



UNIL | Université de Lausanne

Unicentre
CH-1015 Lausanne
<http://serval.unil.ch>

Year: 2024

The relationship between individual behaviour and social norm change

von Flüe Lukas

von Flüe Lukas, 2024, The relationship between individual behaviour and social norm change

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive <http://serval.unil.ch>

Document URN : [urn:nbn:ch:serval-BIB_BA1DBB34D3340](http://nbn:ch:serval-BIB_BA1DBB34D3340)

Droits d'auteur

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.



UNIL | Université de Lausanne

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES
DÉPARTEMENT DE COMPORTEMENT ORGANISATIONNEL

**THE RELATIONSHIP BETWEEN INDIVIDUAL
BEHAVIOUR AND SOCIAL NORM CHANGE**

THÈSE DE DOCTORAT

présentée à la

Faculté des Hautes Études Commerciales
de l'Université de Lausanne

pour l'obtention du grade de
Doctorat en Management

par

Lukas Emanuel von FLÜE

Directrice de thèse
Prof. Sonja Vogt

Co-directeur de thèse
Prof. Charles Efferson

Jury

Prof. Valérie Chavez-Demoulin, Présidente
Prof. Christian Zehnder, expert interne
Prof. Heinrich Nax, expert externe

LAUSANNE
2024



UNIL | Université de Lausanne

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES
DÉPARTEMENT DE COMPORTEMENT ORGANISATIONNEL

**THE RELATIONSHIP BETWEEN INDIVIDUAL
BEHAVIOUR AND SOCIAL NORM CHANGE**

THÈSE DE DOCTORAT

présentée à la

Faculté des Hautes Études Commerciales
de l'Université de Lausanne

pour l'obtention du grade de
Doctorat en Management

par

Lukas Emanuel von FLÜE

Directrice de thèse
Prof. Sonja Vogt

Co-directeur de thèse
Prof. Charles Efferson

Jury

Prof. Valérie Chavez-Demoulin, Présidente
Prof. Christian Zehnder, expert interne
Prof. Heinrich Nax, expert externe

LAUSANNE
2024

IMPRIMATUR

La Faculté des hautes études commerciales de l'Université de Lausanne autorise l'impression de la thèse de doctorat rédigée par

Lukas Emanuel VON FLÜE

intitulée

The Relationship Between Individual Behaviour and Social Norm Change

sans se prononcer sur les opinions exprimées dans cette thèse.

Lausanne, le 17.09.2024



Professeure Marianne Schmid Mast, Doyenne



Thesis committee

Prof. Sonja Vogt
University of Lausanne
Thesis supervisor

Prof. Charles Efferson
University of Lausanne
Thesis co-supervisor

Prof. Christian Zehnder
University of Lausanne
Internal expert

Prof. Heinrich Nax
University of Zürich
External expert

University of Lausanne
Faculty of Business and Economics

Ph.D. in Management

I hereby certify that I have examined the doctoral thesis of

Lukas Emanuel von FLÜE

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature:



Date: 25 August 2024

Prof. Sonja VOGT
Thesis supervisor

University of Lausanne
Faculty of Business and Economics

Ph.D. in Management

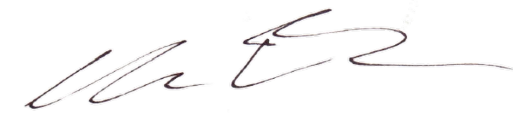
I hereby certify that I have examined the doctoral thesis of

Lukas Emanuel von FLÜE

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature: _____



Date: 22 August 2024

Prof. Charles EFFERSON
Co-supervisor

University of Lausanne
Faculty of Business and Economics

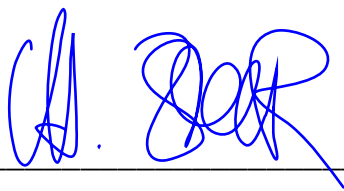
Ph.D. in Management

I hereby certify that I have examined the doctoral thesis of

Lukas Emanuel von FLÜE

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature:  Date: August 21, 2024

Prof. Christian ZEHNDER
Internal expert

University of Lausanne
Faculty of Business and Economics

Ph.D. in Management

I hereby certify that I have examined the doctoral thesis of

Lukas Emanuel von FLÜE

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature: _____



Date: _____



Prof. Heinrich NAX
External expert

Acknowledgements

I am very grateful to my supervisor, Prof. Sonja Vogt, and my co-supervisor, Prof. Charles Efferson, as well as the Swiss National Science Foundation for funding my doctoral thesis. My journey with Charles and Sonja began during my Master's studies at the University of Zurich. Since then, I have had the privilege of working closely with them and learning from their vast expertise.

Sonja and Charles lead a research lab that perfectly blends two fields that captivate me. Sonja's focus on the applied side of research has allowed me to pursue my passion for making meaningful contributions towards the betterment of society. On the other hand, Charles' expertise in evolutionary science has enabled me to develop skills in an area that has fascinated me since childhood. The opportunity to combine these disciplines in my PhD research is a privilege for which I am truly thankful. I look forward to continuing my work with both Sonja and Charles in the future.

I would also like to thank my thesis committee members, Prof. Sonja Vogt, Prof. Charles Efferson, Prof. Christian Zehnder, and Prof. Heinrich Nax, for their invaluable guidance, knowledge, and expertise. Their insightful feedback has been instrumental in shaping this thesis into its final form.

My deepest gratitude also goes to my colleagues and friends in the PACE lab. The countless inspiring discussions we shared over coffee or beer have been a constant source of motivation and intellectual growth.

I am immensely thankful to my friends and family for their unwavering support, especially during times when I needed a break from work. Your encouragement and understanding provided the much-needed psychological support throughout this journey.

Lastly, to my partner, Eva, your patience and understanding during the demanding phases of this research have been deeply supportive. Our moments together in nature have been essential to my journey. For this, and so much more, I am profoundly grateful.

Thank you!

Contents

Introduction	iii
0.1 Initiating cultural change to improve society?	iii
0.2 Complexities and challenges of social tipping	iv
0.3 PhD chapters	v
0.4 Conclusion	vii
1 Heterogeneity in conformist social learning between and within groups	1
1.1 Introduction	3
1.2 Methods and materials	6
1.2.1 Preregistration	6
1.2.2 Participants	6
1.3 Results	16
1.3.1 Flexibility in social learning	16
1.4 Discussion	29
1.5 Conclusion	34
2 Green preferences sustain greenwashing - Challenges in the cultural transition to a sustainable future	41
2.1 Introduction	43
2.2 Applied cultural evolution	47

2.2.1	Cultural evolution of beliefs	47
2.2.2	Cultural evolution of pro-environmental values and norms	48
2.3	Buyer-seller exchange with information asymmetry and heterogeneous sustain-	
	ability preferences	50
2.3.1	Game sequence	54
2.3.2	Game analysis	54
2.3.3	Informing consumers	59
2.3.4	Competition	61
2.4	Discussion	63
3	Enhance threshold models to study the effects of heterogeneity in learning	
	on social norm change	77
3.1	Introduction	79
3.2	Model	81
3.2.1	Game	81
3.2.2	Learning strategies	83
3.2.3	Initial conditions and interventions	87
3.2.4	Structure of simulation	91
3.3	Results	93
3.3.1	Turning agents into individualists and changing preferences	94
3.3.2	Turning agents into individualists without changing preferences	96
3.3.3	Turning agents into success-based learners	97
3.3.4	Turning agents into success-based learners in a polarised population	99
3.4	Discussion	100

Introduction

Man is by nature a social animal...

— Aristotle, *Politics*

0.1 Initiating cultural change to improve society?

As we stand at the precipice of potentially irreversible environmental and social challenges, the need for a unified approach in behavioural science becomes increasingly clear (Constantino et al., 2022; Efferson et al., 2023; Nielsen et al., 2024). Social norms have long been recognised as a potential mechanism for promoting sustainable behaviour (Cialdini et al., 1990; Cialdini and Jacobson, 2021). The array of applications is increasingly broad (Efferson et al., 2023), and ranges from the conservation of critical natural resources (Castilla-Rho et al., 2017), to the widespread adoption of sustainable farming practices (Läpple and Kelley, 2013), and the promotion of pro-environmental behaviour (Berger, 2021; Travers et al., 2021; Constantino et al., 2022).

Social norms are behavioural rules shaped by the collective expectations and observed practices within a population. Specifically, an individual has incentives to follow a given behavioural rule if she thinks sufficiently many others in society adhere to the norm, and if she believes that enough other people agree that the rule should be followed, and deviations

from the norm should potentially be punished (Bicchieri, 2006). These incentives to coordinate and conform can create stable equilibria at an aggregate level, so that a majority of people in a given population follow a certain norm for a long time. Without a change in those coordination and conformity incentives, a population might become stuck in a given equilibrium. This is not a problem if the current norm is beneficial to society. However, traditions such as female genital cutting (Efferson et al., 2015; Vogt et al., 2016), child marriage (Bicchieri et al., 2014), or son bias (United Nations, 2019; Schief et al., 2021) serve as stark reminders of the persistence of norms that are considered harmful (World Health Organization, 2009).

The unsustainable lifestyle that is currently widespread in most high-income countries can be perceived as a set of harmful norms, if not yet always to ourselves, then often to the environment (Otto et al., 2020; Constantino et al., 2022; Masson-Delmotte et al., 2022). The concept of social tipping suggests a pathway to transformative change in these areas. It encapsulates the phenomenon where a norm can rapidly and self-reinforcingly shift once a critical mass of individuals alters their behaviour, thereby inspiring others to do the same (Granovetter, 1978; Efferson et al., 2020, 2023). In fact, the idea of relying on social tipping to promote environmentally friendly behaviour has been discussed with increasing interest recently in the field of cultural evolution (Nyborg, 2020; Otto et al., 2020; Berger, 2021; Constantino et al., 2022; Berger et al., 2023; Efferson et al., 2023; von Flüe et al., 2024). From the perspective of a social planner it seems like an attractive thought, given the urgency of climate change, to target a segment of the population with a preference changing policy intervention, and then let social influence take over to tip society into a more beneficial equilibrium.

0.2 Complexities and challenges of social tipping

While the concept of social tipping holds promise for rapid normative change, the journey to fully understand the complexities of social norm change is just beginning. Even if we imagine a very simple social learning scenario, the number of possible behavioural strategies

a single individual could exhibit is larger than the number of atoms in the universe (Efferson et al., 2023). This is critical because even subtle shifts in preference distribution and the psychological mechanisms underlying social learning can significantly impact the dynamics of social norms (Young, 2009; Efferson et al., 2020; Andreoni et al., 2021; Schimmelpfennig et al., 2021; Efferson et al., 2023). Crucially, we know that individuals differ strongly in their ways of responding to social information (Mesoudi et al., 2016; Kendal et al., 2018), and empirical studies confirm that this type of individual heterogeneity has real-world consequences for social norm change in the domains of public health, sustainability, and economic and political behaviour (Efferson et al., 2015; Vogt et al., 2016; Alvergne and Stevens, 2021; Eriksson et al., 2021; Ehret et al., 2022; Salali et al., 2022).

This complexity is increasingly recognised within the sustainability discourse, highlighting the urgent need for focused research to equip policymakers with effective tools for fostering sustainable practices (Constantino et al., 2022; Berger et al., 2023; Efferson et al., 2023). In recent years, the study of cultural evolution and norm change has taken on new dimensions, driven by advances in understanding the complexity of social learning processes across different contexts (Efferson et al., 2016; Vogt et al., 2016; Efferson et al., 2023). One of the key findings is that people differ greatly in their preferences and their ways of responding to social information (Kendal et al., 2018). However, simply acknowledging the world's complexity is inadequate if behavioural research is to effectively guide policymakers in initiating norm changes toward a sustainable future.

0.3 PhD chapters

The three projects outlined in this thesis collectively expand our understanding of how individual behaviours, influenced by diverse social learning strategies, contribute to cultural evolutionary dynamics. Each project, while distinct in its focus and methodology, contributes a part to a more nuanced portrait of the relationship between individual decision-making and collective behavioural patterns.

The first project supports the notion that individuals differ greatly in social learning (von Flüe, Vogt, and Efferson, unpublished data). Importantly, contrary to recent studies showing how flexibly individuals can use social information by adjusting to different group dynamics and environmental cues (Efferson et al., 2016; Bellamy et al., 2022), the first chapter in this thesis shows that people are often biased by information coming from individuals sharing a group identity. This finding complements research showing that polarisation can prevent norm change (Ehret et al., 2022). This is particularly pertinent in discussions on climate change, where societal divisions often align with normative beliefs (Falkenberg et al., 2022). Future research should explore whether it is feasible to disentangle group identities from sustainability issues, as this separation may be crucial for accelerating the adoption of pro-environmental norms (Efferson et al., 2020; Ehret et al., 2022).

Social learning biases linked to group identities are only one of several factors that may impede cultural shifts toward sustainability. The second project explores how greenwashing impacts situations where consumers and producers have differing environmental concerns (von Flüe et al., 2024). It reveals that information asymmetry, with consumers having less knowledge about the supply chain and its oversight compared to producers, enables firms to exploit environmentally-conscious consumers, potentially increasing the prevalence of greenwashing. Similar to the first project, this study not only acknowledges individual differences in behaviour but also investigates how these differences challenge norm change and explores potential non-invasive, non-paternalistic methods to address these challenges. A crucial insight from this research is that promoting pro-environmental values may not always be effective unless society ensures that consumers are equipped with the necessary information to make informed decisions.

The third project extends the findings from the first two by implementing an agent-based model to explore how different information types influence individual and group decision-making in the context of norm change (von Flüe, unpublished data). It specifically addresses the concept of social tipping, highlighting how minor, targeted interventions could catalyse

widespread societal shifts. This model advances existing norm change models by integrating findings about the heterogeneity of social learning. It demonstrates that interventions are more effective when social planners have a detailed understanding of the specific distribution of preferences and social learning patterns within society.

0.4 Conclusion

At the heart of the difficulty in predicting and facilitating social tipping lies the vast diversity in individual preferences and the psychological underpinnings structuring our responses to social information. The effectiveness of policies aimed at promoting pro-environmental behaviour is often hampered by the insufficient consideration of how these individual differences interact with the structure of social networks. The assumption that populations are homogeneous in their preferences and behaviours leads to overly simplistic models of social change, neglecting the complex web of social influences that shape individual decisions.

The interconnections between the three projects point to broader implications for both theory and practice. The findings of this thesis contribute to a more refined understanding of how norms evolve and how social learning processes can be harnessed to promote positive societal change. For policymakers, the research underscores the need to consider cultural diversity and individual heterogeneity when designing interventions aimed at fostering sustainable behaviours or other desirable social norms.

The key take away from this thesis, which might seem obvious, is that societal change is complex because humans are complex. However, humans are complex to a great extent because they are social animals. And because we are inherently social, we learn from each other. It is my wish that future research will help us better understand how this works so that we can contribute to a cultural change for good.

Bibliography

- Alvergne, A. and Stevens, R. (2021). Cultural change beyond adoption dynamics: Evolutionary approaches to the discontinuation of contraception, *Evolutionary Human Sciences* pp. 1–45.
- Andreoni, J., Nikiforakis, N. and Siegenthaler, S. (2021). Predicting social tipping and norm change in controlled experiments, *Proceedings of the National Academy of Sciences* **118**(16).
- Bellamy, A., McKay, R., Vogt, S. and Efferson, C. (2022). What is the extent of a frequency-dependent social learning strategy space?, *Evolutionary Human Sciences* **4**: e13.
- Berger, J. (2021). Social tipping interventions can promote the diffusion or decay of sustainable consumption norms in the field. evidence from a quasi-experimental intervention study, *Sustainability* **13**(6): 3529.
- Berger, J., Efferson, C. and Vogt, S. (2023). Tipping pro-environmental norm diffusion at scale: opportunities and limitations, *Behavioural Public Policy* **7**(3): 581–606.
- Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*, Cambridge University Press.
- Bicchieri, C., Jiang, T. and Lindemans, J. W. (2014). A social norms perspective on child marriage: The general framework, *New York: UNICEF* .
- Castilla-Rho, J. C., Rojas, R., Andersen, M. S., Holley, C. and Mariethoz, G. (2017). Social tipping points in global groundwater management, *Nature Human Behaviour* **1**(9): 640–649.
- Cialdini, R. B. and Jacobson, R. P. (2021). Influences of social norms on climate change-related behaviors, *Current Opinion in Behavioral Sciences* **42**: 1–8.
- Cialdini, R. B., Reno, R. R. and Kallgren, C. A. (1990). A focus theory of normative conduct:

- Recycling the concept of norms to reduce littering in public places., *Journal of Personality and Social Psychology* **58**(6): 1015.
- Constantino, S. M., Sparkman, G., Kraft-Todd, G. T., Bicchieri, C., Centola, D., Shell-Duncan, B., Vogt, S. and Weber, E. U. (2022). Scaling up change: A critical review and practical guide to harnessing social norms for climate action, *Psychological Science in the Public Interest* **23**(2): 50–97.
- Efferson, C., Lalive, R., Cacault, M. P. and Kistler, D. (2016). The evolution of facultative conformity based on similarity, *PLOS One* **11**(12): e0168551.
- Efferson, C., Vogt, S., Elhadi, A., Ahmed, H. E. F. and Fehr, E. (2015). Female genital cutting is not a social coordination norm, *Science* **349**(6255): 1446–1447.
- Efferson, C., Vogt, S. and Fehr, E. (2020). The promise and the peril of using social influence to reverse harmful traditions, *Nature Human Behaviour* **4**(1): 55–68.
- Efferson, C., Vogt, S. and von Flüe, L. (2023). Activating cultural evolution for good when people differ from each other, in J. J. Tehrani, J. Kendal and R. Kendal (eds), *The Oxford Handbook of Cultural Evolution*.
- Ehret, S., Constantino, S. M., Weber, E. U., Efferson, C. and Vogt, S. (2022). Group identities can undermine social tipping after intervention, *Nature Human Behaviour* pp. 1–11.
- Eriksson, K., Strimling, P., Gelfand, M., Wu, J., Abernathy, J., Akotia, C. S., Aldashev, A., Andersson, P. A., Andrighetto, G., Anum, A. et al. (2021). Perceptions of the appropriate response to norm violation in 57 societies, *Nature Communications* **12**(1): 1481.
- Falkenberg, M., Galeazzi, A., Torricelli, M., Di Marco, N., Larosa, F., Sas, M., Mekacher, A., Pearce, W., Zollo, F., Quattrocioni, W. et al. (2022). Growing polarization around climate change on social media, *Nature Climate Change* **12**(12): 1114–1121.

- Granovetter, M. (1978). Threshold models of collective behavior, *American Journal of Sociology* **83**(6): 1420–1443.
- Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M. and Jones, P. L. (2018). Social learning strategies: Bridge-building between fields, *Trends in Cognitive Sciences* **22**(7): 651–665.
- Läpple, D. and Kelley, H. (2013). Understanding the uptake of organic farming: Accounting for heterogeneities among Irish farmers, *Ecological Economics* **88**: 11–19.
- Masson-Delmotte, V., Zhai, P., Pörtner, H.-O., Roberts, D., Skea, J., Shukla, P. R. et al. (2022). *Global Warming of 1.5 C: IPCC special report on impacts of global warming of 1.5 C above pre-industrial levels in context of strengthening response to climate change, sustainable development, and efforts to eradicate poverty*, Cambridge University Press.
- Mesoudi, A., Chang, L., Dall, S. R. and Thornton, A. (2016). The evolution of individual and cultural variation in social learning, *Trends in Ecology & Evolution* **31**(3): 215–225.
- Nielsen, K. S., Cologna, V., Bauer, J. M., Berger, S., Brick, C., Dietz, T., Hahnel, U. J., Henn, L., Lange, F., Stern, P. C. et al. (2024). Realizing the full potential of behavioural science for climate change mitigation, *Nature Climate Change* **14**(4): 322–330.
- Nyborg, K. (2020). No Man is an Island: Social Coordination and the Environment, *Environmental and Resource Economics* **76**(1): 177–193.
- Otto, I. M., Donges, J. F., Cremades, R., Bhowmik, A., Hewitt, R. J., Lucht, W., Rockström, J., Allerberger, F., McCaffrey, M. and Doe, S. S. (2020). Social tipping dynamics for stabilizing Earth’s climate by 2050, *Proceedings of the National Academy of Sciences* **117**(5): 2354–2365.
- Salali, G. D., Uysal, M. S., Bozyel, G., Akpınar, E. and Aksu, A. (2022). Does social influence affect covid-19 vaccination intention among the unvaccinated?, *Evolutionary Human Sciences* **4**: E32.

- Schief, M., Vogt, S. and Efferson, C. (2021). Investigating the structure of son bias in armenia with novel measures of individual preferences, *Demography* .
- Schimmelpfennig, R., Vogt, S., Ehret, S. and Efferson, C. (2021). Promotion of behavioural change for health in a heterogeneous population, *Bulletin of the World Health Organization* **tbd**(tbd): tbd.
- Travers, H., Walsh, J., Vogt, S., Clements, T. and Milner-Gulland, E. J. (2021). Delivering behavioural change at scale: What conservation can learn from other fields, *Biological Conservation* **257**: 109092.
- United Nations (2019). *World population prospects 2019*, (Report.) New York, NY: United Nations, Department of International Economic and Social Affairs, Population Division. <https://www.un.org/development/desa/pd/news/world-population-prospects-2019-0>.
- Vogt, S., Zaid, N. A. M., Ahmed, H. E. F., Fehr, E. and Efferson, C. (2016). Changing cultural attitudes towards female genital cutting, *Nature* **538**(7626): 506–509.
- von Flüe, L., Efferson, C. and Vogt, S. (2024). Green preferences sustain greenwashing: challenges in the cultural transition to a sustainable future, *Philosophical Transactions of the Royal Society B* **379**(1893): 20220268.
- World Health Organization (2009). Changing cultural and social norms that support violence.
- Young, H. P. (2009). Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning, *American Economic Review* **99**(5): 1899–1924.

Chapter 1

Heterogeneity in conformist social learning between and within groups

Abstract

In this study, we investigate how cognitive load and social learning biases influence frequency-dependent social learning in different subject pools. Our primary research question explores how similarity and group membership affect the social learning process with and without cognitive load. We conducted a highly incentivised behavioural experiment where one group, termed social learners, observed the choice distribution of another group, called demonstrators, before making their decisions. We manipulated two key variables, which were the similarity of social learners to the observed demonstrators and whether they observed ingroup or outgroup demonstrators. This produced four informationally equivalent conditions, which meant that social learners could, in principle, flexibly adjust between the conditions and maximise payoffs. Contrary to our hypothesis that cognitive load would induce heterogeneity and reduce the flexibility of social learning, we found no significant difference in behaviour between the treatments with and without cognitive load. For both treatment conditions with and without cognitive load, we find that participants perform better when they observe social information from the ingroup than from the outgroup. This suggests that social learning biases might be too robust to be altered by cognitive load. We conducted our study with two distinct subject pools from Switzerland and Kenya. Interestingly, our results show a strong difference in the degree of heterogeneity in social learning between the two subject pools. Our study supports the notion that it is important to consider diverse cultural backgrounds in social learning research and to examine both individual- and group-level variations. Our findings challenge the assumption that people are completely flexible in their use of social information and emphasise the need to broaden the scope of research beyond the populations predominantly studied.

Keywords: cultural evolution, frequency-dependent social learning, conformity, ingroup bias, similarity bias, cognitive load

1.1 Introduction

A traditional assumption made in cultural evolution literature is that individuals exhibit fixed strategies when they observe frequency-dependent information about others' behaviour (Boyd and Richerson, 1985; Aoki and Feldman, 2014). In other words, a person who is biased to conform to the majority of her social group is assumed to do so even if information about other aspects in her environment changes. Certain models of social norm change are based on this simplified characterisation of social learning strategies that are fixed at the individual level, which facilitates a mathematical analysis of evolutionary dynamics (Granovetter, 1978; Henrich, 2001).

However, recent research shows that reality is much more complex. On the one hand, specific learning biases exist, but the number of these biases is higher than traditionally assumed, and social learning varies not only between but also within individuals (Mesoudi et al., 2016; Muthukrishna et al., 2016; Kendal et al., 2018). On the other hand, while cognitive biases can sometimes influence behaviour, individuals are not always susceptible to them. For instance, individuals appear to be quite flexible and they often adapt their strategies based on contextual cues such as group membership (Efferson et al., 2016; Bellamy et al., 2022).

While there exists an increasing interest in empirically studying social learning, we have only started discovering the complexities involved with respect to heterogeneity in social learning and its influence on norm change (Young, 2009; Efferson, Vogt and Fehr, 2020; Constantino et al., 2022; Efferson et al., 2023). One of the limiting factors for progress in this area is the fact that empirical research on social learning has predominantly focused on individuals from Western cultures (Henrich et al., 2010). Further, the conflicting evidence about observing biased social learning on the one hand (Kendal et al., 2018) and a high degree of flexibility on the other hand (Efferson et al., 2016; Bellamy et al., 2022) suggests that we have to come up with new study designs that allow us to discern under which circumstances people are more or less prone to social learning biases.

We address those two shortcomings in this study. We investigate frequency-dependent social learning in two distinct subject pools, one from Zurich, Switzerland, and another from Nairobi, Kenya. The most recent evidence of individuals' flexibility in using social cues comes from a laboratory setting in which participants were able to focus on and think strategically about the information provided (Efferson et al., 2016; Bellamy et al., 2022). We seek to test whether this observed flexibility in behaviour persists when individuals do not have the opportunity to calmly think about how to respond to social cues, particularly when under stress or cognitive load. Under such conditions, individuals often exhibit reduced strategic and analytical thinking, resorting to impulsive and habitual behaviours (Duffy and Smith, 2014; Haushofer and Fehr, 2014). Moreover, stress and cognitive load appear to increase reliance on social information (Murray and Schaller, 2012; Delfino et al., 2016; Buckert et al., 2017; Jacquet et al., 2018). However, it remains unclear how stress and cognitive load affect specific social learning strategies, and whether the effects of stress and cognitive load on social learning vary across subjects from two countries with distinct cultures.

