer

Lottery Selection + Feedback

Hypothetical Continuation

Thank you for completing the 8 tasks.

Before you finish the study, we would like you to think about the following **hypothetical scenario**.

Imagine that you had to do **another 8 tasks** like the ones you have just completed. This time, however, you do not get to choose for every individual task whether you want to select the lottery yourself or have the computer select it for you, but you have to **choose one of these two options that will be applied to all 8 tasks**.

Your answer to this question will not affect your payment, we are simply interested in which option you would prefer.

I would want the lottery to be selected by...

o myself o the computer

... in all 8 tasks.

EV-NF & EV-F

Please tell us a bit more about **why you would prefer to select the lottery in all 8 tasks yourself** rather than have the computer select it for you. Your answer to this question will also not affect your payment, so please answer honestly and tick all reasons that apply.

- o I don't trust the computer
- o I think I can do a better job than the computer
- o I don't like relying on anyone but myself
- o I would want to see how well I can do
- o Other

Please tell us a bit more about **why you would prefer to have the computer select the lottery in all 8 tasks** rather than do it yourself. Your answer to this question will also not affect your payment, so please answer honestly and tick all reasons that apply.

- o I don't trust myself
- o I think the computer can do a better job than me
- o I don't like relying on myself
- o I would want to see how well the computer can do
- o Other

RDU-NF & RDU-F

Please tell us a bit more about **why you would prefer to select the lottery in all 8 tasks yourself** rather than have the computer select it for you. Your answer to this question will also not affect your payment, so please answer honestly and tick all reasons that apply.

- ☐ I don't trust the computer
- ☐ I think I can do a better job than the computer
- ☐ I didn't understand Part 1/I did not take Part 1 seriously
- ☐ I don't like relying on anyone but myself
- ☐ I would want to see how well I can do
- ☐ Other

Please tell us a bit more about **why you would prefer to have the computer select the lottery in all 8 tasks** rather than do it yourself. Your answer to this question will also not affect your payment, so please answer honestly and tick all reasons that apply.

- ☐ I don't trust myself
- ☐ I think the computer can do a better job than me
- ☐ I don't like relying on myself
- ☐ I would want to see how well the computer can do
- ☐ Other

Comments

Thank you for completing **Part 2** of our study.

Please let us know if you have any **comments or thoughts** that you would like to share with us.