Therefore, in a highly incentivised behavioural experiment involving two distinct subject pools, we tested whether individuals could adjust between different contexts with and without being under cognitive load. Our study is the first to directly compare participants from two distinct subject pools regarding the effects of different forms of heterogeneity on frequency-dependent social learning. We created two treatments, one to manipulate whether participants observed ingroup or outgroup individuals, and a second that varied whether participants were similar or dissimilar to the observed individuals. This combination resulted in four informationally equivalent treatment conditions, where it was optimal to follow either the majority or the minority of the observed participants.

Having four informationally equivalent treatments means that, in principle, participants possess all information necessary to earn the same expected payoffs in all four treatments. However, there are several reasons why participants might not respond to the observed social information equally in the four treatments. One of these reasons we are interested in is the

possibility that social cognition has evolved in ways that bias participants to learn more easily from similar rather than dissimilar others, from ingroup members rather than outgroup members, and from a majority rather than a minority.

Our hypothesis was that cognitive load, by shifting from goal-directed to habitual behaviour, limits social learners' ability to make optimal use of cues of group membership and similarity and makes them more susceptible to biases in social learning. Contradicting this hypothesis, our findings surprisingly reveal that participants exhibited similarity and ingroup biases regardless of whether they were cognitively distracted. Importantly, when participants observed the decisions of members of their own group, they were significantly better able to adapt to other changing information in order to maximise their payoffs than when they observed the decisions made in the outgroup. This evidence suggests a deeper, perhaps evolutionary, entrenchment of these biases (Efferson et al., 2016). More specifically, individuals might have cognitively evolved social learning biases that make them rely more on social information from ingroup members and enable them to evaluate social information from ingroup members more effectively than from outgroup members.

One of the most significant aspects of our findings concerns heterogeneity in social learning. We show that, in addition to individual variation, the degree of heterogeneity in social learning significantly differs between the two subject pools. Notably, while previous research has discussed that cultural differences in social learning exist and that they influence evolutionary dynamics (Mesoudi et al., 2015; Muthukrishna and Schaller, 2020), our study is the first to so comprehensively describe multiple forms of heterogeneity in social learning and compare them between two distinct subject pools. Our results support the notion that empirical research in social learning should be expanded beyond individuals from Western cultures, focusing more on differences between groups (Mesoudi et al., 2015).

1.2 Methods and materials

1.2.1 Preregistration

We conducted our study with participants from two different subject pools. The pre-registrations for both experiments can be found on <https://osf.io/mzfq5/>. In these documents, we specify the experimental design, the number of subjects we intend to recruit, the research questions and hypotheses, and how we will analyse the results.

1.2.2 Participants

A total of 273 English-speaking students participated at the Decision Science Laboratory at the ETH Zurich, Switzerland (mean age = 24.13, SD = 4.12, females=155, males=114, other=4, average earnings = ca. CHF 40 \approx USD 45), and 266 English-speaking students participated at the Busara Center for Behavioral Economics in Nairobi, Kenya (mean age = 21.58, SD = 1.87, females=112, males=154, other=0, average earnings = ca. KES 1565 \approx USD 10). The monetary incentives were adapted to local contexts in collaboration with the two laboratories, and the two ethics committees, Human Subjects Committee of the Faculty of Business and Economics at the University of Lausanne and the Amref Health Africa in Kenya Ethics and Scientific Review Committee.

In the questionnaire at the end of the experiment, the participants entered their fields of study as free text, and we categorised them into the following three broad categories: “Humanities”, “Natural Sciences” and “Social Sciences”. To simplify the analysis, we kept the number of categories small. We have therefore assigned certain subjects to one of the three categories, even if they belong to another category. For example, mathematics was assigned to “natural sciences” instead of forming the separate category “formal sciences” (see supplementary materials for more details). In Zurich, the distribution is as follows: Humanities: 11; Natural Sciences: 139; Social Sciences: 43. The distribution across subjects in the Nairobi data is as follows: Humanities: 44; Natural Sciences: 42; Social Sciences: 80.

A key aspect of our experimental design is that participants were assigned one of two roles at the beginning of the experiment, one which we call “demonstrators” and one which we call “social learners”. In the experiment we used the more neutral terms “Type A” and “Type B” respectively to avoid any biases that could potentially result from participants’ associations with the less neutral terms. The participants kept their roles throughout the experiment. The demonstrators went through a decision task and received private information about the outcomes. In contrast, social learners first observed the choices of certain demonstrators before making their own decisions, but they did not receive any private information about the outcomes of their choices. We did this because we are interested in how social learners respond to social information only. If social learners also had access to private information, they could have also responded to this type of information, and it would be more complex to isolate the causal effect of social information (Manski, 1993, 2000; Efferson et al., 2016).

The choice task

The task was the same for all participants, except for the information the participants had access to. It is a probabilistic choice task, which was framed in the following way. There are two urns, one of which consists of 3 red marbles and 1 blue marble, while the other urn consists of 3 blue marbles and 1 red marble. Hence, the probability of drawing a marble of a certain colour varies between those two urns. Participants never knew for certain which urn contained which set of marbles because the urns were shuffled before participants made their choices, and the colour of the marbles were hidden as long as the marbles were in the urn. Figure 1.1 below was shown to participants during the instructions to explain the set up of the choice task. In addition to this, an animation was programmed to explain the sequence of the task, and participants went through exercise rounds so that they can learn how the choice task will work during the experiment. The graphics and the animation that we showed to participants during the instructions can be found in section 3.2 of the supplementary materials. During the actual experiment, participants went through several rounds of choosing between the two

urns, and in any given round they could win points for one of the two colours. Hence, one of the two urns was optimal for a given participant in a given round in the sense that it had more marbles of the winning colour.

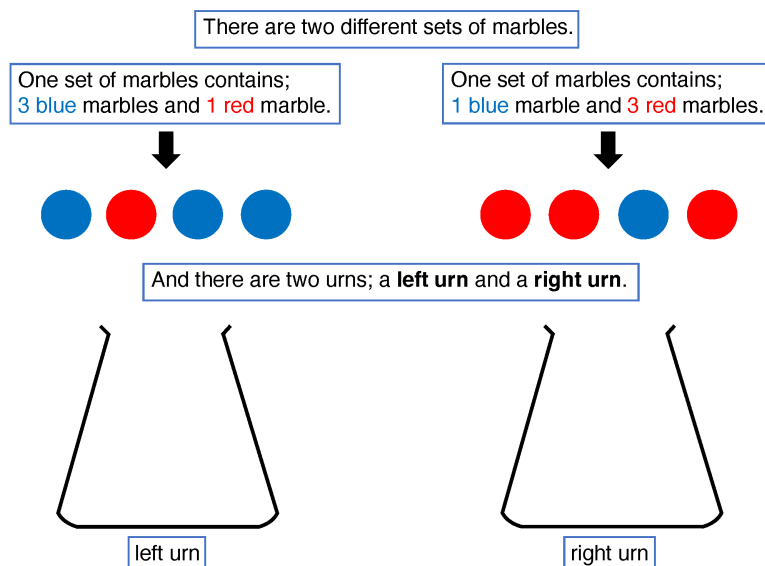


Figure 1.1: Graphic explaining the choice task set up.

Participants who were assigned to the role of the demonstrators went through 20 blocks of 4 rounds of playing the urn choice task. At the beginning of each block the two urns were shuffled and then stayed in the same position for the rest of the block. Hence, the demonstrators had 4 rounds to figure out which urn contained more marbles of their winning colour in a given block. Every time a demonstrator chose an urn, a marble was randomly drawn from the urn by the computer, the colour of the marble was shown to the demonstrator, and it was put back into the urn. A demonstrator received 100 points every time a marble of her winning colour was drawn, and 0 points if not. Not every demonstrator had the same winning colour, and we explain this in more detail in the “Treatments” section. Points were converted to real money at the end of the experiment.

In contrast, participants who were assigned to the role of the social learners only chose once per block. Before making their choice, the social learners observed the choice distribution of

certain demonstrators in the fourth round of blocks, and we explain this in more detail below. Importantly, because social learners only made 20 choices compared to the 80 choices made by demonstrators, this would have meant that they could earn less points had we paid social learners in the same way we paid demonstrators. We solved this problem by randomly drawing four marbles with replacement from the urn chosen by a social learner in each block. Hence, the social learners were able to earn the same average amount of points as the demonstrators.

Treatments

We have a two-by-two-by-two experimental design. Two treatments are within-subjects and one is between subjects. To implement our two within-subjects treatments, we divided all participants into two groups, a “triangle group” and a “square group”. In each of those two groups, there were always 5 demonstrators, and the rest were social learners. An example of an assignment is shown in figure [1.2](#).

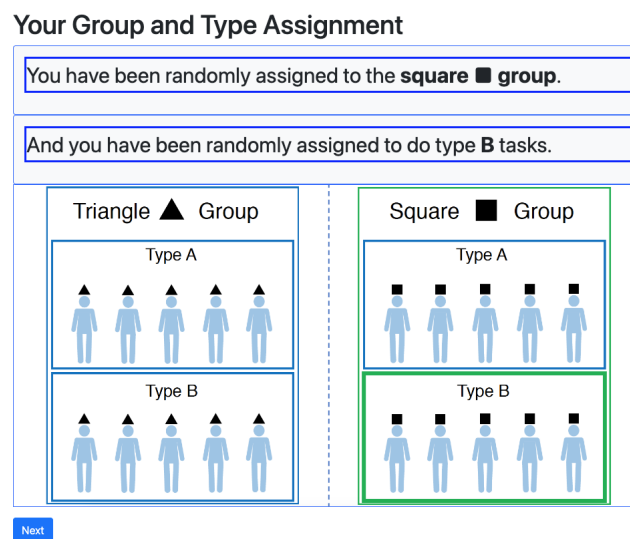


Figure 1.2: Example of group and type assignment.

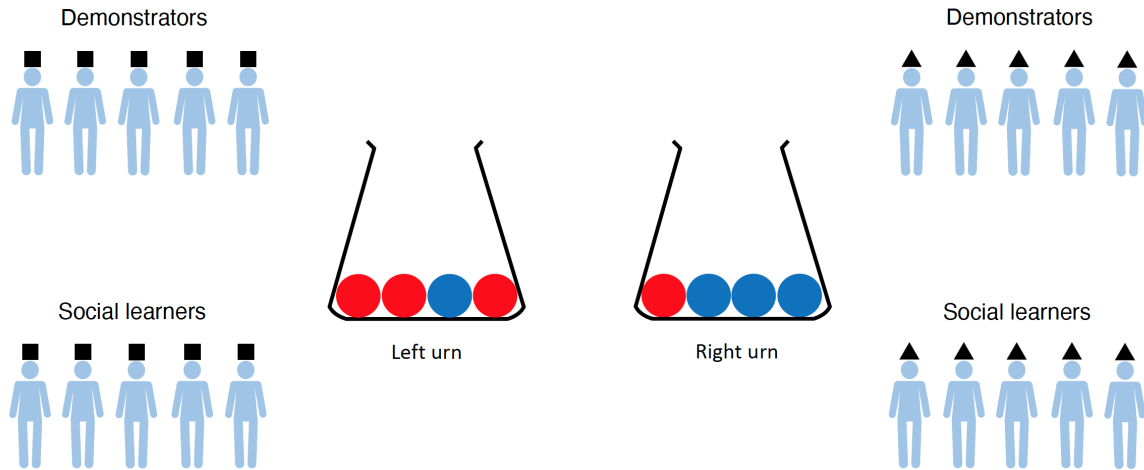
Participants kept their group membership until the end of the experiment. At the beginning of the experiment it was determined which group of demonstrators earns points for which colour of marbles, and social learners were informed about this. The first treatment

varied whether a social learner observed the choices of ingroup demonstrators or outgroup demonstrators. The second treatment varied whether a social learner was very similar or very dissimilar to demonstrators of her ingroup in terms of earning points for a given colour of marbles. More precisely, in any given block the probability for a social learner to earn points for the same colour as demonstrators of her ingroup was either 90% or 10%.

Hence, there are four possible combinations, namely observing ingroup demonstrators and being similar to ingroup demonstrators, observing the ingroup and being dissimilar to the ingroup, observing the outgroup and being similar to the ingroup, and observing the outgroup and being dissimilar to the ingroup. The four combinations of the within-subjects treatments were informationally equivalent in the following sense. The demonstrators of the two groups, triangle and square, received points for marbles of opposite colours. This meant that if the left urn was optimal for the triangle demonstrators in terms of containing more marbles of the winning colour of the triangle demonstrators in a particular block, the right urn was optimal for the square demonstrators in the same block and vice versa. This is illustrated in figure [L.3](#). As a consequence, in any given block it was always optimal to either follow the majority or minority of observed demonstrators.

To illustrate, assume a triangle social learner is similar in terms of the winning colour to her ingroup, and she observes her ingroup in a particular block. If the left urn contains $3/4$ marbles of the colour for which triangle demonstrators win points in this block, and assuming the demonstrators figure this out over the course of the four rounds in this block, then it is best for the triangle social learner to follow the majority behaviour of the observed triangle demonstrators and choose the left urn too. The reason for this is that a triangle social learner who is similar to her ingroup, earns points for the same colour as triangle demonstrators with a probability of 90%. Let us now assume instead that the same social learner in the same block is similar to her ingroup but instead observes square demonstrators from the outgroup. This might be less intuitive than the first case where a triangle social learner who is similar to her ingroup demonstrators observes ingroup triangle demonstrators. Thus, an example of this

If there are 3 red marbles and 1 blue marble in the left urn in this particular block, then there must be 1 red marble and 3 blue marbles in the right urn and this would be true for everyone.



If square demonstrators earn points for red marbles and triangle demonstrators earn points for blue marbles, the left urn is optimal for square demonstrators and the right urn is optimal for triangle demonstrators in this block. Demonstrators have four rounds in each block to find out which urn contains more marbles of their winning colour.

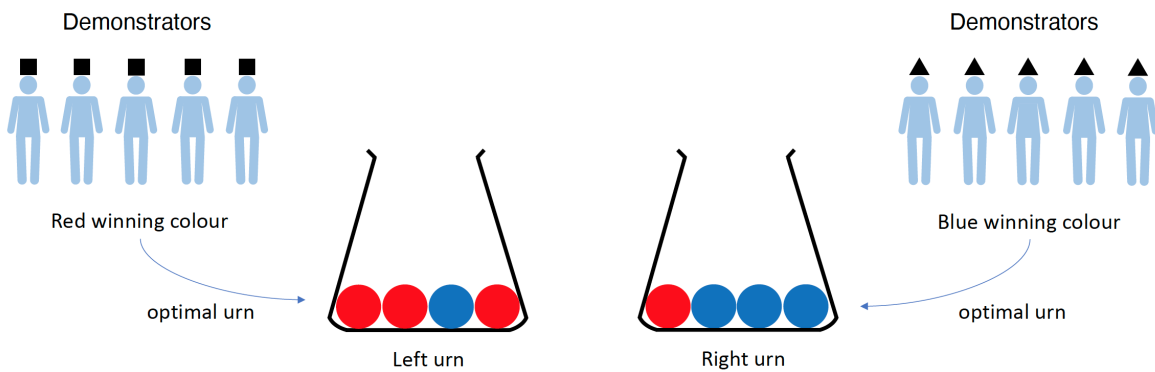


Figure 1.3: Demonstrators of the two groups earn points for opposite colours

situation, as presented to the participants during the experiment, is shown in figure 1.4. Again assume the square demonstrators figure out which urn contains more marbles of the colour for which they win points. For them, the urn containing more marbles of their winning colour is the right urn in a block where the left urn is optimal for triangle demonstrators because they get points for the opposite colour than the triangle demonstrators. Being similar to the demonstrators of her ingroup, the triangle group, implies that the triangle social learner is not similar to the outgroup in terms of the winning colour, as the two groups of demonstrators receive points for opposite colours. Hence, for a triangle social learner who is similar to triangle demonstrators, it is optimal to follow the minority behaviour of square demonstrators in this particular block. For example, if the majority of square demonstrators choose the right urn in this block, then it is optimal for the triangle social learner to do the opposite and choose the left urn.

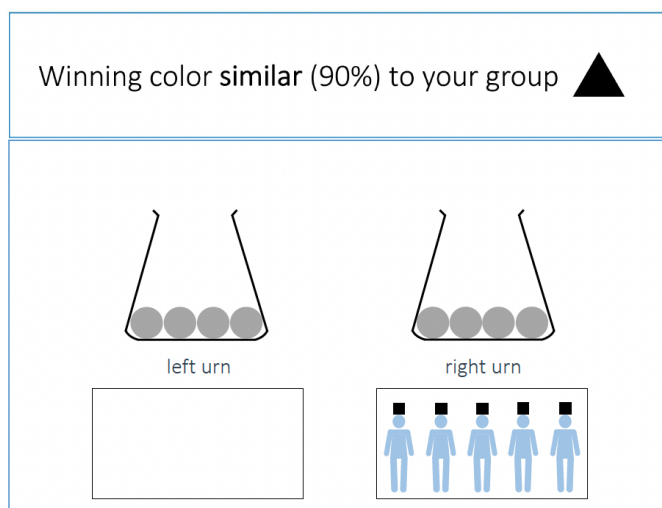


Figure 1.4: The social learner is similar to the ingroup but observes the outgroup.

Similar implications follow for the other two treatment combinations. If the triangle social learner is dissimilar to her ingroup, i.e. the triangle demonstrators, then it is optimal for her to follow the minority when observing triangle demonstrators, but follow the majority when observing square demonstrators.

With the given information about similarity to ingroup demonstrators and seeing either the choices of ingroup or outgroup demonstrators, it is always optimal to either follow the majority or minority of observed demonstrators. The social learners were shown examples of all four types of social information during the instructions (see supplementary materials 3.3-3.5). They therefore had time to consider whether it was best to follow the majority or the minority of the observed demonstrators for each type of social information before the actual experiment began. A social learner who understands this and makes optimal use of this information can, in principle, achieve equivalent payoffs on average in all four within-subjects treatments. This is what is meant when we say that the four treatments are informationally equivalent. However, there are several possible reasons why participants may not utilise the four different types of social information equally, which could reduce the number of points they receive. We discuss these potential reasons in the “Discussion” section.

The within-subjects treatment combinations did not switch between every block. Rather, there were four groups of five blocks. In a group of five blocks, one treatment combination was implemented at a time to allow social learners to become accustomed to the new type of information. We counterbalanced the order of the four groups of blocks, and by extension the four treatments, across social learners. The social learners did not know that the four different types of social information will be shown to them in groups of five blocks. Instead, they only knew that the type of information can change between blocks.

In addition to the four within-subjects treatments, we also implemented a between subjects treatment variation. Half of all social learners were randomly assigned to the “cognitive load” treatment (often abbreviated to “CL” in the following) and the other half to the “no cognitive load” treatment (often abbreviated to “NCL” in the following). In each block, before observing the social information and the choice between the two urns, social learners in the CL treatment were given 8 seconds to memorise a 7-digit number, and they had to recall and enter the number from memory after the urn choice task. In each block, a new 7-digit number was randomly generated, with a different number produced for each social learner. To achieve

the best possible comparability between the CL and the NCL treatment, social learners in the NCL treatment were also shown a 7-digit number. However, they learned during the instructions that they would not have to memorise this number because they will be shown their number after the choice task when they had to enter the number (see section 3.6 in the supplementary materials). Making participants memorise 7-digit numbers is a standard approach to induce cognitive load and strain the working memory (Miller, 1956; Shiv and Fedorikhin, 1999; Duffy and Smith, 2014).

Although memorisation tasks are often not incentivised (Duffy and Smith, 2014), we have done so to increase the likelihood of inducing cognitive load. In particular, social learners were informed that they were equally likely to receive points for either the social learning task or the memorisation task, but not for both. Importantly, we did not disclose which specific task would be rewarded, introducing an element of uncertainty. The rationale was that this uncertainty would incentivise participants to perform well in both tasks, as focusing solely on one task carried the risk of receiving zero points if the ignored task turned out to be the rewarded one. Importantly, if a social learner had only focused on the social learning task, she would have risked receiving 0 points for the memorisation task if she had only received points for memorising the 7-digit numbers. We informed the social learners of this possibility during the instructions and told them that it was in their interest to focus equally on both tasks. While this incentive structure does not inherently prevent a participant from strategically choosing to focus on one task over the other, the answers to the questionnaire at the end of the experiment indicate that social learners spent a substantial fraction of their attention on both tasks (see supplementary materials).

We made sure that the social learners who earned points for the memorisation task could earn the same average amount of points as the social learners who earned points for the urn choice task. Specifically, if a social learner memorised the complete 7-digit number correctly, 4 marbles were randomly drawn with replacement from the urn containing $3/4$ marbles yielding 100 points. If a social learner did not correctly recall the complete number, 4 marbles were

randomly drawn with replacement from the suboptimal urn containing 1/4 marble yielding 100 points. As with the urn choice task, social learners did not receive any feedback about their performance in the memorisation task.

Procedure

All sessions were conducted using the oTree software (Chen et al., 2016) under anonymous laboratory conditions. Each participant in Zurich received a show-up fee of ca. USD 11, and each participant in Nairobi received a show-up fee of ca. USD 4. Those show-up fees are adapted to the local contexts and were set by the respective laboratories. Each participant provided informed consent by signing a consent form. The experiment was approved by the Human Subjects Committee of the Faculty of Business and Economics at the University of Lausanne and by the Amref Health Africa in Kenya Ethics and Scientific Review Committee.

We counterbalanced the winning colours for triangle and square demonstrators across sessions to avoid any uninteresting experimental artefacts related to colour preferences. To avoid experimental artefacts related to spatial biases, we arranged the buttons for choosing between the urns top to bottom instead of orienting them left to right (see supplementary materials). The order was randomly determined for each block and each social learner.

Participants were only able to proceed from the instructions to the actual experiment once they answered a series of quiz questions about the instructions correctly (see supplementary materials 3.7). If a participant did not answer a question correctly, an error message appeared with an explanation why her answer was incorrect and she had to answer the question again.

Participants were instructed not to talk during the experiment and not to use any mobile phone, or any other utensils. In each experimental session in Zurich and in Nairobi, two experimenters, including at least one of the authors, were present in the laboratory to monitor whether the participants followed these rules. Importantly, this allowed us to answer questions during the instructions and to ensure that the participants in the cognitive load treatment could not use any auxiliary means to memorise the multi-digit numbers.

At the end of the experiment, we collected demographic information, and we implemented a questionnaire (see section 4 in the supplementary materials).

1.3 Results

In our analysis, we focus on three main areas. First, we want to know whether participants responded to different social cues at all, and if so, how. In other words, we are interested in how flexible participants were in their use of social information. In a second step, we dive deeper and study how participants' use of social information affected their payoffs. Finally, we also want to find out which social learning strategies the participants used. Before discussing those results, it is important to note that for social information to be valuable, demonstrators have to choose optimally. Optimal choice of the demonstrators means that they choose the urn that contains $3/4$ marbles of the colour for which the demonstrators receive points. This was the case (see section 2.1 in the supplementary materials).

1.3.1 Flexibility in social learning

Based on studies with similar experimental designs (Efferson et al., 2016; Bellamy et al., 2022), we predicted that social learners are flexible in adjusting to the cues of similarity and group membership. More specifically, we expected social learners to flexibly switch between following either the majority or the minority of demonstrators between treatments.

Interestingly, we find that this is only true when social learners observed ingroup demonstrators. We observe this in the Zurich subject pool, and to a lesser extent also for the data from Nairobi. In figure 1.5 we show the proportion of social learners who choose the right urn as a function of the number of observed demonstrators choosing the right urn for the blocks in which social learners observed ingroup demonstrators. We can see that most social learners in Zurich followed the majority of observed ingroup demonstrators when they are similar, regardless of the cognitive load treatment. In contrast, social learners in Zurich followed the minority of observed ingroup demonstrators when they are dissimilar, again irrespective of

the cognitive load treatment. In contrast to the strong conformist social learning response that we observe in the Zurich data for the ingroup treatments, social learners in Nairobi only show this tendency for the blocks in which they are similar to their ingroup when observing ingroup demonstrators (figure 1.7). Surprisingly, we can not observe those clear patterns of conformist social learning for the outgroup treatments. Figures 1.6 and 1.8 show that there was no clear majority of social learners responding strongly to outgroup social information.

Another noteworthy result for the data in Zurich and Nairobi is that cognitive load does not seem to have any effect on how social learners respond to social information. A potential explanation could be that our cognitive load manipulation did not work. However, we tested whether the cognitive load manipulation reduced the social learners' likelihood to recall the numbers correctly, and the results show that it did so strongly and significantly (see supplementary materials 2.2.1). The results of the questionnaire could provide an alternative explanation. Weirdly, social learners in the "no cognitive load" (NCL) treatment answered that they tried to memorise the multi-digit numbers sometimes (see section 4.2 in the supplementary materials). Those social learners did not have to do so because we showed them the number when they had to enter it. If social learners in the "no cognitive load" treatment did try to memorise the number frequently, then their working memory would have been under similar load as the social learners in the "cognitive load" (CL) treatment. This could explain why there was no significant difference in behaviour between those two treatment groups. Importantly, however, social learners knew that there is a 50% probability that they will only earn points for the urn choice task. Hence, it would have been very risky for a social learner in the NCL treatment to focus too much on memorising the number instead of focusing on the urn choice task. Further, we also asked the social learners in the NCL treatment how distracting it was for them that we displayed the number, and most answered that it was only slightly distracting. This suggests that showing the 7-digit numbers to social learners in the NCL treatment did not negatively affect their focus required for the social learning task.

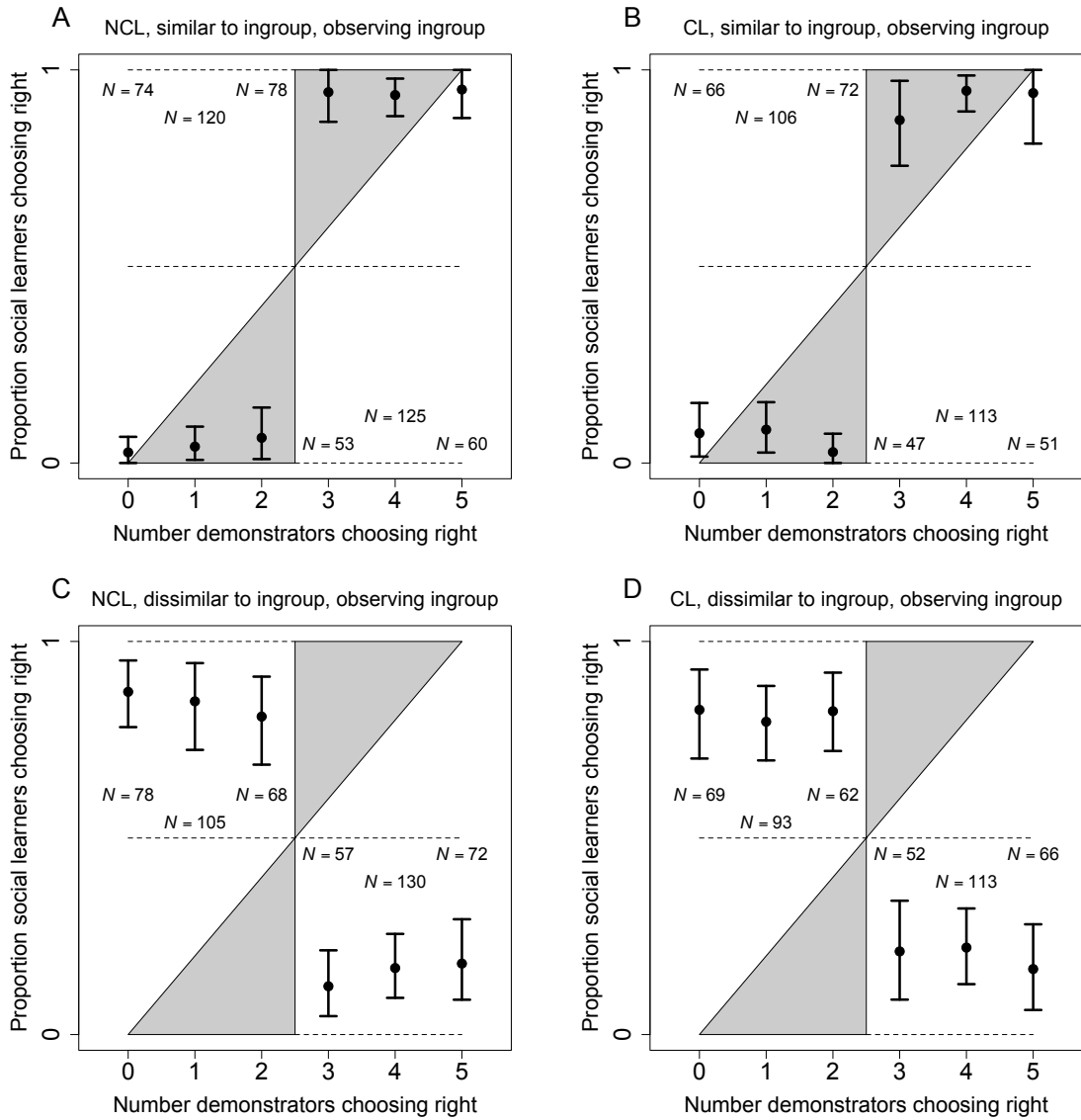


Figure 1.5: Right urn choices by social learners in treatments in which they observed **ingroup** demonstrators (results from **Zurich**). The points show the rate at which social learners chose the right urn as a function of the number of demonstrators choosing the right urn in the final round of blocks. The number of observations is vertically aligned with each point. **A** and **B** show treatments in which social learners are **similar** to ingroup demonstrators and they observe the ingroup. **C** and **D** show treatments in which social learners are **dissimilar** to their ingroup and they observe the ingroup. **A** and **C** represent “no cognitive load” treatments (NCL) and **B** and **D** show “cognitive load” treatments (CL). The error bars are 95% bootstrapped confidence intervals which we clustered on social learner. The gray area in each plot is consistent with conformist cultural transmission (Boyd and Richerson, 1985). The diagonal shown as a solid black line is consistent with unbiased social learning. Finally, the dashed lines provide additional points of reference at 0, 0.5, and 1.

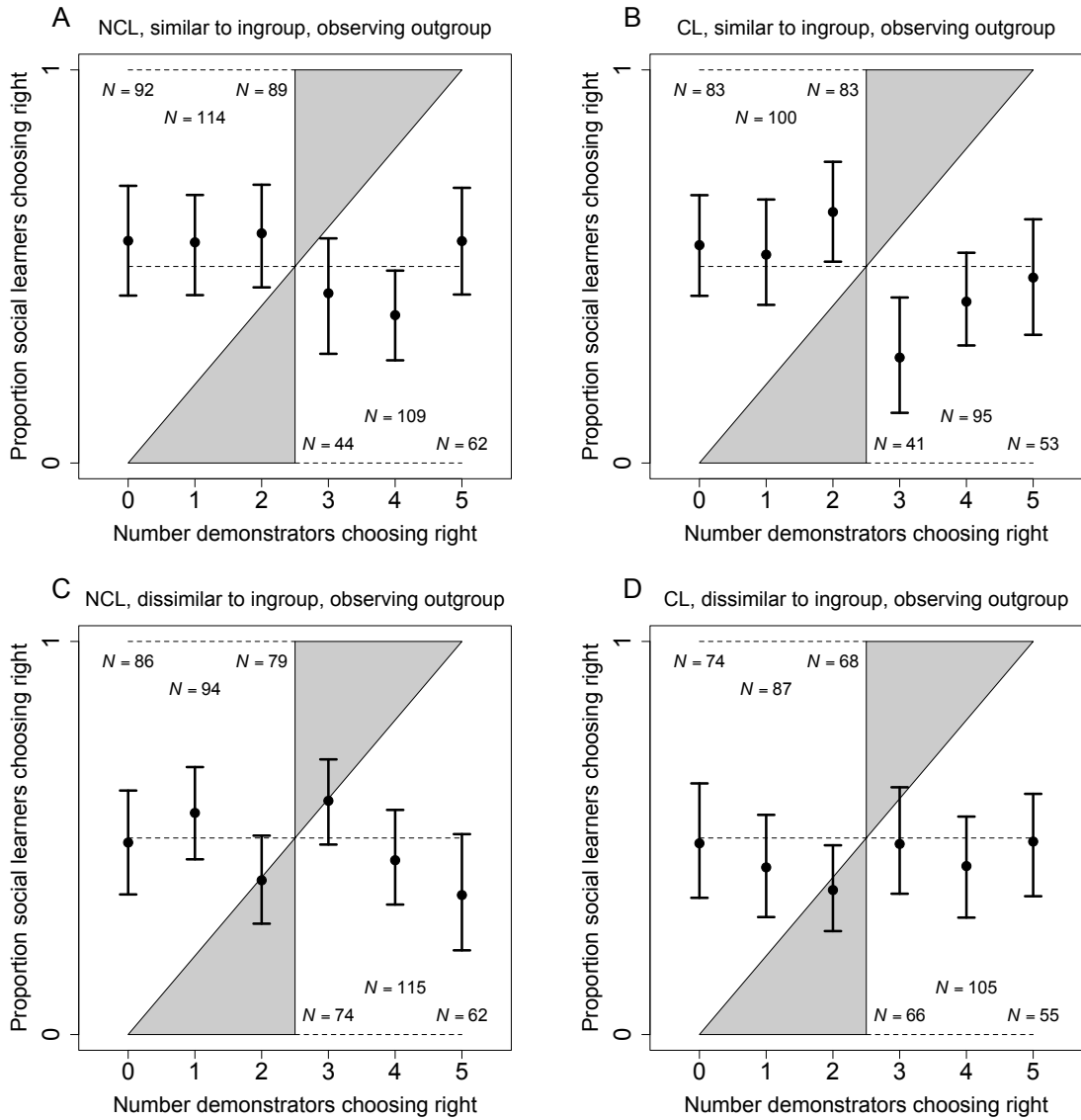


Figure 1.6: Right urn choices by social learners in treatments in which they observed **outgroup** demonstrators (results from **Zurich**). The points show the rate at which social learners chose the right urn as a function of the number of demonstrators choosing the right urn in the final round of blocks. The number of observations is vertically aligned with each point. **A** and **B** show treatments in which social learners are **similar** to the ingroup and they observe the outgroup. **C** and **D** show treatments in which social learners are **dissimilar** to the ingroup and they observe the outgroup. **A** and **C** represent “no cognitive load” treatments (NCL) and **B** and **D** show “cognitive load” treatments (CL). Error bars are 95% bootstrapped confidence intervals clustered on social learner. The gray area in each plot is consistent with conformist cultural transmission (Boyd and Richerson, 1985). The diagonal shown as a solid black line is consistent with unbiased social learning. Finally, the dashed lines provide additional points of reference at 0, 0.5, and 1.

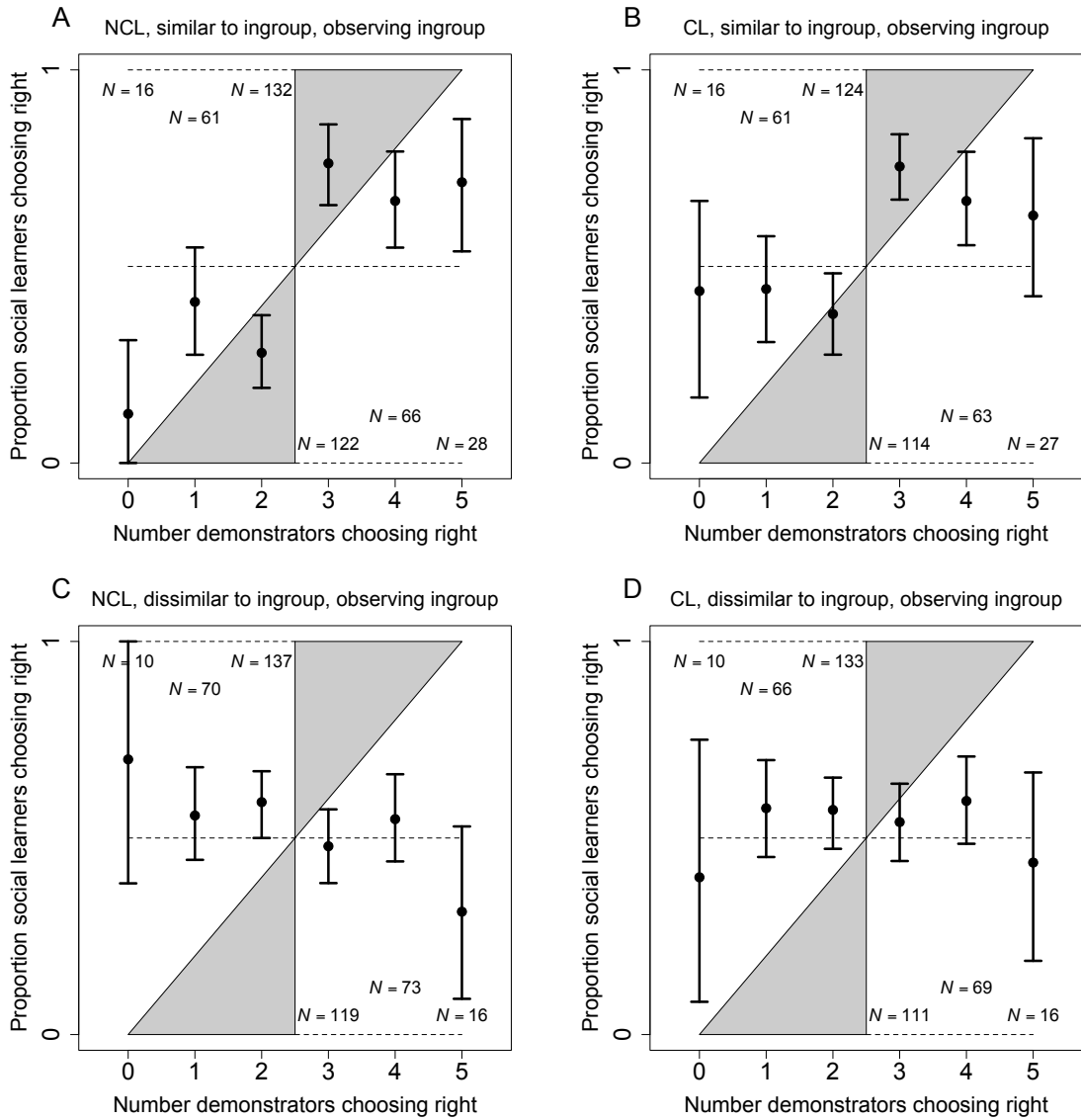


Figure 1.7: Right urn choices by social learners in treatments in which they observed **ingroup** demonstrators (results from **Nairobi**). The points show the rate at which social learners chose the right urn as a function of the number of demonstrators choosing the right urn in the final round of blocks. The number of observations is vertically aligned with each point. **A** and **B** show treatments in which social learners are **similar** to ingroup demonstrators and they observe the ingroup. **C** and **D** show treatments in which social learners are **dissimilar** to their ingroup and they observe the ingroup. **A** and **C** represent “no cognitive load” treatments (NCL) and **B** and **D** show “cognitive load” treatments (CL). The error bars are 95% bootstrapped confidence intervals which we clustered on social learner. The gray area in each plot is consistent with conformist cultural transmission (Boyd and Richerson, 1985). The diagonal shown as a solid black line is consistent with unbiased social learning. Finally, the dashed lines provide additional points of reference at 0, 0.5, and 1.

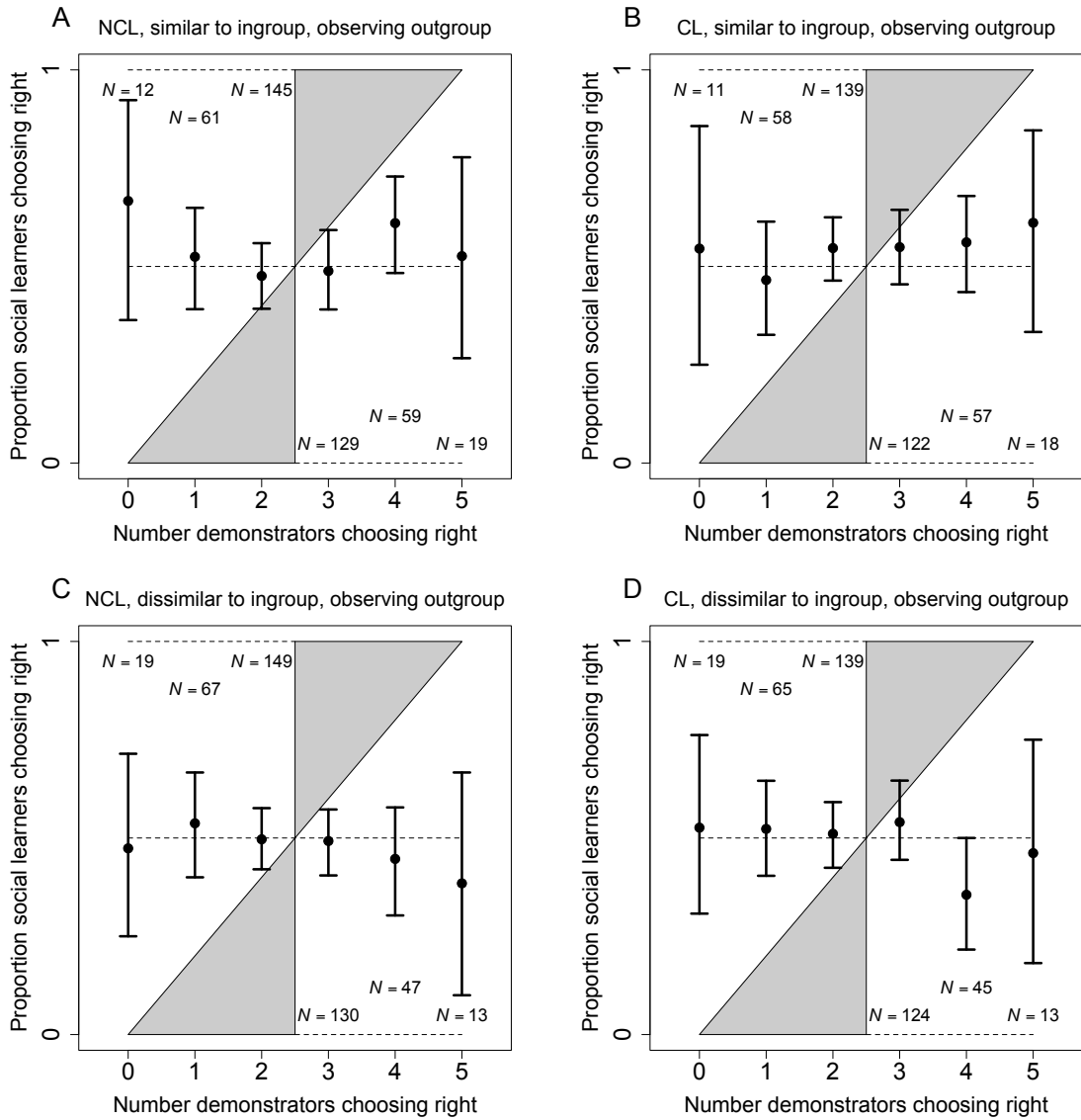


Figure 1.8: Right urn choices by social learners in treatments in which they observed **outgroup** demonstrators (results from **Nairobi**). The points show the rate at which social learners chose the right urn as a function of the number of demonstrators choosing the right urn in the final round of blocks. The number of observations is vertically aligned with each point. **A** and **B** show treatments in which social learners are **similar** to the ingroup and they observe the outgroup. **C** and **D** show treatments in which social learners are **dissimilar** to the ingroup and they observe the outgroup. **A** and **C** represent “no cognitive load” treatments (NCL) and **B** and **D** show “cognitive load” treatments (CL). Error bars are 95% bootstrapped confidence intervals clustered on social learner. The gray area in each plot is consistent with conformist cultural transmission (Boyd and Richerson, 1985). The diagonal shown as a solid black line is consistent with unbiased social learning. Finally, the dashed lines provide additional points of reference at 0, 0.5, and 1.

Optimal behaviour

We want to further analyse whether the fact that social learners did not respond to social information equally affected their payoffs. In other words, we ask whether social learners adapted symmetrically to the different types of social information and thus performed equally well in choosing the urn containing more marbles of the colour yielding points in the different treatments. As mentioned previously, the four within-subjects treatments are informationally equivalent in the sense that social learners have the necessary information to maximise their payoffs in all four treatments by either following the minority or majority of observed choices. For both the Zurich and the Nairobi data we run a multivariate logistic regression of social learners choosing optimally on the centred proportion of demonstrators who chose optimally and on the treatment dummies (see equation (2) in section 1 in the supplementary materials for the model). The results are shown in table [1.1](#).

Based on this regression we conducted pairwise linear hypothesis tests between all eight treatments, which yields 28 comparisons in total. The results of those pairwise linear hypothesis tests are shown in table [1.2](#).

The results shown in table [1.1](#) and table [1.2](#) support our analysis of the pure behaviour of social learners. Importantly, social learners often perform better when observing ingroup demonstrators than outgroup demonstrators, and they sometimes perform better when observing similar demonstrators than when observing dissimilar demonstrators. Overall, these findings hold for the Zurich and the Nairobi data, but the results are less clear in the case of the Nairobi data. When holding the two within-subjects treatments constant, only one treatment comparison revealed a significantly better performance for participants without cognitive load when compared with a treatment where cognitive load was implemented. This was the case for the Zurich data in the treatment where social learners were similar to their ingroup demonstrators, and they observed outgroup demonstrators.

We also run a regression with the control variables age, gender, and field of study (see supplementary materials). The control variables age and gender are not significant for the

Table 1.1: Likelihood of a social learner choosing optimally (log. regression). The omitted category is dissimilar to ingroup, observing the outgroup, and no cognitive load. The proportion of demonstrators choosing optimally is centered by subtracting 0.5. The third and fifth columns show 95% confidence intervals. Those are based on robust standard errors clustered on the social learner (shown in parentheses in the second and fourth columns). Asterisks denote the level of significance of the p-values: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Parameter	Zurich		Nairobi	
	Estimate	95 % CI	Estimate	95 % CI
Intercept	0.066 (0.135)	[-0.198, 0.330]	-0.058 (0.104)	[-0.262, 0.147]
Centered proportion of dem. choosing optimally	2.522 *** (0.260)	[2.012, 3.032]	0.138 (0.188)	[-0.232, 0.507]
Similar to ingroup, observe ingroup, no cognitive load	0.624 *** (0.180)	[0.272, 0.977]	0.350 * (0.138)	[0.080, 0.620]
Dissimilar to ingroup, observe ingroup, no cognitive load	0.377 * (0.149)	[0.086, 0.669]	0.029 (0.134)	[-0.233, 0.291]
Similar to ingroup, observe outgroup, no cognitive load	0.359 * (0.155)	[0.055, 0.664]	0.144 (0.146)	[-0.143, 0.431]
Similar to ingroup, observe ingroup, cognitive load	0.566 ** (0.189)	[0.195, 0.937]	0.165 (0.150)	[-0.128, 0.458]
Dissimilar to ingroup, observe ingroup, cognitive load	0.351 (0.208)	[-0.058, 0.760]	-0.011 (0.147)	[-0.299, 0.276]
Similar to ingroup, observe outgroup, cognitive load	-0.053 (0.205)	[-0.456, 0.350]	0.041 (0.145)	[-0.244, 0.326]
Dissimilar to ingroup, observe outgroup, cognitive load	-0.083 (0.199)	[-0.474, 0.308]	-0.191 (0.144)	[-0.474, 0.091]

Zurich and Nairobi data. For the Zurich data we observe that students with natural sciences or social sciences as their field of study perform significantly better than students with humanities as their field of study. For the data from Nairobi we do not observe any significant results associated with the control variable field of study.

Table 1.2: Pairwise comparisons of treatments based on the regression shown in table [1.1](#). Specifically, we estimate treatment differences and test the linear hypotheses “treatment 1 - treatment 2 = 0”. Asterisks denote the level of significance of the p-values: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Treatment Comparison		Zurich	Nairobi
Similar Ingroup NCL	- Dissimilar Ingroup NCL	0.247	0.321 *
Similar Ingroup NCL	- Similar Outgroup NCL	0.265	0.206
Similar Ingroup NCL	- Similar Ingroup CL	0.059	0.185
Similar Ingroup NCL	- Dissimilar Ingroup CL	0.273	0.362 *
Similar Ingroup NCL	- Similar Outgroup CL	0.677 ***	0.309 *
Similar Ingroup NCL	- Dissimilar Outgroup CL	0.707 ***	0.541 ***
Similar Ingroup NCL	- Dissimilar Outgroup NCL	0.624 ***	0.350 *
Dissimilar Ingroup NCL	- Similar Outgroup NCL	0.018	-0.115
Dissimilar Ingroup NCL	- Similar Ingroup CL	-0.188	-0.136
Dissimilar Ingroup NCL	- Dissimilar Ingroup CL	0.026	0.041
Dissimilar Ingroup NCL	- Similar Outgroup CL	0.431 *	-0.011
Dissimilar Ingroup NCL	- Dissimilar Outgroup CL	0.460 *	0.220
Dissimilar Ingroup NCL	- Dissimilar Outgroup NCL	0.377 **	0.029
Similar Outgroup NCL	- Similar Ingroup CL	-0.206	-0.021
Similar Outgroup NCL	- Dissimilar Ingroup CL	0.008	0.156
Similar Outgroup NCL	- Similar Outgroup CL	0.412 *	0.104
Similar Outgroup NCL	- Dissimilar Outgroup CL	0.442 *	0.335 *
Similar Outgroup NCL	- Dissimilar Outgroup NCL	0.359 *	0.144
Similar Ingroup CL	- Dissimilar Ingroup CL	0.215	0.177
Similar Ingroup CL	- Similar Outgroup CL	0.619 ***	0.124
Similar Ingroup CL	- Dissimilar Outgroup CL	0.648 ***	0.356 **
Similar Ingroup CL	- Dissimilar Outgroup NCL	0.566 ***	0.165
Dissimilar Ingroup CL	- Similar Outgroup CL	0.404	-0.052
Dissimilar Ingroup CL	- Dissimilar Outgroup CL	0.434 *	0.180
Dissimilar Ingroup CL	- Dissimilar Outgroup NCL	0.351 *	-0.011
Similar Outgroup CL	- Dissimilar Outgroup CL	0.030	0.232
Similar Outgroup CL	- Dissimilar Outgroup NCL	-0.053	0.041
Dissimilar Outgroup CL	- Dissimilar Outgroup NCL	-0.083	-0.191

Social learning strategies

In addition to the more general analysis of whether and how well the social learners were able to adjust to the different types of social information in the four treatments, we also wanted to understand whether the social learners followed specific social learning strategies. To do so, we categorised social learning into specific strategies. Surprisingly, this analysis on the individual level does not only reveal variation between participants but also a major difference between the two groups of participants from Zurich and Nairobi. More precisely, figures [1.9](#) and [1.10](#) illustrate two key findings. First, social learners differ in their use of social information. While many participants exhibit the expected behaviour of following the majority or the minority, a substantial number of participants follow other strategies. Second, the degree of heterogeneity is much higher among social learners in Nairobi than in Zurich. The analysis of disaggregated social learning strategies shown in figure [1.9](#) and [1.10](#) is provided for the treatments with cognitive load in which social learners observed ingroup demonstrators. The analysis for the other treatments is shown in section 2.2.4 in the supplementary materials.

Panel A in figure [1.9](#) shows the results for social learners under cognitive load in the treatment in which they were **similar** to their ingroup demonstrators and they observed their ingroup demonstrators. In this treatment it is optimal to always follow the majority because the likelihood of sharing the winning colour with the observed demonstrators is 90%. Most social learners in Zurich were aware of this, and they did in fact always follow the majority of observed demonstrators (see category “Maj”). In contrast, **Panel B** in figure [1.9](#) shows the results for social learners under cognitive load in the treatment in which they were **dissimilar** to their ingroup demonstrators and they observed their ingroup demonstrators. In this treatment it was optimal to always follow the minority because the likelihood of sharing the winning colour with the observed demonstrators was only 10%. Again, most social learners in Zurich recognised that this strategy was optimal, and they did always follow the minority of observed demonstrators (see category “Min”). Very few social learners followed social learning strategies that deviated from following the minority, such as following the majority when it

was optimal to follow the minority and vice versa (see “Maj” and “Min” categories), or always choosing the left or right urn irrespective of the demonstrators’ behaviour (category “U”).

The social learners that we could not categorise into “Maj”, “Min”, or “U” are represented by the $\hat{\beta}_k$ estimate, which is the probability of a social learner choosing the right urn in block j , given that the centred proportion of demonstrators, x_j , chose the right urn. Those social learners who were more likely to follow the observed demonstrators than doing the opposite are represented by positive $\hat{\beta}_k$ values. More social learners have positive $\hat{\beta}_k$ values in Panel A, which is consistent with the optimal strategy in that treatment. In contrast, those social learners who were more likely to do the opposite than observed demonstrators’ behaviour are represented by negative $\hat{\beta}_k$ values. Again, the majority of social learners in Panel B have negative $\hat{\beta}_k$ values which is consistent with the optimal strategy. Social learners with $|\hat{\beta}_k| > 10$ responded clearly to demonstrators’ choices, while others did not (Efferson et al., 2016).

The data from Nairobi shown in figure 1.10 paints a completely different picture than the data from Zurich. Figure 1.10 shows the same treatments as figure 1.9. Again, the mode of social learners followed the majority of observed demonstrators when this was optimal (Panel A) and the mode of social learners followed the minority when that was optimal (Panel B). Surprisingly, however, a large proportion of social learners followed other social learning strategies. For example, as can be seen in both Panel A and B, many social learners in Nairobi always chose either the right or left urn irrespective of observed demonstrators’ behaviour (category “U”). In addition, a high number of social learners could not be categorised into “Maj”, “Min”, or “U”. Even though none of the $\hat{\beta}_k$ estimates are significant, these results indicate that social learners in Nairobi exhibited much more variation in social learning than social learners in Zurich. We explore possible explanations for this in the discussion section.

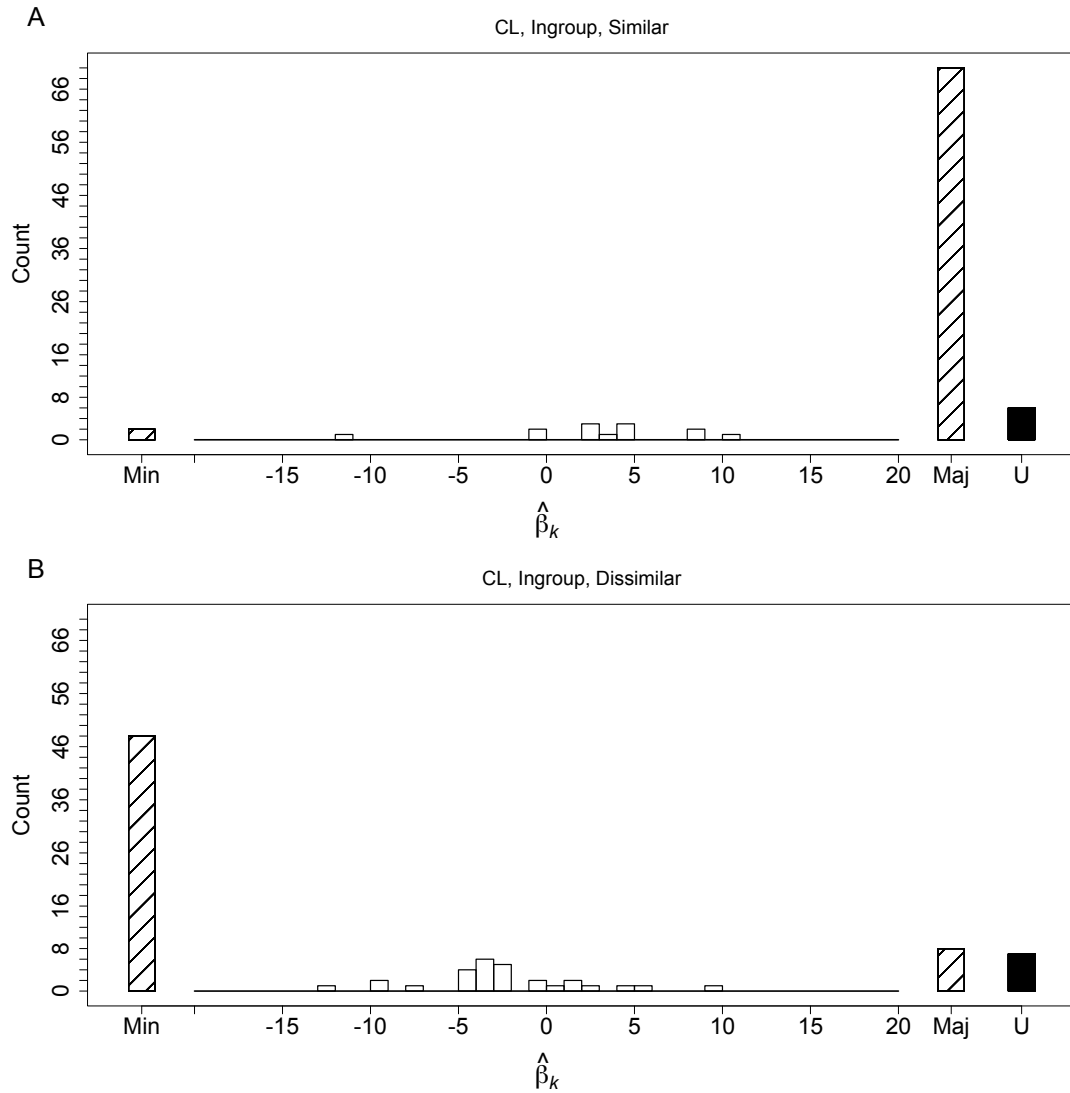


Figure 1.9: Disaggregated social learning strategies (results from **Zurich**). Always following the minority choice among demonstrators is labeled as “Min” type. Always following the majority is a “Maj” type. A social learner who always chose left or always chose right irrespective of demonstrators’ behaviour is described as a “U” type (Unconditional). Some social learners did not fall into one of these three categories. For those we estimated the social learning function in the following way. Let $Y_{jk} \in \{0, 1\}$ indicate if social learner k chose the right urn in block j , and let x_j be the centred proportion of demonstrators choosing the right urn. $\hat{\beta}_k$ was estimated by fitting $P(Y_{jk} = 1) = \exp\{\beta_k x_j\} / (1 + \exp\{\beta_k x_j\})$ with maximum likelihood. Panel **A** shows distributions over types for the blocks in which social learners in the cognitive load treatment are **similar** to their ingroup and they **observe ingroup** demonstrators. Panel **B** shows the blocks in which social learners in the cognitive load treatment are **dissimilar** to their ingroup and they **observe ingroup** demonstrators. Gray bars would show $\hat{\beta}_k$ estimates significant at the 5% level. There are no significant estimates. Social learners who followed the minority (Min) or majority (Maj) and social learners with extreme values of $\hat{\beta}_k$ (e.g. $|\hat{\beta}_k| > 10$) clearly responded to social information, while the rest did not.

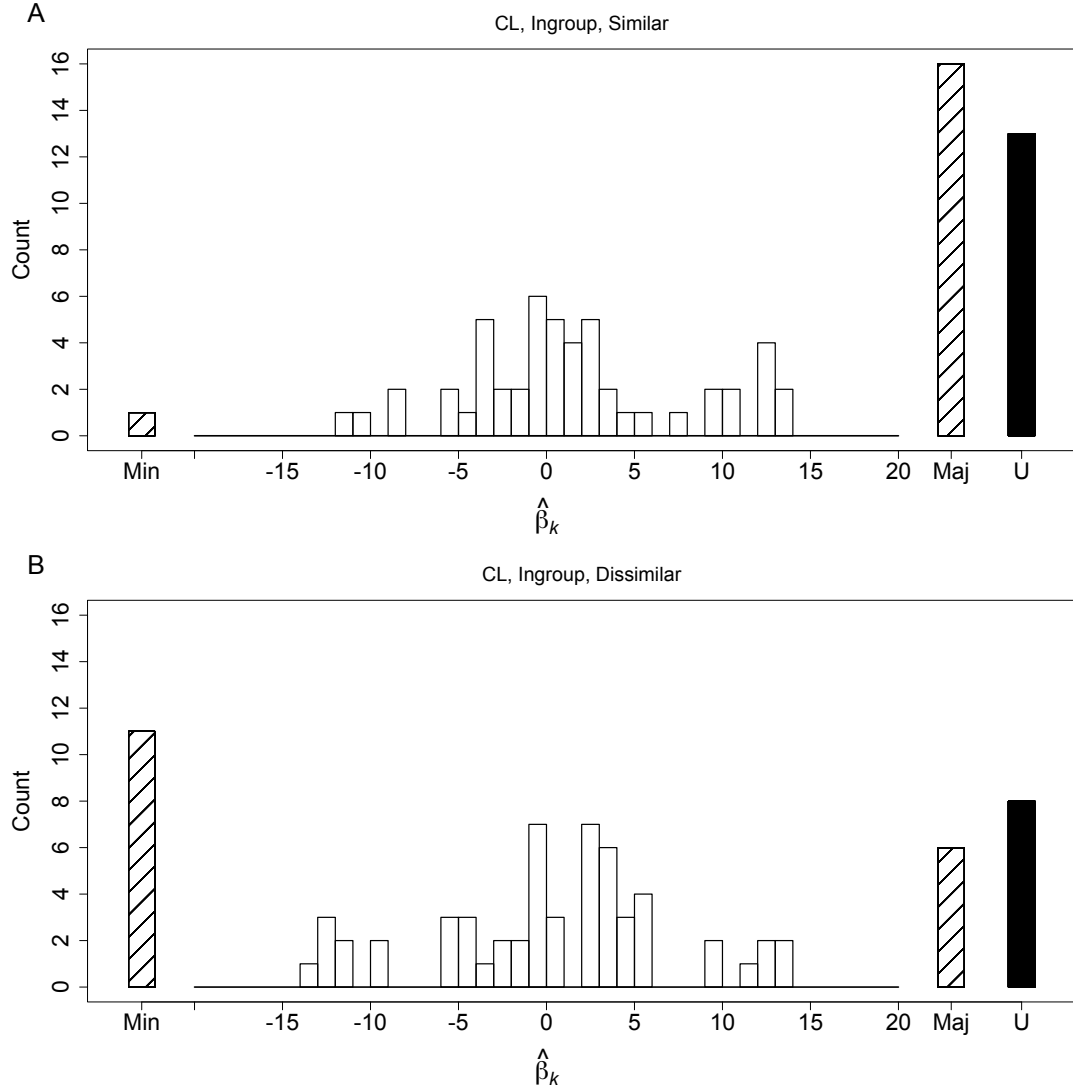


Figure 1.10: Disaggregated social learning strategies (results from **Nairobi**). Always following the minority choice among demonstrators is labeled as “Min” type. Always following the majority is a “Maj” type. A social learner who always chose left or always chose right irrespective of demonstrators’ behaviour is described as a “U” type (Unconditional). Some social learners did not fall into one of these three categories. For those we estimated the social learning function in the following way. Let $Y_{jk} \in \{0, 1\}$ indicate if social learner k chose the right urn in block j , and let x_j be the centred proportion of demonstrators choosing the right urn. $\hat{\beta}_k$ was estimated by fitting $P(Y_{jk} = 1) = \exp\{\beta_k x_j\} / (1 + \exp\{\beta_k x_j\})$ with maximum likelihood. Panel **A** shows distributions over types for the blocks in which social learners in the cognitive load treatment are **similar** to their ingroup and they **observe ingroup** demonstrators. Panel **B** shows the blocks in which social learners in the cognitive load treatment are **dissimilar** to their ingroup and they **observe ingroup** demonstrators. Gray bars would show $\hat{\beta}_k$ estimates significant at the 5% level. There are no significant estimates. Social learners who followed the minority (Min) or majority (Maj) and social learners with extreme values of $\hat{\beta}_k$ (e.g. $|\hat{\beta}_k| > 10$) clearly responded to social information, while the rest did not.

1.4 Discussion

One of our main hypotheses was that cognitive load induces heterogeneity in social learning and causes social learning to be less adaptive. Surprisingly, this was not the case. A potential explanation could be that our cognitive load manipulation did not work. However, our results show that the cognitive load manipulation worked (see supplementary materials 2.2.1). We also think that the potential explanation based on the results of the questionnaire mentioned in section [1.3.1](#) is unlikely. However, we do not currently have the data to completely rule out the possibility that certain social learners in the NCL were distracted to some extent if they tried to memorise the numbers. Studies trying to implement the same experimental design could add more questions to the questionnaire to find out more about the thought processes and motivations of the participants. A better approach is most likely to modify the design in future studies to further reduce the likelihood of social learners in the NCL treatment attempting to memorise the numbers. For instance, one could remove the step in which the social learners in the NCL treatment are shown the 7-digit numbers. This step could be replaced by a pop-up window showing a countdown of the seconds corresponding to the 8 seconds in which the social learners have to memorise the numbers in the CL treatment.

A completely different explanation for the observation that cognitive load did not affect behaviour is that certain social learning biases might be too robust for cognitive load to matter. Importantly, we designed four informationally equivalent treatments in the following sense. The social learners received all the necessary information to always choose the optimal strategy, that is to follow either the minority or the majority of the observed demonstrators. If social learners realise this, they could in principle adjust symmetrically between the four different treatments to maximise payoffs in each treatment. However, this would not be the case if social learners have a developed social cognition that strongly biases them to learn more easily from certain types of social information than from others ([Efferson et al., 2016](#)). For example, social cognition might have evolved under conditions in which individuals

were similar to individuals from whom they learned in crucial ways, such that it was always beneficial to follow the majority. Importantly, the four informationally equivalent treatments only vary by whether participants are shown the choices of ingroup or outgroup demonstrators and by the level of similarity to the ingroup demonstrators. We did this to examine whether social learners exhibit biases towards the ingroup, towards demonstrators with high similarity, or both.

Specifically, if we observe that social learners do not adapt symmetrically between the four informationally equivalent treatments and cognitive load does not influence this effect, one possible explanation is that social learners are cognitively biased to learn more easily from the social information provided in certain treatments than in others. Observing such robust biases does, however, contradict the findings of recent empirical research which found that individuals are quite flexible in their use of social information (Efferson et al., 2016; Bellamy et al., 2022). What we found is that social learners might be especially prone to ingroup bias. The fact that social learners managed relatively well to distinguish between similar and dissimilar demonstrators when observing ingroup demonstrators, but not when observing outgroup demonstrators, suggests the following hypothesis. Human psychology may have evolved to be more attuned to social cues from ingroup members due to frequent interactions with them. This hypothesis is consistent with the evidence for ingroup favouritism and outgroup aversion, which shows that it is easier for individuals to trust, interact and cooperate with ingroup members than with outgroup members (Bernhard et al., 2006; Efferson et al., 2008; Romano et al., 2017; Smaldino et al., 2017; Romano et al., 2024). Importantly, recent studies indicate that individuals rely more on social information from ingroup members than from outgroup members (Buttelmann et al., 2013; Howard et al., 2015; Kang et al., 2021; Zou and Xu, 2023). Many of these studies are conducted with infants, reinforcing the idea that humans' superior ability to evaluate social information from ingroup members could stem from an evolutionary predisposition to prioritise interpreting and integrating cues from familiar sources, which are perceived as more reliable.

However, our results should be interpreted with caution due to possible limitations, including the abstract use of minimal groups based on symbols (Diehl, 1990). Minimal groups may not fully capture the complexity of real-world group dynamics and social identities, which could influence the observed ingroup bias. Importantly, since we only simulate an abstraction of group membership, it is possible that the observed behaviour is not due to a cognitively evolved ingroup bias. Instead, general learning mechanisms may explain why social learners struggle to adopt optimal strategies when observing outgroup demonstrators compared to ingroup demonstrators. We indicated similarity regarding the winning colour by informing social learners whether they were very similar or dissimilar to ingroup demonstrators. Participants also learned that ingroup and outgroup demonstrators received points for different colours, meaning that in any given block, the optimal urn for one group is the opposite of the other. Thus, observing ingroup demonstrators requires only one cognitive step, whereas observing outgroup demonstrators necessitates two steps. For example, if a participant knows she is similar to ingroup demonstrators in a certain block but observes outgroup demonstrators, she must deduce that she is not similar to the observed demonstrators because the two groups earn points for different colours.

The cognitive efforts required to make this logical inference might also help explain the most interesting finding in our study, namely that the degree of heterogeneity in social learning is much higher in Nairobi than Zurich. In our experiment this means that a higher fraction of social learners in Nairobi followed suboptimal social learning strategies than in Zurich. One potential explanation for this difference between participants coming from the two subject pools in Zurich and Nairobi might be cultural. Cultural factors, both historical and recent, may cause people in Kenya to respond more strongly to social information from ingroup members than from outgroup members, regardless of other available information. And this response might be more pronounced than the Swiss tendency to react to the behaviour of ingroup versus outgroup members. Unfortunately, our data is insufficient to support this cultural explanation. Disentangling the effects of culture on behaviour is very challenging.

Recent studies in cultural evolution have begun using innovative approaches, such as regression discontinuity at language borders (Faessler et al., 2024). Future research on social learning biases could benefit from adopting these methods.

Beyond cultural differences, other factors may explain why social learning heterogeneity is significantly higher in Nairobi than in Zurich. As noted, social learners were given time during the instructions to understand the optimal strategies for each of the four treatments involving different types of social information. However, the quiz results indicate that this understanding varied among participants. At the end of the instructions, we tested participants' comprehension with a quiz, focusing on their first attempts (see section 3.7 in the supplementary materials). A majority of social learners in both Zurich and Nairobi answered all or most quiz questions correctly on their first attempt. For instance, 65% of the participants in Zurich and 79% in Nairobi correctly identified when they would earn points for the same colour of marbles as their ingroup. However, only about 42% of participants in Nairobi correctly answered that the demonstrators from the two groups received points for opposite colours on their first attempt, compared to 80% in Zurich.

Participants were given explanations for incorrect answers and had to retake the quiz until they answered correctly. However, the initial misunderstandings in Nairobi might have led to suboptimal learning strategies in the outgroup treatments. Our data on participants' fields of study provide further insights into these results. Specifically, a higher proportion of social learners in Zurich study natural sciences compared to those in Nairobi. Although this is not equivalent to an IQ test, students in natural sciences, including those studying mathematics, which we categorised under natural sciences, may be more accustomed to solving logical problems. If the observed outcomes are not due to an ingroup bias but rather require general learning mechanisms to infer optimal strategies in the outgroup treatments, then this familiarity could help students in the natural sciences in Zurich more easily grasp the strategy of following the majority of outgroup demonstrators when they are dissimilar to ingroup members compared to their Nairobi counterparts. Indeed, participants studying

natural sciences in Zurich achieved the highest fractions of correct first attempts across all quiz questions, outperforming their peers from other study fields in Zurich and all students in Nairobi (see Tables 6 and 7 in Section 3.7 of the supplementary materials).

Our study cannot entirely rule out the possibility that the observed ingroup bias and differences between the two subject pools are attributable to general learning mechanisms, such as educational background or variations in understanding the instructions. To address these potential confounding factors, future studies could incorporate IQ tests and expand the questionnaire to account for influences beyond social cognition that might affect behaviour. Importantly, to better understand the gene-culture interactions in evolved social cognition, future research should focus on making group identities more salient and credible instead of relying on minimal groups using abstract symbols (Diehl, 1990). One promising approach is to use group memberships based on political affiliation, as this group marker seems to elicit strong responses (Ehret et al., 2022). Therefore, a straightforward improvement to our study design and similar research on social learning biases would be to replace abstract symbols with real-world group markers. Laboratory studies using this approach could be combined with field studies to enhance external validity. The increasing political polarisation surrounding climate change debates (Falkenberg et al., 2022) provides an urgent context for exploring how individuals respond to social cues from ingroup versus outgroup members. This polarisation may be further intensified by the spread of misinformation, which frequently accompanies climate change discussions (Treen et al., 2020). Examining these dynamics is particularly relevant for social learning research, as evolved cognitive biases can amplify the acceptance of unfounded beliefs (Efferson, McKay and Fehr, 2020; Sulik et al., 2021). Hence, future studies could for example investigate whether individuals are more adept at recognising misinformation on social media from their political peers before adopting it, compared to recognising it from their political outgroup.

1.5 Conclusion

We show that individuals display considerable variation in their responses to social information. This diversity in social learning may be rooted in evolved social cognition or general learning mechanisms. Importantly, these individual-level variations lead to substantial distinctions in the distribution of social learning strategies at the group level. The greater heterogeneity in social learning observed in Nairobi compared to Zurich might be partially explained by cultural factors, educational backgrounds, or differing levels of comprehension during the experiment. If cultural factors explain why individuals vary in social learning, our study underscores the importance of expanding cultural evolution research beyond the predominantly Western populations typically studied (Henrich et al., 2010).

However, even if other factors besides culture contribute to the diversity in social learning, our study suggests that future research involving diverse subject pools will be crucial. Behavioural variation between groups plays a crucial role in cultural group selection (Henrich and Boyd, 1998; Henrich, 2004). Improving our understanding of the mechanisms at the individual level that lead to group-level patterns might help us better comprehend why social norm changes can or cannot occur (Efferson et al., 2023). This insight is not only valuable from a theoretical standpoint but also has significant implications for applied research. An urgent area is climate change, where debates are often linked to political affiliations and are becoming increasingly polarised (Falkenberg et al., 2022). Such divisions along normative lines, especially when tied to group identities, could hinder cultural shifts toward sustainable norms (Efferson, Vogt and Fehr, 2020; Ehret et al., 2022). A combination of basic research on social learning involving different subject pools, as in this study, and field studies focusing more on cultural mechanisms is most likely needed to address the challenges associated with the transition to a sustainable future (Efferson et al., 2023; Constantino et al., 2022). We therefore suggest that the research presented here should be considered as one of the first steps of such a potential research agenda.

Acknowledgements and funding

For their helpful comments, we thank seminar participants at the Lisbon School of Economics and Management, the University College London, and the University of Lausanne. The authors acknowledge support from the Swiss National Science Foundation (Grant Nr. 100018_185417/1 and 100018_215540/1).

Author contributions

LvF designed the study with feedback from SV and CE. LvF programmed the experiment. LvF analysed the data with feedback from CE. LvF wrote the article with feedback from CE and SV.

Research transparency and reproducibility

The supplementary materials, including the data and data analysis, and the preregistration are available on <https://osf.io/mzfq5/>. The code for the experiment can be found here: https://github.com/LukasVonFluee/oTree_SocialLearning_experiment.git.

Bibliography

- Aoki, K. and Feldman, M. W. (2014). Evolution of learning strategies in temporally and spatially variable environments: a review of theory, *Theoretical population biology* **91**: 3–19.
- Bellamy, A., McKay, R., Vogt, S. and Efferson, C. (2022). What is the extent of a frequency-dependent social learning strategy space?, *Evolutionary Human Sciences* **4**: e13.
- Bernhard, H., Fischbacher, U. and Fehr, E. (2006). Parochial altruism in humans, *Nature* **442**(7105): 912–915.
- Boyd, R. and Richerson, P. J. (1985). *Culture and the Evolutionary Process*, University of Chicago press.

- Buckert, M., Oechssler, J. and Schwieren, C. (2017). Imitation under stress, *Journal of Economic Behavior & Organization* **139**: 252–266.
- Buttelmann, D., Zmyj, N., Daum, M. and Carpenter, M. (2013). Selective imitation of in-group over out-group members in 14-month-old infants, *Child Development* **84**(2): 422–428.
- Chen, D. L., Schonger, M. and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments, *Journal of Behavioral and Experimental Finance* **9**: 88–97.
- Constantino, S. M., Sparkman, G., Kraft-Todd, G. T., Bicchieri, C., Centola, D., Shell-Duncan, B., Vogt, S. and Weber, E. U. (2022). Scaling up change: A critical review and practical guide to harnessing social norms for climate action, *Psychological Science in the Public Interest* **23**(2): 50–97.
- Delfino, A., Marengo, L. and Ploner, M. (2016). I did it your way. an experimental investigation of peer effects in investment choices, *Journal of Economic Psychology* **54**: 113–123.
- Diehl, M. (1990). The minimal group paradigm: Theoretical explanations and empirical findings, *European Review of Social Psychology* **1**(1): 263–292.
- Duffy, S. and Smith, J. (2014). Cognitive load in the multi-player prisoner’s dilemma game: Are there brains in games?, *Journal of Behavioral and Experimental Economics* **51**: 47–56.
- Efferson, C., Lalive, R., Cacault, M. P. and Kistler, D. (2016). The evolution of facultative conformity based on similarity, *PLOS One* **11**(12): e0168551.
- Efferson, C., Lalive, R. and Fehr, E. (2008). The coevolution of cultural groups and ingroup favoritism, *Science* **321**(5897): 1844–1849.
- Efferson, C., McKay, R. and Fehr, E. (2020). The evolution of distorted beliefs vs. mistaken choices under asymmetric error costs, *Evolutionary Human Sciences* **2**.

- Efferson, C., Vogt, S. and Fehr, E. (2020). The promise and the peril of using social influence to reverse harmful traditions, *Nature Human Behaviour* **4**(1): 55–68.
- Efferson, C., Vogt, S. and von Flüe, L. (2023). Activating cultural evolution for good when people differ from each other, in J. J. Tehrani, J. Kendal and R. Kendal (eds), *The Oxford Handbook of Cultural Evolution*.
- Ehret, S., Constantino, S. M., Weber, E. U., Efferson, C. and Vogt, S. (2022). Group identities can undermine social tipping after intervention, *Nature Human Behaviour* pp. 1–11.
- Faessler, L., Lalive, R. and Efferson, C. (2024). How culture shapes choices related to fertility and mortality: Causal evidence at the swiss language border, *Evolutionary Human Sciences* **6**: e28.
- Falkenberg, M., Galeazzi, A., Torricelli, M., Di Marco, N., Larosa, F., Sas, M., Mekacher, A., Pearce, W., Zollo, F., Quattrocioni, W. et al. (2022). Growing polarization around climate change on social media, *Nature Climate Change* **12**(12): 1114–1121.
- Granovetter, M. (1978). Threshold models of collective behavior, *American Journal of Sociology* **83**(6): 1420–1443.
- Haushofer, J. and Fehr, E. (2014). On the psychology of poverty, *Science* **344**(6186): 862–867.
- Henrich, J. (2001). Cultural transmission and the diffusion of innovations: Adoption dynamics indicate that biased cultural transmission is the predominate force in behavioral change, *American Anthropologist* **103**(4): 992–1013.
- Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation, *Journal of Economic Behavior & Organization* **53**(1): 3–35.
- Henrich, J. and Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences, *Evolution and Human Behavior* **19**(4): 215–241.

- Henrich, J., Heine, S. J. and Norenzayan, A. (2010). The weirdest people in the world?, *Behavioral and Brain Sciences* **33**(2-3): 61–83.
- Howard, L. H., Henderson, A. M., Carrazza, C. and Woodward, A. L. (2015). Infants' and young children's imitation of linguistic in-group and out-group informants, *Child development* **86**(1): 259–275.
- Jacquet, P. O., Wyart, V., Desantis, A., Hsu, Y.-F., Granjon, L., Sergent, C. and Waszak, F. (2018). Human susceptibility to social influence and its neural correlates are related to perceived vulnerability to extrinsic morbidity risks, *Scientific Reports* **8**(1): 1–18.
- Kang, P., Burke, C. J., Tobler, P. N. and Hein, G. (2021). Why we learn less from observing outgroups, *Journal of Neuroscience* **41**(1): 144–152.
- Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M. and Jones, P. L. (2018). Social learning strategies: Bridge-building between fields, *Trends in Cognitive Sciences* **22**(7): 651–665.
- Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem, *The Review of Economic Studies* **60**(3): 531–542.
- Manski, C. F. (2000). Economic analysis of social interactions, *Journal of Economic Perspectives* **14**(3): 115–136.
- Mesoudi, A., Chang, L., Dall, S. R. and Thornton, A. (2016). The evolution of individual and cultural variation in social learning, *Trends in Ecology & Evolution* **31**(3): 215–225.
- Mesoudi, A., Chang, L., Murray, K. and Lu, H. J. (2015). Higher frequency of social learning in China than in the West shows cultural variation in the dynamics of cultural evolution, *Proceedings of the Royal Society B: Biological Sciences* **282**(1798): 20142209.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information., *Psychological Review* **63**(2): 81.

- Murray, D. R. and Schaller, M. (2012). Threat (s) and conformity deconstructed: Perceived threat of infectious disease and its implications for conformist attitudes and behavior, *European Journal of Social Psychology* **42**(2): 180–188.
- Muthukrishna, M., Morgan, T. J. and Henrich, J. (2016). The when and who of social learning and conformist transmission, *Evolution and Human Behavior* **37**(1): 10–20.
- Muthukrishna, M. and Schaller, M. (2020). Are collectivistic cultures more prone to rapid transformation? Computational models of cross-cultural differences, social network structure, dynamic social influence, and cultural change, *Personality and Social Psychology Review* **24**(2): 103–120.
- Romano, A., Balliet, D., Yamagishi, T. and Liu, J. H. (2017). Parochial trust and cooperation across 17 societies, *Proceedings of the National Academy of Sciences* **114**(48): 12702–12707.
- Romano, A., Gross, J. and De Dreu, C. K. (2024). The nasty neighbor effect in humans, *Science Advances* **10**(26): eadm7968.
- Shiv, B. and Fedorikhin, A. (1999). Heart and mind in conflict: The interplay of affect and cognition in consumer decision making, *Journal of Consumer Research* **26**(3): 278–292.
- Smaldino, P. E., Janssen, M. A., Hillis, V. and Bednar, J. (2017). Adoption as a social marker: Innovation diffusion with outgroup aversion, *The Journal of Mathematical Sociology* **41**(1): 26–45.
- Sulik, J., Efferson, C. and McKay, R. (2021). Collectively jumping to conclusions: Social information amplifies the tendency to gather insufficient data., *Journal of Experimental Psychology: General* .
- Treen, K. M. d., Williams, H. T. and O’Neill, S. J. (2020). Online misinformation about climate change, *Wiley Interdisciplinary Reviews: Climate Change* **11**(5): e665.

Young, H. P. (2009). Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning, *American Economic Review* **99**(5): 1899–1924.

Zou, W. and Xu, X. (2023). Ingroup bias in a social learning experiment, *Experimental Economics* **26**(1): 27–54.

Chapter 2

Green preferences sustain
greenwashing - Challenges in the
cultural transition to a sustainable
future

Abstract

Discussions of the environmental impact that revolve around monetary incentives and other easy-to-measure factors are important, but they neglect culture. Pro-environmental values will be crucial when facing sustainability challenges in the Anthropocene, and demand among green consumers is arguably critical to incentivise sustainable production. However, due to asymmetric information, consumers might not know whether the premium they pay for green production is well-spent. Reliable monitoring of manufacturers is meant to solve this problem. To see how this might work, we develop and analyse a game theoretic model of a simple buyer-seller exchange with asymmetric information, and our analysis shows that greenwashing can exist exactly because reliable monitoring co-exists with unreliable monitoring. More broadly, promoting pro-environmental values among consumers might even amplify the problem at times because a manufacturer with significant market power can exploit both consumer preferences for sustainability and trustworthy monitoring to gouge prices and in extreme cases green wash in plain sight. We discuss several strategies to address this problem. Promoting accurate beliefs and a large-scale behavioural change based on pro-environmental values might be necessary for a rapid transition to a sustainable future, but recent evidence from the cultural evolution literature highlights many important challenges.

Keywords: cultural evolution, frequency-dependent social learning, conformity, ingroup bias, similarity bias, cognitive load

2.1 Introduction

A few years ago, one of us happened to be in the room when a member of staff at a large international firm based in Switzerland gave a presentation. The grand theme of the presentation was sustainability. In practical terms, however, the presentation centred largely around certification programmes designed to identify goods that have been sustainably produced. After some introductory remarks about changing consumer preferences for socially and environmentally responsible production, and the need for firms to respond, the real star of the presentation appeared.

Specifically, the firm in question, like many others (Giuliani et al., 2017), had recently introduced its own in-house certification programme, a kind of proprietary imprimatur to signal that the consumer could buy the firm's self-certified products in good conscience. The staff member giving the presentation offered in-house certification as a great innovation, a win-win mechanism for everyone up and down the supply chain, from primary manufacturers to end consumers. She largely ignored the obvious counterargument. Namely, in-house certification faces a credibility problem. It allows the firm to obscure the criteria for certification, and it allows the firm to obscure the protocols for verifying that manufacturers have actually met the criteria in place, whatever they may be. In-house certification, in short, is vulnerable to greenwashing. Greenwashing may undermine the meaning of certification for the end consumer and by extension call into question the value of in-house certification for the firm.

That said, how exactly do incentives work when consumers face a suite of certification schemes, some of which are relatively credible, and some of which are not? A number of recent papers have highlighted the possibility that a sustainable future will require the kind of large-scale behaviour change that follows when social norms change (Nyborg et al., 2016; Brooks et al., 2018; Farmer et al., 2019; Travers et al., 2021; Constantino et al., 2022). One version of this argument is that consumer populations undergo a cultural evolutionary process that causes sustainability preferences to spread. After this has happened, consumers can

exert pressure up supply chains via their purchasing power and in turn force firms to respond (Aghion et al., 2020).

This sounds like an important mechanism, but verification and enforcement present key challenges. Many consumers may be willing to pay a premium for sustainably produced goods, but they cannot or will not invest the substantial time and energy necessary to verify that the goods in question were in fact sustainably produced (Young et al., 2010). This is where certification schemes enter the picture (Giovannucci and Ponte, 2005; Valkila et al., 2010). Product labels based on such certification schemes are meant to help consumers make informed buying decisions (Golan et al., 2001). However, it might often be difficult for consumers to verify the information provided by labels.

Credibility and reliability are not only questionable for in-house certification. Even with third-party certifications, the reliability of verifying the production process is not always perfect (Crespi and Marette, 2003b; Nilsson et al., 2004; Crespi and Marette, 2005). Importantly, when a market like the food market is flooded with eco-labels, consumers can become overwhelmed and confused by the multitude of similar labels that all indicate sustainable production (Crespi and Marette, 2003b; Nilsson et al., 2004; Horne, 2009). In other words, consumers generally possess less information about the reliability of monitoring and the production processes than manufacturers, regardless of the label in question (Golan et al., 2001; Horne, 2009). To the extent that certification is unreliable and information asymmetry exists, greenwashing may be possible (Walker and Wan, 2012; Bowen, 2014; Marquis et al., 2016; Seele and Gatti, 2017).

Hence, we ask the following questions. What effect does unreliable monitoring have when combined with culturally evolved preferences for sustainability among consumers and manufacturers? Does unreliable monitoring contaminate reliable programmes, undermine trust in monitoring in general, and thus limit the willingness of consumers to pay a premium for sustainable production? Alternatively, do unreliable schemes ride the coat tails of reliable schemes, which would mean that consumers enrich greenwashing firms by paying a premium

for unsustainable production?

To address those questions, a setting many readers will perhaps recognise from their daily lives, we develop and analyse a game theoretic model of a simple buyer-seller exchange with asymmetric information. The basic structure of the model is straightforward. The manufacturer makes two choices, namely the extent to which production is sustainable and the asking price for the good. The consumer does not observe the extent to which production is sustainable. Rather, she only observes the asking price, and after doing so she chooses either to buy the good or not. Although the consumer does not directly observe production, production is subject to one of two types of certification. On the one hand, certification is reliable in the sense that the manufacturer pays an enormous cost if she chooses low sustainability and a high price. On the other hand, certification is unreliable, and a high price does not necessarily indicate sustainable production.

Thus, our model represents a version of a well-known problem. In particular, there is a large literature studying the effects of asymmetric information on buyer-seller exchange, where the manufacturer knows the production process but the consumer does not (Darby and Karni, 1973; Laffont and Maskin, 1987). We want to examine this problem applied to greenwashing. The interest in studying greenwashing with economic analysis is relatively recent (Kirchoff, 2000; Lyon and Maxwell, 2011). One focus in this literature seems to lie on regulations. Examples include mandatory certifications and punishments in case the conditions of third-party certifications are violated (Kirchoff, 2000; Gatti et al., 2019; Garrido et al., 2020). In addition, competition among firms and green preferences among consumers are thought to put sufficient pressure on firms to produce sustainably (Aghion et al., 2020). Common among all those studies is the understanding of the firm as a profit-maximising agent who does not care about sustainability. This view stands in contrast to the literature on corporate social responsibility, which explicitly addresses the possibility that firms have values that extend beyond profit maximisation (Bénabou and Tirole, 2010). Corporate social responsibility is offered as a possible way to address the issue of greenwashing (de Freitas Netto et al., 2020).

However, to the best of our knowledge, integrating sustainability preferences of both the customer and manufacturer in a game theoretic model to analyse greenwashing is a novel approach that has not been previously explored. Importantly, the manufacturer and the consumer do not necessarily have the same preferences for sustainability in our model. In this sense, apart from the normal tension in any buyer-seller exchange with asymmetric information (Kirchoff, 2000; Garrido et al., 2020), we introduce an additional potential source of tension by allowing the two parties to disagree about the value of sustainable production. In a model similar to ours, the firm and the consumer both have different types of preferences linked to corporate social responsibility investments, but those preferences are not both specifically linked to sustainability (Wu et al., 2020). Further, the focus in that study lies on economic aspects such as different degrees of information asymmetry, the consumer’s bargaining power, and budget constraints. In contrast, we emphasise how cultural evolution of pro-environmental preferences and beliefs affect greenwashing.

Linking the discussion of greenwashing to the literature on cultural evolution is our main goal. The array of domains where cultural evolutionary theory is applied to policy is quite spectacular (Efferson et al., 2023). However, as far as we know, there do not exist any studies in the cultural evolution literature on the topic of greenwashing. Crucially, here we do not explicitly model the cultural evolution of sustainability preferences (Waring et al., 2017; Kline et al., 2018). We simply treat these preferences as parameters in the utility functions of the two types of agent. This allows us to analyse the model under different parameter values as a kind of comparative statics exercise. Our task is to examine how these preferences, once in place, shape incentives for sustainable production when manufacturers know what production methods they are using, but consumers do not.

In our model, the consumer’s beliefs are crucial to determine the circumstances under which pro-environmental values of the consumer and the manufacturer can prevent greenwashing. What happens, however, if beliefs get distorted? We look at the effects on sustainable production when cultural evolution of beliefs causes subjective probabilities to deviate

from objective probabilities (Efferson, McKay and Fehr, 2020).

More generally, we discuss recent evidence on the cultural evolution of beliefs and pro-environmental values and link them to the type of buyer-seller exchange with asymmetric information described above. With this discussion, we want to highlight some of the challenges that must be overcome culturally to address the threats that greenwashing poses to any transition to a sustainable future.

2.2 Applied cultural evolution

An interesting concept discussed in the cultural evolution literature is that a benevolent social planner can recruit endogenous evolutionary mechanisms to initiate cultural change for good (Efferson, Vogt and Fehr, 2020; Berger et al., 2023; Schimmelpfennig et al., 2021; Travers et al., 2021; Constantino et al., 2022; Ehret et al., 2022; Schimmelpfennig and Muthukrishna, 2023; Efferson et al., 2023). While regulations and economic incentives are certainly important mechanisms to address the issue of greenwashing, a social planner could try to intervene in the cultural evolution of behaviours related to sustainability. Our model accounts for two individual characteristics that might be relevant in this regard, namely pro-environmental values and beliefs. Before analysing our model, we want to highlight some challenges that a social planner might face when she tries to improve the accuracy of beliefs or increase pro-environmental values in the population.

2.2.1 Cultural evolution of beliefs

In today's digital society with countless online platforms, people face both opportunities and risks associated with learning from others. A social planner who does not want to resort to extreme measures such as banning certain communication platforms might nevertheless be interested in providing incentives to curb the spread of misinformation (Treen et al., 2020). Importantly, recent insights from the cultural evolution literature point to several mechanisms that might prove critical for such a policy intervention. A consumer's belief can be influenced

by the social information available in her network and the way the network is structured. The accuracy of information often increases in more decentralised networks, while the wisdom of crowds can be distorted in more centralised networks (Golub and Jackson, 2010; Becker et al., 2017). In addition to network features, cognitively evolved biases, such as frequency-dependent social learning, or the cognitive attraction to certain types of information, can potentially cause people to adopt mistaken beliefs or misinformation (Acerbi, 2019; Efferson, McKay and Fehr, 2020). The spread of misinformation might be amplified through homophily, polarisation, and echo chambers (Del Vicario et al., 2016; Treen et al., 2020). These insights suggest several potential ways of improving the diffusion of accurate beliefs. A social planner could for example incentivise critical thinking and warn people about misinformation (Treen et al., 2020). Alternatively, she could act preemptively and rely on social learning mechanisms, such as prestige bias, by recruiting influential people and having them share accurate information (Banerjee et al., 2020).

In our analysis, we will compare benchmark cases in which the consumer forms accurate beliefs by matching the objective probabilities in the game, to situations in which the consumer has subjective beliefs that deviate from the objective probabilities.

2.2.2 Cultural evolution of pro-environmental values and norms

A large-scale change in behaviour might be necessary if we want to achieve a sustainable future (Nyborg et al., 2016; Brooks et al., 2018; Farmer et al., 2019; Otto et al., 2020; Travers et al., 2021; Constantino et al., 2022). Hence, a better understanding of the mechanisms underlying social tipping dynamics could be critical (Bentley et al., 2014; Nyborg et al., 2016; Berger et al., 2023; Constantino et al., 2022; Efferson et al., 2023). The idea of social tipping is that once a critical mass of people commit to a particular behaviour, a self-reinforcing dynamic can be set in motion that transforms the behaviour into a new social norm (Gladwell, 2000; Constantino et al., 2022).

In the analysis we discuss further below, we model only one consumer. The consumer in

our model can care a lot or not at all about sustainability. In other words, our model might be interpreted as a comparison between a world with a minority of consumers having strong pro-environmental values to one where a majority cares about the environment.

Importantly, there are several challenges that threaten such a transition to a world with a majority of the consumer population caring about sustainability. Recent evidence points to a number of factors that may prevent tipping. Among those are group identities, out-group aversion, certain institutional settings, and heterogeneous forms of social learning (Smaldino et al., 2017; Efferson, Vogt and Fehr, 2020; Andreoni et al., 2021; Berger et al., 2023; Smaldino and Jones, 2021; Ehret et al., 2022; Efferson et al., 2023). Despite those insights helping us understand tipping dynamics better, the complexity of tipping could be much greater than we have assumed so far (Constantino et al., 2022; Efferson et al., 2023).

Given the complexity of tipping mechanisms, the question is, What can a social planner do to initiate the rapid cultural transition to a sustainable future that is so urgently needed? Clearly, we need more research to improve our understanding. However, simply waiting for better insights from research is not practical given the urgency of many environmental problems. An important insight that can inform a social planner who wants to act now is the idea that pre-existing values and preferences in a population are often much more important than the indirect effects that follow a policy intervention (Efferson, Vogt and Fehr (2020)). Importantly, mechanisms of indirect effects such as social learning can be so complex that it is impossible to predict tipping dynamics (Efferson et al. (2023)).

An approach of inducing a change in pre-existing values and preferences that has recently gained interest in the cultural evolution literature is edutainment (Efferson et al., 2023). Edutainment is a communications strategy that combines education and entertainment. The advantage of this approach is that today's technologies make it possible to reach most people at low costs and avoid biased targeting of interventions (DellaVigna and La Ferrara, 2015; La Ferrara, 2016). Avoiding biased targeting of interventions is important because biased targets in general, and amenable targets in particular, seem to reduce the effectiveness of an

intervention (Efferson, Vogt and Fehr, 2020; Efferson et al., 2023).

The discussion on the cultural evolution of beliefs makes it clear that edutainment also carries the danger of being misused to spread misinformation. Therefore, a policy intervention based on edutainment should be accompanied by careful efforts to avoid the diffusion of misinformation.

2.3 Buyer-seller exchange with information asymmetry and heterogeneous sustainability preferences

We model a trade interaction between two agents. The manufacturer-retailer in our model combines the decision making of both a manufacturer and a retailer. This agent makes choices about production and prices. We can think about this type of agent as either a single entity, such as a small-scale farmer, or a long-standing and stable partnership between a manufacturer and retailer who coordinate their choices with respect to the end consumer. In what follows, we refer to the manufacturer-retailer as the manufacturer for short and we indicate this type of agent with the subscript m . The second type of agent is the consumer, which we indicate with subscript c . Agents of both types have a preference for sustainability. We capture these preferences with $\alpha \geq 0$ for the manufacturer and with $\beta \geq 0$ for the consumer. Importantly, manufacturer and consumer preferences can be different. We allow for this possibility because a key question concerns what happens when culturally evolved preferences and norms are heterogeneous and do not align in trade interactions. This is a one-shot game, where one manufacturer produces and sells one product to one end consumer. The manufacturer chooses the sustainability of production, $s \geq 0$, and a price, $p > 0$. The consumer observes the price but not the sustainability level. The action space for the consumer is $b \in A_c = \{0, 1\}$. Specifically, after observing the price, the consumer makes a buying decision, where $b = 0$ means “not buying” and $b = 1$ “buying”. Not buying the product is equivalent to dropping out of the trade interaction. In that case, the consumer gets a utility associated with an

outside option, u_c^0 . We make the simplifying assumption that the manufacturer does not have an outside option. If the consumer does not buy the product, the manufacturer absorbs the cost of the production that has already occurred but receives no price. We derive the following utility function for the manufacturer, where the “0” superscripts on u indicate that the consumer does not buy, and the “1” superscripts on u indicate that the consumer buys.

$$\begin{aligned} u_m(s, p, | b) &= b[u_m^1] + (1 - b)[u_m^0] \\ u_m(s, p, | b) &= b[p - s + \alpha s] + (1 - b)[-s + \alpha s]. \end{aligned} \tag{2.1}$$

This is a unit-free version of the model derived from a dimensional analysis (Supplementary Information). In short, s is the sustainability level chosen by the manufacturer, and it enters her utility function once associated with a cost, which is normalised to one, and once associated with a utility based on her sustainability preference, α . Thus, from the manufacturer’s point of view, increasing the sustainability level leads to higher costs but may also yield higher utility if her sustainability preference is sufficiently strong. We assume a maximum level of sustainability, $\bar{s} > 0$, that the manufacturer can choose.

Sustainability and price are binary choice variables. Specifically, we define an exogenous cut-off value for prices such that $p_L < p_H$. Similarly, we define an exogenous cut-off value for sustainability, such that $s_L < s_H$, where we assume $s_H \leq \bar{s}$. In other words, for simplicity, we assume here that the product produced and sold by the manufacturer can be classified as having either high or low sustainability. A similar simplification is assumed for the price.

The reason for defining low and high levels for sustainability and prices is that prices can act as signals for sustainability in our model. Importantly, the consumer does not observe the manufacturer’s choice of sustainability, but she observes the price. The notion that prices can be used effectively as signals for quality and sustainability is supported by research (Rao and Monroe, 1989; Kirmani and Rao, 2000; Shiv et al., 2005; Mahenc, 2008; Brécard et al., 2009). This signalling function of prices might be especially relevant when the market of a given sector

becomes flooded with eco-labels, and consumers become overwhelmed and confused by the multitude of similar labels all indicating sustainable production (Crespi and Marette, 2003b; Nilsson et al., 2004; Horne, 2009). Hence, this setting, where prices are the only signals for sustainability, is one possible interpretation of our model. This idea is consistent with the observation that goods marketed as sustainably produced often involve a price premium (Nimon and Beghin, 1999; Mahenc, 2008). An alternative interpretation is that prices in our model are a simplified way of modelling both a label and the price itself. In that case, in addition to acting as a price itself, a relatively high price can also be understood as a label for a relatively high sustainability level. Conversely, a relatively low price can also be perceived as labelling a relatively low sustainability level.

Depending on the consumer's beliefs and other specifications in the game, prices can act as signals because of the following aspect of our model. We assume an outside authority who is either highly reliable or unreliable in monitoring the manufacturer's choices. Only the manufacturer is able to observe the level of reliability, and so she has an informational advantage over the consumer. Specifically, if reliability is high, a manufacturer who chooses a high price, p_H , but low sustainability, s_L , receives a large punishment, which essentially renders the manufacturer's payoff $-\infty$. If reliability is low, there is an incentive for the manufacturer to exploit her information advantage and opt for low sustainability and a high price without having to fear the risk of a penalty. Intentional and misleading signals of sustainable production like this represent one form of greenwashing (Seele and Gatti, 2017). Following this definition, we will refer to the term greenwashing in the rest of this text as fraudulently signalling high sustainability with a high price, while in reality choosing low sustainability. Regardless of the reliability, a manufacturer can choose a low price and high sustainability or a high price and high sustainability without any punishment. Whatever the incentives facing the manufacturer, the consumer can condition her buying decision on the price of the good and any inferences she draws from the price.

We define the following utility function for the consumer, where b again acts as a dummy

variable determining whether the consumer receives the utility for buying the product or for not buying the product. The “0” superscripts on u indicate that the consumer does not buy. If the consumer does buy, we use the “1” superscripts on u .

$$\begin{aligned} u_c(b | s, p) &= b[u_c^1] + (1 - b)[u_c^0] \\ u_c(b | s, p) &= b[\pi - p + \beta s] + (1 - b)[u_c^0]. \end{aligned} \tag{2.2}$$

The parameter π represents the consumer’s maximum willingness to pay for a conventionally produced product with $s = 0$. For the moment, we do not further specify the outside utility, u_c^0 , of the consumer. This outside utility can take any value, and we will discuss how different values can lead to different equilibria. We will introduce a more specific definition of the consumer’s outside option at a later point.

The utilities shown in equations [2.1](#) and [2.2](#) represent utilities that occur when the consumer chooses to buy or not to buy the product. Note, however, that the two types of agents base their decisions in the game on expected utilities. Those expected utilities are determined by the agents’ beliefs. For instance, the consumer has a prior belief that the reliability of the monitoring authority is low. By extension, the consumer has a prior belief that the reliability is high. More generally, the consumer has prior beliefs about every move in the game, which is shown in [Figure 2.1](#). She uses Bayesian updating to form posterior beliefs about the sustainability of production after observing the producer’s chosen price (section 3 in the supplementary materials). The consumer buys if her expected utility from buying weakly exceeds the value of her outside option.

$$\mathbb{E}[u_c^1(b = 1 | p)] \geq u_c^0. \tag{2.3}$$

In other words, even though the consumer does not observe the sustainability level chosen by the manufacturer during the game and has to base her buying decision on expected utilities,

she will learn the true sustainability level after choosing to buy the product. As we analyse a one-shot game, this aspect of our model can be understood in the following way. The end of the game represents the long term result of the trade interaction, at which point the consumer has gained enough information to know the manufacturer’s production process. In the discussion section we suggest an alternative modelling approach in a repeated setting.

2.3.1 Game sequence

The game unfolds according to the following sequence, which is also shown in Figure [2.1](#). First, Nature draws a level of reliability, $R = \{r_L, r_H\}$. Defining this move in the game as Nature’s draw is a way of modelling incomplete information as imperfect information. In other words, the fact that the consumer does not observe the specific level of monitoring reliability is modelled as the consumer not observing a given move in the game. We refer to low reliability with “ r_L ” and we refer to high reliability with “ r_H ”. Specifically, r_L is drawn with probability g and r_H is drawn with probability $(1 - g)$. The manufacturer observes Nature’s choice and then chooses a sustainability level, $s \geq 0$. The manufacturer then chooses a price, $p > 0$. The action space of the manufacturer is $A_m = \{(s_L, p_L), (s_L, p_H), (s_H, p_L), (s_H, p_H)\}$. Finally, the consumer observes the price, but not the sustainability level, and then chooses to buy if her expected utility from buying weakly exceeds the value of her outside option, u_c^0 .

Figure [2.1](#) shows our buyer-seller exchange game of incomplete information, which is modelled as a game of imperfect information in which Nature chooses the reliability of an outside monitoring authority.

2.3.2 Game analysis

In our model, the manufacturer and consumer can have different preferences for sustainability. A case of particular interest is when the manufacturer does not care about sustainability, but the consumer does. We compare this scenario to benchmark cases where preferences align. Specifically, two benchmark cases are relevant (see sections 7.1 and 7.2 in the online

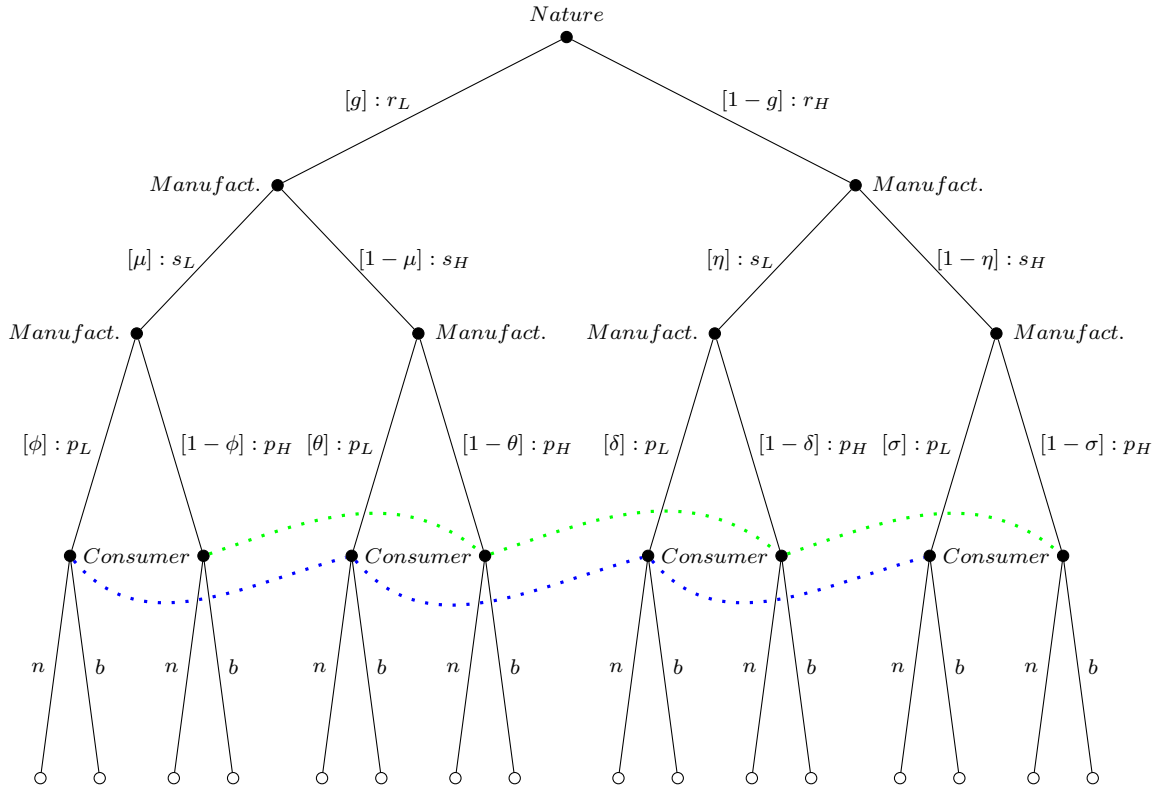


Figure 2.1: The game tree. The quantities g , μ , η , ϕ , θ , δ , and σ all specify subjective priors for the consumer along different paths through the game tree. The green dashed line connects four nodes that form one information set, and the blue dashed line connects four nodes that form the other information set.

supplementary materials). In one case, both the manufacturer and the consumer care little about sustainability. In this case, an equilibrium can exist in which the manufacturer sells cheap products with low sustainability to the consumer who is happy to buy (see section 7.1.3 in the online supplementary materials). In the other case, both the manufacturer and the consumer have sufficiently strong preferences for sustainability. In that equilibrium, the manufacturer sells highly sustainable products at high prices under both levels of reliability to a consumer who is happy to buy (see section 7.2.4 in the online supplementary materials).

What happens, however, when the manufacturer cares little about sustainability, but the consumer cares a lot? In this situation (section 7.1 in the online supplementary materials), two

perfect Bayesian pooling equilibria exist in which the manufacturer chooses a high price under both levels of reliability. Specifically, the manufacturer subject to a monitoring authority with low reliability chooses low sustainability and a high price, and the manufacturer subject to a highly reliable monitoring authority chooses high sustainability and a high price. These equilibria exist if the consumer's expected utility for purchasing the product after observing a high price is at least as large as the utility from her outside option. The two equilibria of this sort are slightly different (section 7.1 in the online supplementary materials). In one, the expected utility for buying after observing a high price is at least as large as the outside option for the consumer, but the expected utility for buying after observing a low price is strictly smaller. In the second equilibrium, the outside option is worse than the expected utilities for buying after observing any price.

Crucially, however, in both cases the consumer's tolerance of getting an expensive product with low sustainability increases as her preference for sustainability increases. In other words, strong pro-environmental values of the consumer sustain the value of greenwashing for the manufacturer. This result seems quite surprising at first glance. It is less surprising when looking at the equilibrium analysis more closely. Doing so clarifies how a manufacturer who is monitored with low reliability, when facing a consumer with strong sustainability preferences, can exploit information asymmetries to green wash.

First, we derive the consumer's expected utility for buying the product after observing a high price because this is the price choice in the two mentioned perfect Bayesian pooling equilibria. The consumer's expected utility is based on her prior beliefs about the moves in the game. The consumer's belief that the manufacturer has chosen low sustainability, after observing a high price, can be represented by the following conditional probability,

$$P(s_L | p_H) = \frac{[g\mu(1 - \phi)] + [(1 - g)\eta(1 - \delta)]}{[g\mu(1 - \phi)] + [g(1 - \mu)(1 - \theta)] + [(1 - g)\eta(1 - \delta)] + [(1 - g)(1 - \eta)(1 - \sigma)]}, \quad (2.4)$$

where g , for example, is the consumer's belief that the monitoring reliability of the manufacturer is low. In Figure [2.1](#) we show which probabilities correspond to which moves in the game. The beliefs about the monitoring reliability, g and $(1 - g)$, are what we will focus on in the following in addition to the pro-environmental values of the two types of agents. We assume that g and $(1 - g)$ are exogenous probabilities, corresponding to the true underlying distribution of low versus high monitoring reliability in the market. In other words, even though the consumer does not know for sure whether the manufacturer she is dealing with is monitored with low or high reliability, we assume for the moment that she forms accurate beliefs about the corresponding probabilities.

The consumer is aware of the sustainability preference of the manufacturer, α , and the utilities that the manufacturer receives for the different possible choice combinations in the game. The consumer also knows that the manufacturer is aware of the consumer's sustainability preference, β . This allows the consumer to infer the manufacturer's preference ordering under low and high monitoring reliability (section 6 in the supplementary materials). For instance, the consumer knows whether a manufacturer prefers choosing high or low sustainability under high or low monitoring reliability, conditional on the value of the manufacturer's sustainability preference. The assumption that the two types of agents in the game know each other's preferences is a simplification typically made in game theoretical models. A possible interpretation of such a setting is that the consumer has information about the reputation of the manufacturer with respect to pro-environmental values. Similarly, the manufacturer can conduct market research to find out about the consumer's preferences.

To derive the expected utility of the consumer we also need the consumer's belief that the manufacturer has chosen high sustainability, after observing a high price. This is simply the complementary probability to the one shown in [2.4](#), which is $P(s_H | p_H) = [1 - P(s_L | p_H)]$. If we substitute the consumer's beliefs about the manufacturer's preference orderings into the conditional probabilities, $P(s_L | p_H)$ and $P(s_H | p_H)$, we derive the following expected utility of the consumer for choosing to buy the product after observing a high price,

$$\begin{aligned}
\mathbb{E}[u_c^1(b = 1 \mid p_H)] &= \mathbb{E}[\pi - p_H + \beta\{P(s_L \mid p_H)s_L + P(s_H \mid p_H)s_H\}] \\
&= \pi - p_H + \beta\{gs_L + (1 - g)s_H\}.
\end{aligned} \tag{2.5}$$

To illustrate why the consumer's preference for sustainability supports greenwashing, consider the condition that must be satisfied for the consumer to buy the product.

$$\begin{aligned}
\mathbb{E}[u_c^1(b = 1 \mid p_H)] &\geq u_c^0 \\
\pi - p_H + \beta\{gs_L + (1 - g)s_H\} &\geq u_c^0.
\end{aligned} \tag{2.6}$$

For the moment, we think of the outside option as a conventionally produced substitute, i.e. $s = 0$, for the product in the focal trade interaction. For example, the manufacturer in the focal trade interaction could be a potential pioneer in terms of sustainability, and thus sustainability ceases to be an issue if the consumer abandons the current trade opportunity. As long as sustainability preferences captured by β are not relevant when the consumer reverts to the outside option, the focal manufacturer has the opportunity to invest a small amount into sustainability, i.e. $s_L > 0$, to ensure that increasing β will cause the consumer to eventually buy from her all else equal.

Apart from β , an interesting parameter to interpret is g . Given $\beta > 0$, for large g , a consumer with accurate beliefs matching this objective probability thinks that monitoring of the manufacturer is unreliable, and if g is sufficiently large she may not buy, depending on the magnitudes of $\pi - p_H$ and βs_L . In contrast, if g is low and the consumer derives her belief from this objective probability, then the likelihood that condition (2.6) is satisfied increases all else equal. With accurate beliefs, the consumer can make informed decisions because she is aware of the true distribution of low versus high monitoring reliability in the market, represented by the probabilities g and $(1 - g)$ respectively. To summarise, if most monitoring systems are reliable and the consumer knows this, the rate of greenwashing will also be low.

However, even if most monitoring systems are reliable, the first few manufacturers with low sustainability preferences that realise that they are being monitored with low reliability can grab the opportunity to profit with high prices that fraudulently signal sustainable production.

Crucially, however, the consumer's beliefs do not necessarily have to match the objective probabilities g and $(1 - g)$. In a market sector that has been flooded with labels signalling a similar or equal amount of sustainability, the consumer might no longer be able to evaluate which of them are reliably monitored. Instead, the consumer might base her decisions on subjective probabilities, let's call them \hat{g} and $(1 - \hat{g})$, that possibly deviate from the objective probabilities due to misinformation or the adoption of mistaken beliefs (Acerbi, 2019; Efferson, McKay and Fehr, 2020). We can see in equation 2.6 that a consumer is more likely to buy the product if she thinks that monitoring tends to be reliable. In other words, if the consumer underestimates the probability that monitoring is unreliable, i.e. $\hat{g} < g$, she mistakenly believes that green washing occurs less than it actually does. This result reveals that, in addition to the consumer's preferences for sustainability, prior beliefs about the reliability of an external monitoring authority are also important. If the combination of sustainability preferences and trust in this reliability are jointly adequate, the consumer pays a premium to buy in equilibrium, and greenwashing occurs in tandem.

More broadly, condition (2.6) in its entirety must hold for the consumer to buy. Accordingly, π , p , s_L , and s_H also matter. We simply focus on β and g because they are the quantities naturally associated with any cultural evolutionary process related to sustainability.

2.3.3 Informing consumers

The results of the previous section have shown that, if most monitoring systems are reliable and the consumer has accurate beliefs, greenwashing is less likely. Depending on the consumer's sustainability preferences, however, greenwashing might still occur. Specifically, the consumer cannot be certain about the true monitoring reliability of the manufacturer she is dealing with. Instead, she bases her decisions on the probabilities g and $(1 - g)$. The manu-

facturer on the other hand knows the monitoring reliability and can exploit a consumer with sufficiently strong sustainability preferences who is willing to accept the risk of greenwashing.

Consumer information and education are possible ways of reducing the problem of information asymmetry (Crespi and Marette, 2003b,a). We conducted an analysis where we imagine an extreme scenario, namely that a social planner knows the true reliability of the manufacturer's monitoring and successfully informs the consumer about this reliability (see section 7.3 in the online supplementary materials). In other words, the consumer observes Nature's move and thus knows with certainty whether she faces a manufacturer who is monitored with high or low reliability. This might constitute an unrealistic setting because a social planner does not always possess true information, and she is most likely not able to inform and educate consumers perfectly. However, we wanted to see whether removing information asymmetry completely, at least with respect to monitoring reliability, can prevent greenwashing.

Intuitively, we might expect that it does. Sometimes it does indeed, but not always. If the consumer knows that monitoring is reliable, and if the manufacturer cares little about sustainability, the consumer controls the outcome in all four equilibria that exist (see sections 7.3.1-7.3.4 in the online supplementary materials). If the consumer cares about sustainability, the manufacturer produces a highly sustainable product. If the consumer cares about money but not about sustainability, the manufacturer happily follows along with low sustainability. Thus, pro-environmental values of the consumer become more important when information asymmetry decreases. Perhaps the most surprising case arises when a consumer with strong sustainability preferences knowingly accepts a low-sustainability product for a high price.

$$\begin{aligned} \mathbb{E}[u_c^1(b = 1 \mid p_H, R = r_L)] &\geq u_c^0 \\ \pi - p_H + \beta s_L &\geq u_c^0. \end{aligned} \tag{2.7}$$

This equilibrium can occur if the product from the focal manufacturer is sufficiently valu-

able to the consumer relative to her outside option, u_c^0 (see section 7.3.4 in the online supplementary materials). Such an equilibrium only exists when a manufacturer under low reliability is relatively powerful. Assuming there is indeed a monopolist with large market power, our analysis suggests that the only way to achieve sustainable production is to increase that manufacturer's pro-environmental values. What happens, however, if there is competition? The situation might change completely because the consumer's outside option potentially improves, and we will discuss this next.

2.3.4 Competition

Here we extend the model by defining the outside option as a second manufacturer. Now, the consumer not only has the option to buy a certain type of product from the focal manufacturer, but she also has the option to leave the focal trade interaction and buy a substitute product from another manufacturer. Let us call the focal manufacturer $M1$ and the outside manufacturer $M0$. We do not conduct a full equilibrium analysis for the setting with competition. Instead, we conduct comparative statics exercises for some interesting scenarios. Importantly, we assume exogenous values for the choices of $M0$ and focus on the decision-making of $M1$. For now, assume that both manufacturers care little about sustainability, i.e. $\alpha < 1$ and that both manufacturers choose the same high price, p_H , and the same values for s_L and s_H . The consumer chooses to buy the product from $M1$ if the following condition is satisfied,

$$\begin{aligned} \mathbb{E}[u_c^1(b = 1 | p_H)] &\geq u_c^0 \\ \pi - p_H + \beta\{gs_L + (1 - g)s_H\} &\geq \pi - c - p_H + \beta\{gs_L + (1 - g)s_H\}. \end{aligned} \tag{2.8}$$

The parameter $c > 0$ is a cost. Such a cost arises if the consumer decides to abandon the focal trade interaction and searches for a new manufacturer. Everything else equal, an increase in c makes it more likely that the consumer is willing to buy from $M1$. We discuss

effects of different s_L and s_H values in the online supplementary materials (see section 7.4).

First, we assume again that the consumer has the accurate beliefs, g and $(1 - g)$, about monitoring reliability. An interesting scenario to look at is one where the outside manufacturer has strong pro-environmental values, $\alpha > 1$, and the focal manufacturer cares very little about sustainability, $0 < \alpha < 1$. With a sufficiently strong sustainability preference, the outside manufacturer, $M0$, will offer high sustainability, regardless of monitoring reliability and irrespective of whether the consumer buys from that manufacturer or not (Supplementary Information). To simplify, we assume $M0$ chooses p_H . The consumer has corresponding beliefs, and the outside option simplifies to $u_c^0 = \pi - c - p_H + \beta s_H$. The focal manufacturer, $M1$, is aware of this. We assume a consumer with a sufficiently high sustainability preference who is not willing to buy from $M1$ after observing p_L , because this would imply the utility $u_c^1 = \pi - p_L + \beta s_L$. After observing p_H , the consumer buys from $M1$ if g is sufficiently low or c sufficiently high (Supplementary Information). The question is whether $M1$ prefers pooling over a high price and selling the product, over pooling over a low price and not selling but having lower costs associated with low sustainability. For certain values of s_H , the price, and α , $M1$ is willing to pool over a high price (Supplementary Information). In this case, strong pro-environmental preferences of a competing manufacturer, $M0$, can induce a manufacturer with low sustainability preferences, who is monitored under high reliability, to produce sustainably (Supplementary Information).

Now imagine a scenario where both manufacturers care little about sustainability, both have already chosen the same high price, but they can spread misinformation. In other words, the consumer forms subjective probabilities about the monitoring reliability for the two manufacturers. We indicate the consumer's belief for low reliability associated with $M1$ as g^1 . Similarly, we indicate the belief for low reliability associated with $M0$ as g^0 . Importantly, the consumer's beliefs about the monitoring reliability may or may not differ between the two competing manufacturers, and they might deviate from the objective probability g . We derive the following comparison of expected utilities for the consumer,

$$\begin{aligned} \mathbb{E}[u_c^1(b = 1 | p_H)] &\geq u_c^0 \\ \pi - p_H + \beta\{g^1 s_L + (1 - g^1) s_H\} &\geq \pi - c - p_H + \beta\{g^0 s_L + (1 - g^0) s_H\}. \end{aligned} \tag{2.9}$$

We observe in equation (2.8) that increasing g^1 , all else equal, decreases the consumer's willingness to buy from $M1$. To illustrate that more clearly, assume the consumer believes with certainty that the second manufacturer, $M0$, is monitored with high reliability, i.e., $g^0 = 0$, and the focal manufacturer, $M1$, with low reliability, i.e., $g^1 = 1$. If that consumer's β is sufficiently high, this implies that she would be willing to buy from $M0$ because she can be certain to receive a product with high sustainability from that second manufacturer.

If, on the other hand, $M1$ succeeds in spreading misinformation that leads the consumer to believe that g^0 is very high but g^1 very low, the consumer would be more willing to buy from $M1$ than $M0$, everything else equal. Hence, mistaken beliefs can induce a consumer to underestimate or overestimate the likelihood that certain manufacturers green wash. Specifically, manufacturers monitored under low reliability can spread misinformation and exploit the consumer's pro-environmental values to increase their scope for greenwashing.

2.4 Discussion

Altogether, reliable monitoring can have two countervailing effects. On the one hand, if reliable monitoring is increasingly common, and if the prior beliefs of consumers reflect this, then g decreases. Declines in g increase the scope for an equilibrium in which environmentally-minded consumers tolerate the risk of greenwashing. In this sense, reliable monitoring compromises the ability of green consumers to exert pressure on the manufacturer. On the other hand, in a greenwashing equilibrium, green consumers believe and accept that greenwashing occurs with probability g . If this belief is accurate, say because it rests on accurate reporting about greenwashing among firms, or because the consumer learns this information from

other well-informed consumers, then the actual rate of fraudulent signalling in equilibrium is also g . In this sense, increasingly reliable monitoring decreases g and supports the ability of green consumers to exert pressure on the manufacturer to produce sustainably. Importantly, our analysis reveals that the worst outcome of all is when consumers overestimate the reliability of monitoring of manufacturers. Such an outcome is likely to support an equilibrium with greenwashing, but the actual rate of green washing would be much higher than environmentally-conscious consumers realise. This is an especially interesting possibility as manufacturers start exploiting unreliable monitoring. Do changes in consumer beliefs reflect this appropriately? Do consumers underestimate reliability, overestimate reliability, or get it just right? As unreliable monitoring proliferates, consumers may decide that all monitoring is untrustworthy. Alternatively, consumers may simply hope for the best. Crucially, however, a social planner could intervene and support the dissemination of accurate beliefs among consumers, i.e. beliefs that better reflect the true reliability of monitoring systems.

In addition to beliefs, pro-environmental values can be critical too. Most trivially, our model shows that sustainable production will follow regardless of the presence of competition or reliable monitoring if both types of agents, the manufacturer and the consumer, care about sustainability. If the manufacturer does not care about sustainability, our analysis shows that the consumer's green preference becomes more important as competition increases, as long as her belief about monitoring reliability matches objective probabilities. If the consumer adopts mistaken beliefs, however, greenwashing can even occur in the presence of competition.

Crucially, we have considered the possibility of differences in environmental preferences between manufacturer and consumer and within manufacturers, but we have ignored the possibility of heterogeneity within consumers. A manufacturer with unsustainably produced products might be able to find consumers with low sustainability preferences and a sustainably producing manufacturer might find consumers with strong pro-environmental preferences. Thus, our finding that a second manufacturer can incentivise another manufacturer to produce sustainably might not be a relevant finding in a world where no majority of green consumers

yet exists. Importantly, however, we wanted to focus on a scenario where we imagine what happens in terms of sustainable outcomes when there are, in fact, either a majority or a minority of green consumers. We did so because it has been argued recently that a large-scale behavioural change will be necessary to achieve a transition to a sustainable future. In other words, collapsing demand into one consumer allowed us to ask what happens when this aggregate demand is in fact green. Our results show that when a majority of people have strong pro-environmental preferences, sustainable outcomes do indeed follow under certain circumstances. Nonetheless, we typically expect individual heterogeneity to have important effects on social tipping dynamics (Constantino et al., 2022; Efferson et al., 2023). Hence, evolutionary models that take into account the heterogeneity of both manufacturer and consumer preferences could contribute to a better understanding of tipping in consumption and production behaviour.

The importance of a consumer's pro-environmental values decreases if her outside option is sufficiently weak, or equivalently if the manufacturer enjoys a kind of short-side power (Bowles, 2009). In that case, our analysis suggests that the cultural diffusion of pro-environmental values among consumers does not reliably support sustainable production. Any effort to amplify sustainability preferences among consumers further would only exacerbate greenwashing in such a situation. If a manufacturer with significant market power has weak preferences for sustainability, asymmetric information implies that preferences for sustainability among consumers support greenwashing in equilibrium. Crucially, the existence of reliable monitoring does not prevent greenwashing. Indeed, reliable monitoring creates the opportunity for manufacturers operating under unreliable monitoring to free-ride on the reputations of their trustworthy counterparts. Under those circumstances, the ability of the consumer to exert pressure on manufacturers is limited, and targeting the cultural evolution of sustainability preferences among manufacturers could be an effective way to promote sustainable outcomes. Imagine business schools that emphasise environmental responsibility just as much as profits and firm value (Collier, 2018; Mayer, 2018).

Of the many ways of addressing unsustainable production and consumption, we necessarily neglected a number of potentially important mechanisms. For example, a mechanism that we have not modelled is the manufacturer's choice of certification. A manufacturer might be incentivised to choose a certain certification due to the reliability of monitoring, the costs associated with the system of monitoring or the credibility towards the consumer (Golan et al., 2001; Cason and Gangadharan, 2002). We have ignored the choice of a monitoring system because we wanted to focus on the simple possibility that the consumer has less information about monitoring reliability than the manufacturer. In addition, we have simplified our model by using prices as signals. In other words, a manufacturer does not have costs directly linked to choosing a given label, but she faces costs if she chooses a high price and low sustainability when monitoring reliability is high. Choosing certain certification schemes can be important for the outcomes of production (Garrido et al., 2020), and future models could combine the signalling aspect of prices implemented in our model with such a choice between different certifications.

Further, we have only discussed manufacturer competition in a very simplified setting with strong assumptions. In reality, manufacturers can compete via prices and not only by choosing different levels of sustainability. Results are likely to change if the manufacturers can compete via prices. In fact, with price competition, the focal manufacturer becomes less likely to offer high sustainability in the presence of a competing manufacturer (see section 7.4 in the supplementary materials). On the other hand, this might again change if there are multiple consumers. This might allow manufacturers with different sustainability preferences to target consumers who also differ with respect to pro-environmental values. These are interesting questions for future research.

Finally, we have also ignored mechanisms in our model that are likely important in a repeated setting. For example, consumers might not learn about the true underlying sustainability level after buying a product. Instead, we could imagine a dynamic setting where greenwashing of a manufacturer might only be uncovered by consumers with a certain prob-

ability every period. This mechanism could also introduce a feedback loop between the consumer's sustainability preference and finding out about greenwashing. Buying products with sustainable labels and later finding out that the product was not sustainably produced might cause consumers to trust less in labels and lead to an overall lower willingness to pay for such labels. Such and other mechanisms could for example be studied in behavioural experiments, where multiple consumers have the option to choose from multiple manufacturers.

From a cultural evolutionary perspective, another interesting question that could be examined in a behavioural experiment would be to see the effects of consumers being able to observe the buying decisions of other consumers. An important notion in the cultural evolution literature is that individuals are more likely to interact and socially learn from similar others (Efferson et al., 2008; Kendal et al., 2018). In a behavioural experiment, one could study whether consumers are more likely to buy a given product after observing that someone with similar sustainability preferences bought the product. Importantly, one could examine whether such a social learning effect is stronger than other factors such as beliefs about the probabilities of monitoring reliability.

In broad terms, the upshot of our model for improving sustainability via pro-environmental values is the following. Preferences for sustainability can differ between manufacturer and consumer. When they do, increasingly green preferences among consumers open the door ever more widely to greenwashing as a likely outcome. Surprisingly, this is even true in extreme cases when the consumer knows with certainty whether or not monitoring is reliable. Should a social planner promote the cultural evolution of preferences for sustainability among manufacturers, consumers, or both? Our analysis suggests that the information structure, competition among manufacturers, the actual distribution of reliable and unreliable monitoring, and consumer beliefs about reliability, might all be important. These factors determine whether the cultural evolution of sustainability preferences actually translates into improved sustainability.

The cultural evolution of pro-environmental values and beliefs involves complex interre-

relationships that we are only now beginning to uncover. We have highlighted some of the challenges posed by the spread of false beliefs and the mechanisms that prevent large-scale behaviour change based on pro-environmental norms. Our simple model illustrates what those risks imply for the sustainability of production and consumption in certain contexts. For a rapid transition to a sustainable future, it is crucial that we accelerate our research on possible ways to address these challenges.

Acknowledgements and funding

The authors acknowledge support from the Swiss National Science Foundation (SNSF Project no. 100018_185417/1).

Author contributions

L.v.F.: conceptualization, formal analysis, methodology, validation, visualization, writing: original draft, review and editing; C.E.: conceptualization, formal analysis, funding acquisition, project administration, supervision, writing: original draft, review and editing; S.V.: conceptualization, funding acquisition, project administration, supervision.

Research transparency and reproducibility

The supplementary material is available on: <https://doi.org/10.6084/m9.figshare.c.6858568.v2>.

Bibliography

- Acerbi, A. (2019). Cognitive attraction and online misinformation, *Palgrave Communications* **5**(1): 1–7.
- Aghion, P., Bénabou, R., Martin, R. and Roulet, A. (2020). Environmental preferences and technological choices: Is market competition clean or dirty?, *Technical report*, National Bureau of Economic Research.
- Andreoni, J., Nikiforakis, N. and Siegenthaler, S. (2021). Predicting social tipping and norm change in controlled experiments, *Proceedings of the National Academy of Sciences* **118**(16).
- Banerjee, A., Alsan, M., Breza, E., Chandrasekhar, A. G., Chowdhury, A., Duflo, E., Goldsmith-Pinkham, P. and Olken, B. A. (2020). Messages on covid-19 prevention in

- india increased symptoms reporting and adherence to preventive behaviors among 25 million recipients with similar effects on non-recipient members of their communities, *Technical report*, National Bureau of Economic Research.
- Becker, J., Brackbill, D. and Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds, *Proceedings of the National Academy of Sciences* **114**(26): E5070–E5076.
- Bénabou, R. and Tirole, J. (2010). Individual and corporate social responsibility, *Economica* **77**(305): 1–19.
- Bentley, R. A., Maddison, E. J., Ranner, P. H., Bissell, J., Caiado, C., Bhatanacharoen, P., Clark, T., Botha, M., Akinbami, F. and Hollow, M. (2014). Social tipping points and Earth systems dynamics, *Frontiers in Environmental Science* **2**: 35.
- Berger, J., Efferson, C. and Vogt, S. (2023). Tipping pro-environmental norm diffusion at scale: opportunities and limitations, *Behavioural Public Policy* **7**(3): 581–606.
- Bowen, F. (2014). *After Greenwashing: Symbolic Corporate Environmentalism and Society*, Cambridge University Press.
- Bowles, S. (2009). *Microeconomics: Behavior, Institutions, and Evolution*, Princeton University Press.
- Brécard, D., Hlaimi, B., Lucas, S., Perraudau, Y. and Salladarré, F. (2009). Determinants of demand for green products: An application to eco-label demand for fish in europe, *Ecological Economics* **69**(1): 115–125.
- Brooks, J. S., Waring, T. M., Mulder, M. B. and Richerson, P. J. (2018). Applying cultural evolution to sustainability challenges: An introduction to the special issue, *Sustainability Science* **13**(1): 1–8.
- Cason, T. N. and Gangadharan, L. (2002). Environmental labeling and incomplete consumer

- information in laboratory markets, *Journal of Environmental Economics and Management* **43**(1): 113–134.
- Collier, P. (2018). *The Future of Capitalism: Facing the New Anxieties*, Penguin UK.
- Constantino, S. M., Sparkman, G., Kraft-Todd, G. T., Bicchieri, C., Centola, D., Shell-Duncan, B., Vogt, S. and Weber, E. U. (2022). Scaling up change: A critical review and practical guide to harnessing social norms for climate action, *Psychological Science in the Public Interest* **23**(2): 50–97.
- Crespi, J. M. and Marette, S. (2003a). “Does Contain” vs. “Does Not Contain”: Does it matter which gmo label is used?, *European Journal of Law and Economics* **16**: 327–344.
- Crespi, J. M. and Marette, S. (2003b). Some economic implications of public labeling, *Journal of Food Distribution Research* **34**(856-2016-57144): 83–94.
- Crespi, J. M. and Marette, S. (2005). Eco-labelling economics: Is public involvement necessary, *Environment, Information and Consumer Behavior* pp. 93–110.
- Darby, M. R. and Karni, E. (1973). Free competition and the optimal amount of fraud, *The Journal of Law and Economics* **16**(1): 67–88.
- de Freitas Netto, S. V., Sobral, M. F. F., Ribeiro, A. R. B. and Soares, G. R. d. L. (2020). Concepts and forms of greenwashing: A systematic review, *Environmental Sciences Europe* **32**(1): 1–12.
- Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E. and Quattrocioni, W. (2016). The spreading of misinformation online, *Proceedings of the National Academy of Sciences* **113**(3): 554–559.
- DellaVigna, S. and La Ferrara, E. (2015). Economic and social impacts of the media, *Handbook of media economics*, Vol. 1, Elsevier, pp. 723–768.

- Efferson, C., Lalive, R. and Fehr, E. (2008). The coevolution of cultural groups and ingroup favoritism, *Science* **321**(5897): 1844–1849.
- Efferson, C., McKay, R. and Fehr, E. (2020). The evolution of distorted beliefs vs. mistaken choices under asymmetric error costs, *Evolutionary Human Sciences* **2**.
- Efferson, C., Vogt, S. and Fehr, E. (2020). The promise and the peril of using social influence to reverse harmful traditions, *Nature Human Behaviour* **4**(1): 55–68.
- Efferson, C., Vogt, S. and von Flüe, L. (2023). Activating cultural evolution for good when people differ from each other, in J. Kendal, R. Kendal and J. Tehrani (eds), *Oxford Handbook of Cultural Evolution*, Oxford University Press, chapter TBD.
- Ehret, S., Constantino, S. M., Weber, E. U., Efferson, C. and Vogt, S. (2022). Group identities can undermine social tipping after intervention, *Nature Human Behaviour* pp. 1–11.
- Farmer, J. D., Hepburn, C., Ives, M. C., Hale, T., Wetzer, T., Mealy, P., Rafaty, R., Srivastav, S. and Way, R. (2019). Sensitive intervention points in the post-carbon transition, *Science* **364**(6436): 132–134.
- Garrido, D., Espínola-Arredondo, A. and Muñoz-García, F. (2020). Can mandatory certification promote greenwashing? a signaling approach, *Journal of Public Economic Theory* **22**(6): 1801–1851.
- Gatti, L., Seele, P. and Rademacher, L. (2019). Grey zone in–greenwash out. a review of greenwashing research and implications for the voluntary–mandatory transition of csr, *International Journal of Corporate Social Responsibility* **4**(1): 1–15.
- Giovannucci, D. and Ponte, S. (2005). Standards as a new form of social contract? sustainability initiatives in the coffee industry, *Food Policy* **30**(3): 284–301.
- Giuliani, E., Ciravegna, L., Vezzulli, A. and Kilian, B. (2017). Decoupling standards from

- practice: The impact of in-house certifications on coffee farms' environmental and social conduct, *World Development* **96**: 294–314.
- Gladwell, M. (2000). *The Tipping Point: How little things can make a big difference*, Little, Brown and Company.
- Golan, E., Kuchler, F., Mitchell, L., Greene, C. and Jessup, A. (2001). Economics of food labeling, *Journal of Consumer Policy* **24**(2): 117–184.
- Golub, B. and Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds, *American Economic Journal: Microeconomics* **2**(1): 112–49.
- Horne, R. E. (2009). Limits to labels: The role of eco-labels in the assessment of product sustainability and routes to sustainable consumption, *International Journal of Consumer Studies* **33**(2): 175–182.
- Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M. and Jones, P. L. (2018). Social learning strategies: Bridge-building between fields, *Trends in Cognitive Sciences* **22**(7): 651–665.
- Kirchoff, S. (2000). Green business and blue angels: A model of voluntary overcompliance with asymmetric information, *Environmental and Resource Economics* **15**(4): 403–20.
- Kirmani, A. and Rao, A. R. (2000). No pain, no gain: A critical review of the literature on signaling unobservable product quality, *Journal of Marketing* **64**(2): 66–79.
- Kline, M. A., Waring, T. M. and Salerno, J. (2018). Designing cultural multilevel selection research for sustainability science, *Sustainability science* **13**(1): 9–19.
- La Ferrara, E. (2016). Mass media and social change: Can we use television to fight poverty?, *Journal of the European Economic Association* **14**(4): 791–827.
- Laffont, J.-J. and Maskin, E. (1987). Monopoly with asymmetric information about quality: Behavior and regulation, *European Economic Review* **31**(1): 483–489.

- Lyon, T. P. and Maxwell, J. W. (2011). Greenwash: Corporate environmental disclosure under threat of audit, *Journal of Economics & Management Strategy* **20**(1): 3–41.
- Mahenc, P. (2008). Signaling the environmental performance of polluting products to green consumers, *International Journal of Industrial Organization* **26**(1): 59–68.
- Marquis, C., Toffel, M. W. and Zhou, Y. (2016). Scrutiny, norms, and selective disclosure: A global study of greenwashing, *Organization Science* **27**(2): 483–504.
- Mayer, C. (2018). *Prosperity: Better Business Makes the Greater Good*, Oxford University Press.
- Nilsson, H., Tunçer, B. and Thidell, Å. (2004). The use of eco-labeling like initiatives on food products to promote quality assurance—is there enough credibility?, *Journal of Cleaner Production* **12**(5): 517–526.
- Nimon, W. and Beghin, J. (1999). Are eco-labels valuable? evidence from the apparel industry, *American Journal of Agricultural Economics* **81**(4): 801–811.
- Nyborg, K., Anderies, J. M., Dannenberg, A., Lindahl, T., Schill, C., Schlüter, M., Adger, W. N., Arrow, K. J., Barrett, S. and Carpenter, S. (2016). Social norms as solutions, *Science* **354**(6308): 42–43.
- Otto, I. M., Donges, J. F., Cremades, R., Bhowmik, A., Hewitt, R. J., Lucht, W., Rockström, J., Allerberger, F., McCaffrey, M. and Doe, S. S. (2020). Social tipping dynamics for stabilizing Earth’s climate by 2050, *Proceedings of the National Academy of Sciences* **117**(5): 2354–2365.
- Rao, A. R. and Monroe, K. B. (1989). The effect of price, brand name, and store name on buyers’ perceptions of product quality: An integrative review, *Journal of Marketing Research* **26**(3): 351–357.

- Schimmelpfennig, R. and Muthukrishna, M. (2023). Cultural evolutionary behavioural science in public policy, *Behavioural Public Policy* pp. 1–31.
- Schimmelpfennig, R., Vogt, S., Ehret, S. and Efferson, C. (2021). Promotion of behavioural change for health in a heterogeneous population, *Bulletin of the World Health Organization* **tbd**(tbd): tbd.
- Seele, P. and Gatti, L. (2017). Greenwashing revisited: In search of a typology and accusation-based definition incorporating legitimacy strategies, *Business Strategy and the Environment* **26**(2): 239–252.
- Shiv, B., Carmon, Z. and Ariely, D. (2005). Placebo effects of marketing actions: Consumers may get what they pay for, *Journal of Marketing Research* **42**(4): 383–393.
- Smaldino, P. E., Janssen, M. A., Hillis, V. and Bednar, J. (2017). Adoption as a social marker: Innovation diffusion with outgroup aversion, *The Journal of Mathematical Sociology* **41**(1): 26–45.
- Smaldino, P. E. and Jones, J. H. (2021). Coupled dynamics of behaviour and disease contagion among antagonistic groups, *Evolutionary Human Sciences* **3**.
- Travers, H., Walsh, J., Vogt, S., Clements, T. and Milner-Gulland, E. J. (2021). Delivering behavioural change at scale: What conservation can learn from other fields, *Biological Conservation* **257**: 109092.
- Treen, K. M. d., Williams, H. T. and O’Neill, S. J. (2020). Online misinformation about climate change, *Wiley Interdisciplinary Reviews: Climate Change* **11**(5): e665.
- Valkila, J., Haaparanta, P. and Niemi, N. (2010). Empowering coffee traders? the coffee value chain from nicaraguan fair trade farmers to finnish consumers, *Journal of Business Ethics* **97**(2): 257–270.

- Walker, K. and Wan, F. (2012). The harm of symbolic actions and green-washing: Corporate actions and communications on environmental performance and their financial implications, *Journal of Business Ethics* **109**(2): 227–242.
- Waring, T. M., Goff, S. H. and Smaldino, P. E. (2017). The coevolution of economic institutions and sustainable consumption via cultural group selection, *Ecological Economics* **131**: 524–532.
- Wu, Y., Zhang, K. and Xie, J. (2020). Bad greenwashing, good greenwashing: Corporate social responsibility and information transparency, *Management Science* **66**(7): 3095–3112.
- Young, W., Hwang, K., McDonald, S. and Oates, C. J. (2010). Sustainable consumption: green consumer behaviour when purchasing products, *Sustainable Development* **18**(1): 20–31.

Chapter 3

Enhance threshold models to study
the effects of heterogeneity in learning
on social norm change

Abstract

Individuals vary greatly in how they respond to information about the behaviour of others, which can significantly influence the effects of policy interventions aimed at changing social norms. Despite its significance, our understanding of this diversity in social learning and its implications for evolutionary dynamics is still in its early stages. This paper implements an agent-based model to investigate social norm change, setting itself apart by allowing agents to access and utilise a variety of information types. This approach enhances typical norm change models which assume that agents solely react to choice distributions. The dynamics of social norm change are modelled with a coordination game in which agents choose between two options, the status quo and an alternative. While coordination on the alternative behaviour would enhance social welfare, agents remain entrenched in a status quo norm equilibrium, which is suboptimal for each individual. The study evaluates various interventions aimed at catalysing endogenous shifts towards the alternative norm. By expanding traditional models to include multiple learning mechanisms, a wider array of interventions can be assessed. The effectiveness of these interventions depends crucially on accurately targeting specific groups of agents and carefully selecting the social information highlighted. This research underscores that diversity in social learning adds complexity to the dynamics of norm change, suggesting that achieving rapid norm change by targeting a critical mass of influential individuals may not always be feasible.

Keywords: cultural evolution, social learning, heterogeneity in social learning, social norms, social norm change, social tipping, coordination game, agent-based model

3.1 Introduction

Cultural change is deeply influenced by the intricate ways individuals interact and learn from each other. Our understanding of this process often relies on models that assume everyone behaves as a conformist (Granovetter, 1978; Boyd and Richerson, 1985; Efferson et al., 2020; Andreoni et al., 2021). Conformist social learning involves people adopting common behaviours more frequently and shying away from less common ones. As certain behaviours gain popularity, their adoption rates accelerate, reinforcing the dominant trends. Conversely, behaviours that are rare often become even rarer or disappear completely.

Despite the widespread use of these models, they don't fully align with real-world observations, which show that people engage in many different forms of social learning (Mesoudi et al., 2016; Kendal et al., 2018). While models simplify assumptions to make analysis manageable, it's becoming clear that individual differences significantly influence how norms change and evolve (Young, 2009; Efferson et al., 2020; Andreoni et al., 2021; Newton, 2021; Efferson et al., 2023). This discrepancy highlights the need to develop models that better capture the nuanced ways people actually behave.

Crucially, the misalignment between empirical research and theoretical models of norm change carries implications that extend beyond academia, significantly affecting policy-making across various sectors (Efferson et al., 2023). In particular, there is growing interest in applying insights from cultural change research to encourage sustainable behaviours (Nyborg et al., 2016; Otto et al., 2020; Constantino et al., 2022). This body of research often views pro-environmental behaviour as one potential stable outcome in a binary coordination scenario. The concept suggests that targeted interventions might motivate a small yet pivotal segment of the population to adopt sustainable practices. In turn, this could influence others to do the same, leading to widespread adoption. This phenomenon, where a shift from a minority adopting a sustainable lifestyle to broad societal engagement occurs, is commonly referred to as social tipping.

While social tipping presents a promising strategy for policymakers aiming to promote widespread adoption of sustainable behaviours through targeted interventions on a small but critical subgroup, the effectiveness of these mechanisms is still uncertain. Our current understanding of the factors that facilitate or impede the adoption of sustainable norms remains limited (Efferson et al., 2023). The diversity in individual preferences and their responses to social influences introduces significant complexity. For example, the relationship between an individual’s preferences and their social group identity can greatly influence their likelihood of adopting new social norms (Ehret et al., 2022). Furthermore, recent models caution that even if norm change occurs, it may not benefit everyone in a society characterised by diverse preferences (Efferson et al., 2023, 2024).

This paper introduces a new modelling approach that embraces the diversity in social learning, moving beyond the assumption that all individuals conform. An agent-based model is implemented, centred on a coordination game that models decision making as a function of three different types of information that influence the decision between maintaining a status quo or adopting an alternative norm. By considering the impact of successful individuals and majority behaviours, alongside the personal benefits of each choice, a model is presented that more realistically simulates decision-making processes. A particular focus lies on how changes in social norms affect social welfare, balancing between the benefits these norms may bring and the inequalities they might introduce (Efferson et al., 2024).

Our findings highlight that effectively shifting societal norms requires targeted interventions that not only consider who is influenced but also strategically emphasise the types of information to which people are most responsive. Importantly, this model underscores the complexities involved in crafting effective policy interventions, particularly due to the varied ways individuals respond to different information types. Therefore, a deeper understanding of these dynamics is crucial before policymakers can confidently apply insights from the social tipping literature to their strategies.

3.2 Model

3.2.1 Game

Consider a population of N agents who must repeatedly choose between a status quo (SQ) and an alternative (Alt) over t_{max} time periods. In each period, agents are randomly paired to play a coordination game, making their choices without knowing their partner's preference in advance. Each agent has an idiosyncratic preference, denoted as x_i , which influences the incentives they face in the coordination game, as shown in Table [3.1](#).

	SQ	Alt
SQ	$a + x_i$	$b + x_i$
Alt	a	d

Table 3.1: Coordination game: Row player's payoffs are shown.

The x_i values are defined over the open interval $(0, d - b)$, where $b < d$, and are drawn from a left-skewed beta distribution (see supplementary materials). The game depicted in table [3.1](#) is a strict coordination game because, for each agent i , $a < a + x_i$ and $b + x_i < d$. Based on choices observed in period $t - 1$, agent i forms the belief, $\tilde{q}_{i,t}$, that her partner will play Alt in period t . Specifically, the belief $\tilde{q}_{i,t}$ corresponds to the actual fraction of Alt choices in period $t - 1$. An agent i 's belief $\tilde{q}_{i,t}$, together with her preference, x_i , determines the expected payoffs for choosing SQ and Alt . The expected payoff from choosing SQ , $E[\Pi_{i,t}(SQ)]$, and the expected payoff from choosing Alt , $E[\Pi_{i,t}(Alt)]$, are the following,

$$\begin{aligned}
 E[\Pi_{i,t}(SQ)] &= (1 - \tilde{q}_{i,t})(a + x_i) + \tilde{q}_{i,t}(b + x_i) \\
 E[\Pi_{i,t}(Alt)] &= (1 - \tilde{q}_{i,t})a + \tilde{q}_{i,t}d.
 \end{aligned}
 \tag{3.1}$$

It follows from Equation [3.1](#) that agent i is indifferent between the two choice options, SQ and Alt , if $\tilde{q}_{i,t} = x_i/(d - b)$. If agent i believes $\tilde{q}_{i,t} > x_i/(d - b)$, she prefers Alt , and vice versa.

This approach is often referred to as a “threshold model” (Granovetter, 1978; Efferson et al., 2020, 2024), where the beliefs, $\tilde{q}_{i,t}$, represent idiosyncratic preferences, commonly termed threshold values.

Importantly, in the model described thus far, agents respond exclusively to information regarding the frequency of choices. This behaviour can be interpreted in various ways. For instance, agents may be viewed as myopically best responding to their beliefs about their partners’ decisions (Mäs and Nax, 2016). Alternatively, they could be seen as strong conformists who disproportionately follow the majority (Boyd and Richerson, 1985; Efferson et al., 2008). If the agents are interpreted as myopically best responding, the x_i values represent idiosyncratic preferences that shape their expected payoffs in the coordination game. Conversely, if the agents are viewed as conformists, the x_i values can be considered content biases (Kendal et al., 2018), which differ among agents and determine the extent of their tendency to align with the majority.

If agents are interpreted as myopically best responding, specific inferences about the system’s dynamics can be made. For example, one can evaluate the relative costs of miscoordination and how these affect norm change. If agent i ’s partner selects SQ , coordinating on SQ results in a payoff of $a + x_i$, while choosing Alt yields a payoff of a . Therefore, the cost of not matching the partner’s SQ choice is $x_i = (a + x_i) - a$. Similarly, the cost of miscoordination when the partner opts for Alt is $d - (b + x_i)$. This creates relative costs associated with miscoordination. Specifically, if $x_i > (d - b)/2$, agent i is more inclined to prefer SQ over Alt . In technical terms, this indicates that, for such an individual, coordinating on SQ risk-dominates coordinating on Alt (Harsanyi and Selten, 1988). If $x_i < (d - b)/2$, the opposite is true. Given that the x_i values are drawn from a left-skewed beta distribution, the majority of agents perceive coordination on SQ as risk dominant.

However, these inferences about risk dominance are valid only in models where all agents respond exclusively to information about expected payoffs. Importantly, alternative forms of social learning, such as success-based learning exist (Kendal et al., 2018), but are overlooked

in typical threshold models. The model presented in this paper retains the interpretation of agents as maximising their expected payoffs in the coordination game but extends the framework by incorporating two additional types of information to which agents respond. More precisely, while agents in this model engage in the coordination game shown in Table 3.1 and have idiosyncratic preferences based on their x_i values, these heterogeneous preferences primarily influence the decisions of agents who respond to expected payoffs. For agents responding to the other two types of information, these preferences might not be as significant. This will be elaborated upon in the next section.

3.2.2 Learning strategies

Each agent observes the following three types of information in every period: 1) private information about the expected payoff, 2) majority information, and 3) success-based information. The information observed by each agent can be represented as an array, $\mathbf{I}_{i,t}$, consisting of three binary values. The process by which these values are determined is explained in detail for each of the three types of information.

1) Private information corresponds to what is shown in equation 3.1. This corresponds to the characterisation of agents' decision-making shown in Efferson et al. (2024), with the slight variation here that agents form their beliefs, $\tilde{q}_{i,t}$, based on a sample of $n \leq N$ choices of other agents observed in period $t - 1$. Determining beliefs based on the previous period is a simplified form of what is sometimes called the "recency effect" (Young, 2015). If agent i observes that her expected payoff for SQ is at least as high as for Alt , the first index of her information array will be defined as $\mathbf{I}_{i,t}[1] = 0$, and it will be defined as $\mathbf{I}_{i,t}[1] = 1$ if her expected payoff for Alt is higher.

2) Agents can exhibit a simple form of conformity bias (Boyd and Richerson, 1985). Specifically, if agent i observes that the majority of agents in her sample of n observations, i.e. a fraction > 0.5 , chose Alt in $t - 1$, then $\mathbf{I}_{i,t}[2] = 1$, and $\mathbf{I}_{i,t}[2] = 0$ otherwise.

3) An agent observes whether the most successful individual in her sample of n observations

chose SQ or Alt in period $t-1$. The “most successful” refers to the agent with the highest payoff in that period. Crucially, the assumption that agents can observe the payoffs or preferences of others is not necessary for the success-based learning modelled here. Each agent has idiosyncratic preferences affecting their payoff, $a + x_i$, from coordinating on SQ . If two agents coordinate on Alt , they receive a fixed payoff d . Importantly, it is assumed that $\forall i, (a + x_i) < d$. Consequently, if an agent i observes two agents coordinating on Alt , those two agents are necessarily perceived as the most successful agents, leading i to copy one of their behaviours. Furthermore, agents do not need to observe payoffs and preferences when responding to success-based information, even when observing two agents coordinating on SQ . Specifically, coordination on SQ yields a strictly higher payoff than miscoordination for each agent. Therefore, if all agents in agent i 's sample of n observations coordinate on SQ , it follows that they are the most successful. Consequently, if agent i is an individualist learner, they will simply replicate the behaviour of one of those agents coordinating on SQ .

Together, the three types of information yield 8 possible combinations of information shown in table [3.2](#)

Private	Majority	Success
$\mathbf{I}_{i,t}[1]$	$\mathbf{I}_{i,t}[2]$	$\mathbf{I}_{i,t}[3]$
0	0	0
0	0	1
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Table 3.2: Information combinations.

The strategies, $S_i(\mathbf{I}_i)$, are mappings, $\mathbf{I}_i \rightarrow S_i$, from observed information, \mathbf{I}_i , to choice probabilities. In the model presented here, the action space is simplified by allowing the agents to only have deterministic strategies, where the probability of choosing an action for each observed combination of information is either 0, meaning the agent never chooses that

action for a given information combination, or 1, meaning the agent always chooses that action if she observes a particular information combination. Hence, there are $2^8 = 256$ possible pure strategies (see equation (2) in the supplementary materials).

Not all of the possible strategies resemble well-studied learning strategies. Hence, instead of randomly sampling N strategies from the pool of 256 strategies, we initialise the system with five particular types of learners, each making up 1/5 of the population. Those five learner types are labelled as ‘‘Conservatives’’, ‘‘individualists’’, ‘‘conformists’’, ‘‘success-oriented’’, and ‘‘other’’ learners. Conservatives always choose SQ regardless of the observed information combination. Their strategy looks like the following:

$$\text{Conservative } S_i(\mathbf{I}_i) = \left\{ \begin{array}{l} SQ \mid \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 0 \\ SQ \mid \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 1 \\ SQ \mid \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 0 \\ SQ \mid \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 1 \\ SQ \mid \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 0 \\ SQ \mid \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 1 \\ SQ \mid \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 0 \\ SQ \mid \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 1 \end{array} \right\} \quad (3.2)$$

Individualists only pay attention to their private information about expected payoffs, $\mathbf{I}_i[1]$. The strategy of individualist learners is presented below in equation [3.3](#). As shown, anytime an individualist observes that the expected payoff for SQ is at least as high as for Alt , they choose SQ , and Alt otherwise.

$$\text{Individualist } S_i(\mathbf{I}_i) = \left\{ \begin{array}{l} SQ \mid \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 0 \\ SQ \mid \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 1 \\ SQ \mid \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 0 \\ SQ \mid \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 1 \\ Alt \mid \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 0 \\ Alt \mid \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 1 \\ Alt \mid \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 0 \\ Alt \mid \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 1 \end{array} \right\} \quad (3.3)$$

Conformists mimic the majority's choice and solely focus on that social information, $\mathbf{I}_i[2]$, ignoring any other type of information. Their strategy looks as follows.

$$\text{Conformist } S_i(\mathbf{I}_i) = \begin{pmatrix} SQ & | & \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 0 \\ SQ & | & \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 1 \\ Alt & | & \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 0 \\ Alt & | & \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 1 \\ SQ & | & \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 0 \\ SQ & | & \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 1 \\ Alt & | & \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 0 \\ Alt & | & \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 1 \end{pmatrix} \quad (3.4)$$

Success-oriented learners copy the peers receiving the highest payoff from playing the coordination game by only considering $\mathbf{I}_i[3]$. This form of learning is typically described as success bias or payoff bias in the cultural evolution literature (Kendal et al., 2018). The strategy of agents exhibiting this learning bias is shown in equation 3.5 below.

$$\text{Success-based } S_i(\mathbf{I}_i) = \begin{pmatrix} SQ & | & \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 0 \\ Alt & | & \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 1 \\ SQ & | & \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 0 \\ Alt & | & \mathbf{I}_i[1] = 0, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 1 \\ SQ & | & \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 0 \\ Alt & | & \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 0, \mathbf{I}_i[3] = 1 \\ SQ & | & \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 0 \\ Alt & | & \mathbf{I}_i[1] = 1, \mathbf{I}_i[2] = 1, \mathbf{I}_i[3] = 1 \end{pmatrix} \quad (3.5)$$

The “other” learners are generated by randomly assigning a probability of either 0 or 1 to each of the eight possible combinations of information. In other words, agents of this learner type represent individuals whose responses to specific information combinations are difficult for the policy maker to understand or predict. This aligns with recent research suggesting that individual heterogeneity in preferences and social learning is likely much greater than traditionally assumed (Duflo et al., 2015; Vivaldi, 2015; Mesoudi et al., 2016; Vogt et al., 2016; Kendal et al., 2018).

3.2.3 Initial conditions and interventions

The model starts with the system in an SQ norm equilibrium, where no agents have adopted the alternative Alt before the first period. This initial condition presents a challenge for the social planner, whose goal is to shift the system from the SQ equilibrium to an Alt equilibrium. The social planner believes that an Alt norm equilibrium would improve social welfare, defined in utilitarian terms as the aggregate or average payoffs across the population. To simplify, the model assumes all agents agree that coordinating on Alt is payoff dominant, meaning that $\forall i, (a + x_i) < d$, regardless of their preferences x_i .

Although all agents agree that coordinating on Alt is payoff dominant, the transition to an Alt equilibrium is not straightforward. This is because payoffs alone do not drive the dynamics of norm change in the model presented here. Agents who respond to expected payoff information, including all individualists and possibly some of the “other” learners, view SQ as risk dominant. Specifically, given the assumption that no one chose Alt in the period preceding the first period, this represents a corner case of risk dominance, where the cost of miscoordination for agent i for choosing SQ is zero based on the belief $\tilde{q}_{i,1} = 0$. In contrast, the expected cost of miscoordination for an agent i with belief $\tilde{q}_{i,1} = 0$ when choosing Alt is $x_i = (a + x_i) - a$.

Conservative learners will only ever choose SQ regardless of the observed information. Success biased learners will only choose Alt if they observe that the agent experiencing the highest payoff in their network has chosen Alt , which is impossible if we assume the fraction of agents choosing Alt to be 0 in the period prior to the first period. Similarly, conformists will stick with SQ as long as there is no majority choosing Alt .

The initial SQ equilibrium and the fact that most agents are resistant to change is the reason why the social planner designs interventions to initiate change which will be analysed here. Importantly, unlike in comparable models where each agent only responds to distributions in choices (Efferson et al., 2020; Andreoni et al., 2021; Efferson et al., 2023, 2024), the social planner in the model here can implement multifaceted strategies that adjust not

just preferences but also the informational landscape of the agents. The intervention, which targets a fraction of ϕ of agents, consist of two main aspects.

Firstly, the target group criteria determine which group of agents the intervention is aimed at. More precisely, an agent can be targeted based on her learner type and/or her preference, i.e. her x_i value. If the social planner chooses to target based on learner type, she can direct the intervention to one of the five types described in section [3.2.2](#). If the social planner can not or does not want to target based on learner types, she can also target the intervention of size ϕ across all types. Targeting based on the x_i values is also possible, and three cases are analysed. Interventions may target either a fraction of ϕ of the most amenable agents, i.e. with the lowest x_i values, or a fraction of ϕ of the most resistant agents, with the highest x_i values, or not condition on x_i values at all. Targeting based on learner type and targeting conditional on x_i values can be combined. For example, the social planner might choose to target all conformists but only the most resistant of those.

The second aspect of the intervention in addition to the targeting criteria is the content. Specifically, interventions can influence to which social information agents pay attention to and/or the preferences, i.e. the x_i values. If the social planner chooses to change preferences, this is modelled as setting the x_i values to 0, which implies that an agent targeted by such an intervention always prefers to choose *Alt* if this agent responds to private information about expected payoffs. In addition to changing preferences, it is also possible to influence the learning of agents by making certain types of information more salient. In the simulation program that is implemented by modifying the social learning strategy such that it resembles the behaviour of one the following three learner types; Individualists, conformists, or success-oriented learners. Transforming agents into conservatives or “other” learners is not considered, as having individuals consistently choose *SQ* offers no benefit, and converting them into “other” learners with random responses to the eight information combinations is too unpredictable to be useful for the social planner.

Table [3.3](#) outlines all 108 possible combinations of interventions. These include six ap-

proaches for targeting by learner type, three methods for targeting based on x_i values, three types of intervention content aimed at modifying responses to information, and two strategies for altering x_i values.

Targeting criteria		Content of intervention	
Target based on learner types	Target based on x_i values	Increase salience of information types	Change x_i values
Individualists	Most amenable	Individualist cues	Change x_i values to 0
Conformists	Most resistant	Conformist cues	No changes to x_i values
Success oriented	Randomly	Success based cues	
Conservatives			
Other learners			
Across all types			

Table 3.3: Interventions

The aim of this conceptual study is not to offer practical implementation strategies for the interventions listed in Table 3.3. Instead, the model is employed to explore the theoretical outcomes of implementing the strategies described in Table 3.3 under ideal conditions managed by a social planner. Still, it is important to note that the interventions proposed are grounded in previous research about the influence of individual heterogeneity on norm change interventions and real-world applicability.

Previous findings suggest that the structure of individual heterogeneity in preferences and social learning strongly affects whether interventions aiming to tip a population from one norm equilibrium to another one are successful or not (Vogt et al., 2016; Castilla-Rho et al., 2017; Efferson et al., 2020; Andreoni et al., 2021; Ehret et al., 2022). Hence, it is important for policy makers to be careful whom to target with an intervention and also what the content of the intervention itself is.

Research has demonstrated that preferences, social learning strategies, and network structures can be effectively measured in both laboratory and field settings (Efferson et al., 2008; Morgan et al., 2012; Efferson et al., 2015, 2016; Vogt et al., 2016; Efferson and Vogt, 2018; Alvergne and Stevens, 2021; Bellamy et al., 2022). Consequently, targeting different learner

types may become feasible if social planners intensify their efforts to conduct surveys and gather data regarding people's preferences and their responsiveness to specific types of social information. Indeed, firms appear to be already targeting specific groups on social networks with content designed to modify preferences and learning behaviours (Brady et al., 2023). In line with recent proposals advocating the use of behavioural research findings to benefit society, rather than allowing these tools to remain exclusively in corporate hands (Thaler and Sunstein, 2021; Nielsen et al., 2024), a similar case can be made for harnessing online data to tailor interventions on social networks. Policymakers could leverage this approach to target specific segments of the population with precise information campaigns.

Altering preferences in practice might be challenging. Hence, a key aspect of this study is to explore alternative approaches for initiating norm change when a social planner is unable to modify individuals' preferences. While direct preference modification may be difficult, there are documented methods for influencing attitudes, such as through edutainment. This approach, which combines entertainment with educational content to address specific issues, has been examined in various studies (La Ferrara et al., 2012; DellaVigna and La Ferrara, 2015; La Ferrara, 2016; Vogt et al., 2016). Although this study does not assess the ethical implications of using edutainment, previous research has highlighted its potential benefits, including reaching a broad and unbiased audience (Efferson et al., 2023). Such attributes may appeal to a benevolent social planner interested in fostering normative change.

Numerous studies have investigated how the prominence of specific information types can influence social learning processes. For instance, promoting individualistic learning strategies is discussed as a way of enhancing critical thinking and reducing susceptibility to misinformation (Cook et al., 2017, 2018). Importantly, evidence suggests that companies are already leveraging online data to increase the visibility of targeted information and intensify certain learning biases (Brady et al., 2023). Similarly, social planners could utilise this approach to strategically inform and influence public behaviours.

Crucially, however, social planners might not have the necessary tools to measure prefer-

ences and social learning strategies in the population. Instead, they might have to implement an intervention that targets a random segment in the population. This scenario is analysed and the results section will discuss how random targeting compares to situations in which a social planner might have more information about the preferences and learning strategies in the population.

A social planner may also face challenges in effectively altering preferences, or may possess tools that enhance the salience of certain types of information while struggling to do the same with others. These scenarios underpin the focus on the three interventions analysed in this study. Importantly, these interventions demonstrate the contributions of this model beyond traditional threshold models, which typically assume agents respond only to expected payoffs.

The first intervention explores scenarios where a social planner successfully modifies preferences of agents solely driven by expected payoffs, paralleling similar strategies in prior research (Efferson et al., 2020, 2024). The second intervention considers scenarios targeting individualists, where attempts to change preferences do not succeed, thus presenting a different dynamic in the intervention's impact. The third intervention, distinguished by its focus on enhancing the salience of success-based information, showcases the innovative aspect of the model presented here. It provides a nuanced approach for policy makers who acknowledge that individuals' decisions can extend beyond straightforward economic calculations. This third intervention is then also analysed in an environment where the population is structured into a polarised network. Recent research has shown that polarisation can prevent tipping (Efferson et al., 2020; Ehret et al., 2022), and the analysis in section 3.3.4 discusses what this means for a policy maker who can and wants to implement a success-based intervention.

3.2.4 Structure of simulation

Agents choose between SQ and Alt and play the coordination game shown in 3.1 repeatedly over $t_{max} = 100$ time periods. The system is initialised by creating $N = 100$ agents with idiosyncratic x_i values and learning strategies. At the beginning of time period $t_{max}/2$, a

policy intervention is implemented among a fraction ϕ of agents. The analysis focuses on a population that consists of the five learner types discussed in section [3.2.2](#), where each of the five learner types makes up 1/5 of the population.

Before participating in the repeated coordination game, agents are organized into networks of two distinct types. The analysis explores scenarios involving a fully connected network and a segregated one, where agents are divided into two groups based on homophily. In the latter case, agents with the lowest x_i values form one group, while those with the highest x_i values form another. This setup presents a particularly challenging environment for norm change, potentially hindering the shift to an alternative norm ([Efferson et al., 2020](#); [Ehret et al., 2022](#)).

The sequence of events within a time period, including the initialisation before the first time period, is summarised in figure [3.1](#).

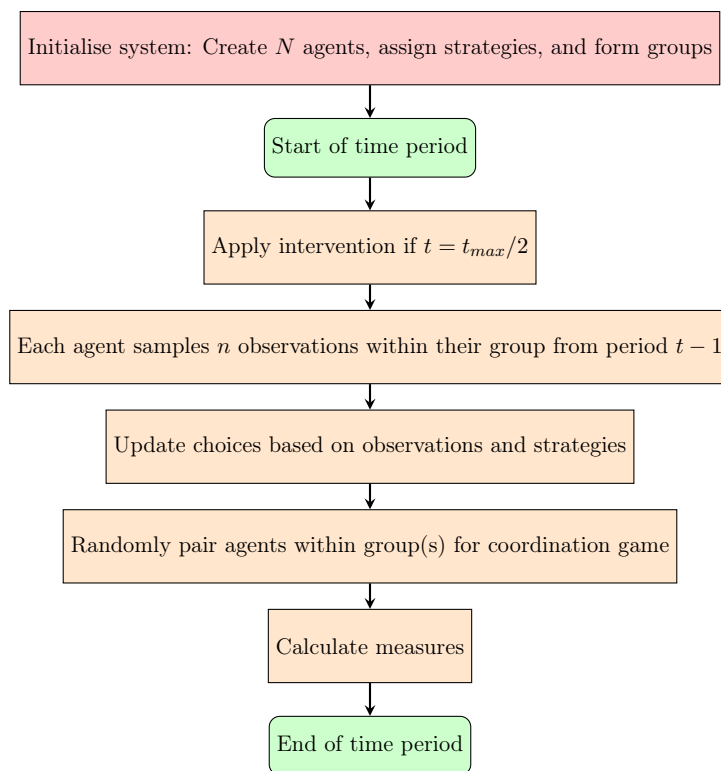


Figure 3.1: Simulation structure

3.3 Results

The main objective of this study is exploratory in nature, and the research question associated with this endeavour is formulated as follows:

Research question: *How do different types of policy intervention affect tipping the population to the *Alt* equilibrium, and what are their implications for miscoordination, welfare, and inequality, given agents' access to information on expected payoff, majority behaviour, and the behaviour of the most successful agent in terms of payoff?*

However, an auxiliary hypothesis can be formulated based on findings of two recent studies with comparable models. Specifically, the results of those models show that targeting resistant or a random sample of agents is generally more successful in tipping the population of agents to the *Alt* equilibrium than targeting the most amenable agents (Efferson et al., 2020, 2024). Although the model here includes multiple different learner types, one form of intervention that is analysed targets randomly across all those learner types and conditional on the x_i values. In other words, that specific analysis shows how targeting the policy measure to more resistant, more amenable, or a random fraction of ϕ of agents performs in terms of tipping the population of agents to the *Alt* equilibrium, miscoordination, average payoff, and inequality. Hence, the following hypothesis is put forward here: Interventions targeting a random sample of agents or targeting the most resistant agents lead to more tipping and less miscoordination compared to interventions targeting the most amenable agents. The hypothesis does not concern average payoffs and inequality because the intervention simulated in the paper by Efferson et al. (2024) does not change x_i values and instead changes the payoff structure in a different way. As shown in section 3 in the supplementary material, the results support this auxiliary hypothesis.

A description of all the parameters are shown in table 3.4. The results for the third form of intervention content, namely the transformation of agents into conformists, are presented

in the supplementary material (section 3.4). It is also important to note that the proportion of people targeted is always $\phi = 0.2$. This ensures that the analysis of those interventions targeting specific learner types, each of which make up 1/5 of the population, is comparable to the intervention targeting across all learner types.

Parameter	Definition
$t_{max} = 100$	Total number of time periods.
$N = 100$	Population size.
$n = 10$	Number of observations each agent makes within their group to derive their information array \mathbf{I}_i .
$\phi = 0.2$	Size of intervention as a fraction of the agent population.
$\alpha = 3, \beta = 2$	Shape parameters of beta distribution, from which the x_i values are drawn.
$a = 0.75, b = 1, d = 2$	Incentives in coordination game.
$\mu \in \{0.01\}$	Error in decision-making. A focal agent i chooses the norm defined by her strategy $S_i(\mathbf{I}_i)$ and her belief $\tilde{q}_{i,t}$ with probability $1 - \mu$.

Table 3.4: Definition of parameters.

3.3.1 Turning agents into individualists and changing preferences

Figure [3.2](#) shows the results of the intervention that turns a fraction ϕ of agents into individualist learners and changes their preferences to $x_i = 0$. The intervention targeting conservative learners is most effective in increasing the fraction of *Alt* choices, decreasing miscoordination, increasing average payoff, and decreasing inequality of payoffs.

This makes sense, given that those learners are most resistant to change with $x_i = 1$. The dynamics for targeting individualists, other learners, and targeting across all types are not significantly different. Those targets all perform better than targeting conformists and success based learners. This can be explained as follows. Changing preferences of individualists is effective because those learners pay attention to expected payoff, which is influenced by the x_i values. Targeting across learner types and targeting other learners is also effective in achieving a majority of agents choosing *Alt* because it utilises the indirect effects of norm change resulting from conformists and success based learners. In other words, the direct effects

of the intervention first convinces a fraction of agents to adopt *Alt*, which then increases the fraction of *Alt* sufficiently enough to make conformists and success based learners adopt *Alt* as well. Based on similar reasoning, targeting conformists and success based learners is less effective because it prevents part of said indirect effects.

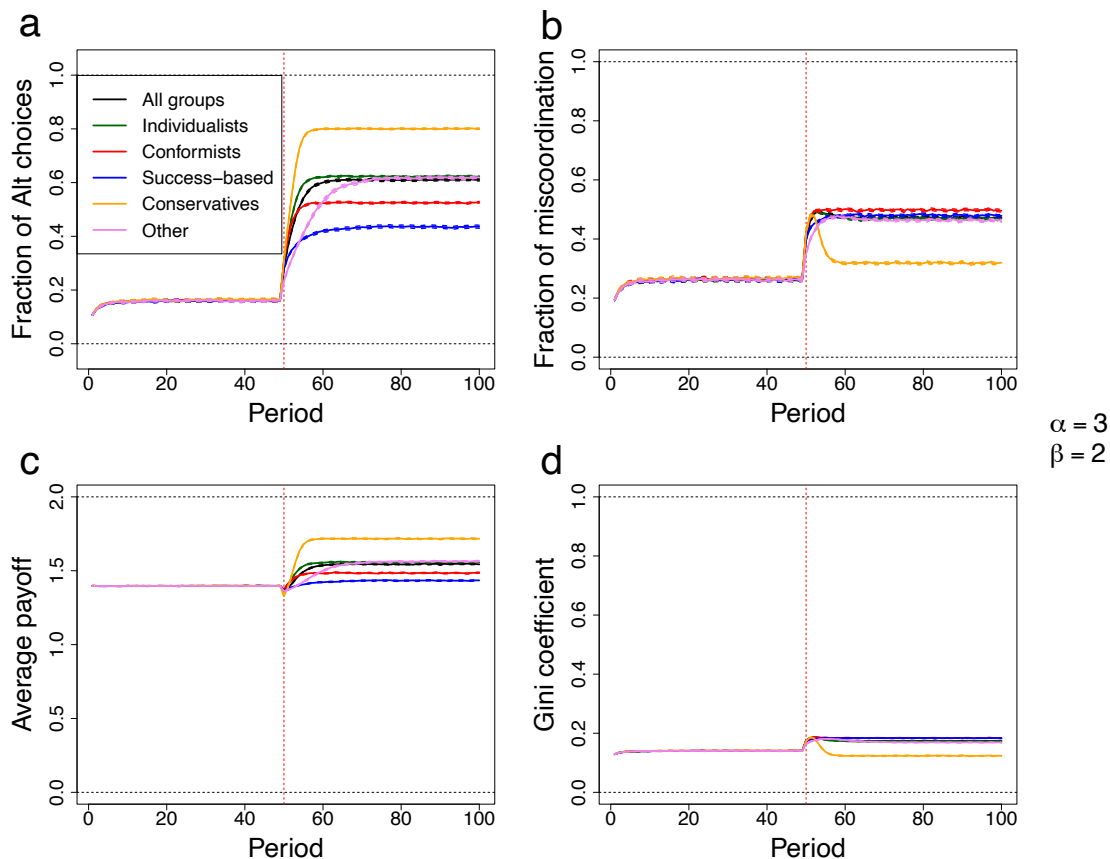


Figure 3.2: The figure shows the dynamics in a fully connected network for an intervention that turns a fraction of ϕ of agents into individualist learners with $x_i = 0$. Targeting different types of learners is shown in different colours. The black line shows the results for targeting a fraction of ϕ agents across all learning types. The solid lines show values averaged over 1000 simulations. The dashed lines show 95% bootstrapped confidence intervals. Graphs show the fraction of agents choosing *Alt* (a), the fraction of miscoordination (b), average payoff (c) and the Gini coefficient (d). Aside from α and β , parameter values are $N = 100$, $n = 10$, and $\phi = 0.2$.

3.3.2 Turning agents into individualists without changing preferences

Changing preferences might be challenging or even impossible. To have a direct comparison to the intervention shown in the previous section, figure 3.3 shows the results for an intervention that also turns a fraction of agents into individualist learners, but does not change preferences.

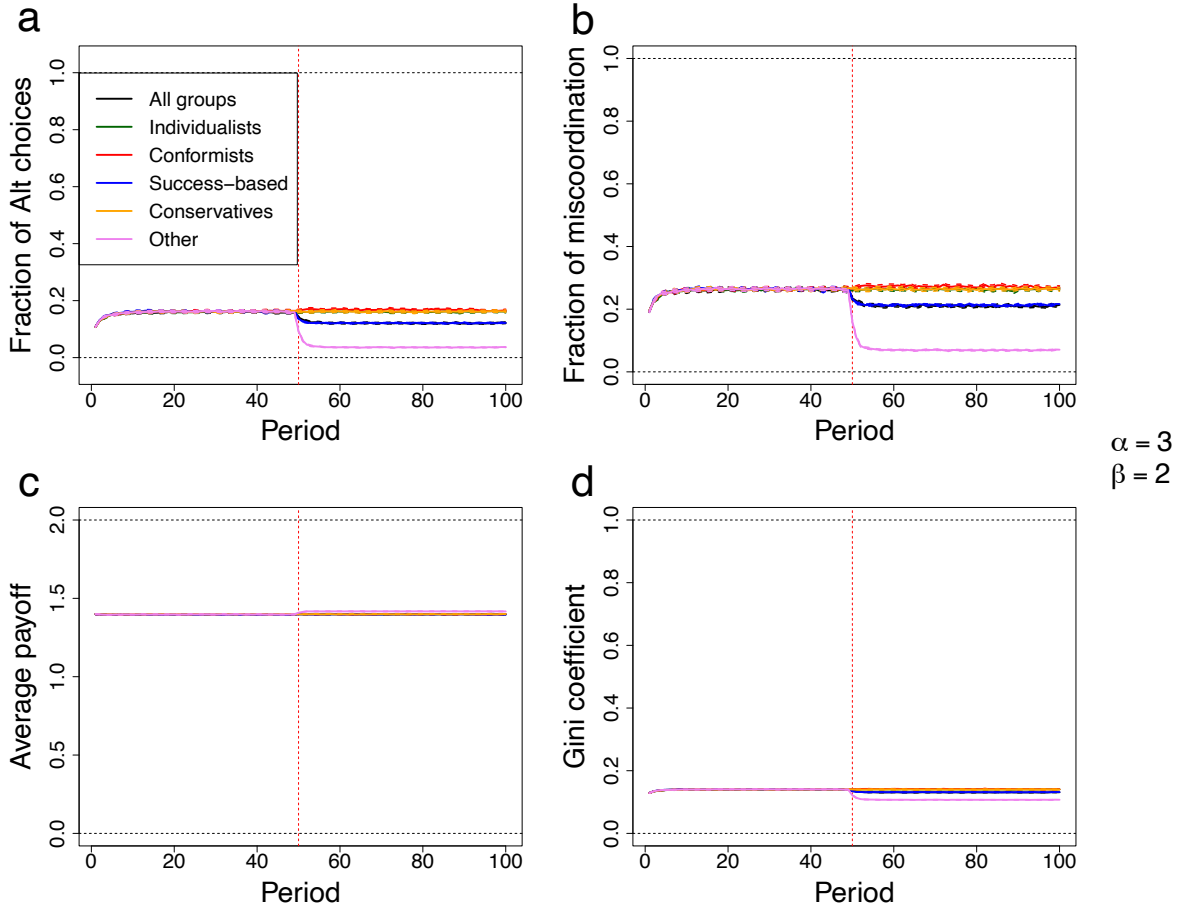


Figure 3.3: The figure shows the dynamics in a fully connected network for an intervention that turns a fraction of ϕ of agents into individualist learners without changing their x_i values. Targeting different types of learners is shown in different colours. The black line shows the results for targeting a fraction of ϕ agents across all learning types. The solid lines show values averaged over 1000 simulations. The dashed lines show 95% bootstrapped confidence intervals. Graphs show the fraction of agents choosing *Alt* (a), the fraction of miscoordination (b), average payoff (c) and the Gini coefficient (d). Aside from α and β , parameter values are $N = 100$, $n = 10$, and $\phi = 0.2$.

It is not surprising that this type of intervention is less effective than the one presented in figure [3.2](#) because no preferences are changed here. The more interesting result concerns the intervention targeting other and success-oriented learners. Importantly, both of those two types of learners include some agents who had tendencies of choosing *Alt* pre-intervention. Specifically, some other learners chose *Alt* simply because they would choose *Alt* regardless of the information observed. Some success oriented learners chose *Alt* pre-intervention because they observed other agents coordinating on *Alt*. Hence, turning those agents into individualists who only consider information about expected payoff causes some of them to stop choosing *Alt* because of the low *Alt* fraction.

3.3.3 Turning agents into success-based learners

An intriguing question is whether the effects of an intervention that changes preferences can be replicated by simply making certain types of information more salient, without altering the preferences themselves. The results in the previous section show that this is not possible if the intervention increases the salience of information about expected payoffs. In contrast to typical models of norm change which assume that every agent is a conformist, this new model here adds success-based learning. This is especially interesting in this context where an alternative norm is payoff dominant. Crucially, if some agents lead the way and coordinate on *Alt* despite the *SQ* equilibrium, agents biased by success-based information might copy that behaviour and adopt *Alt* as well.

This is exactly what happens for certain intervention targets. Specifically, figure [3.4](#) shows that the fraction of *Alt* choices increases for all targets except for targeting success based learners and other learners. It is obvious that targeting success oriented learners has no effect. Targeting other learners has a negative effect in terms of increasing *Alt* choices. The other learners are the ones increasing the fraction of *Alt* choices pre-intervention above zero. Turning them into success-oriented learners while not having many agents yet coordinating on *Alt* results in a backlash in terms of the *Alt* fraction. However, it also decreases miscoordination and inequality, because the fraction of *Alt* approaches zero.

In contrast, the other targeting strategies all increase the fraction of *Alt*. All conservatives, conformists, and most individualists did not choose *Alt* pre-intervention due to a majority choosing *SQ*. Having a fraction of other learners in the population implies that some agents coordinate on *Alt* from time to time. Hence, turning conservatives, individualists, and conformists into success-based learners makes many of them adopt *Alt* post-intervention.

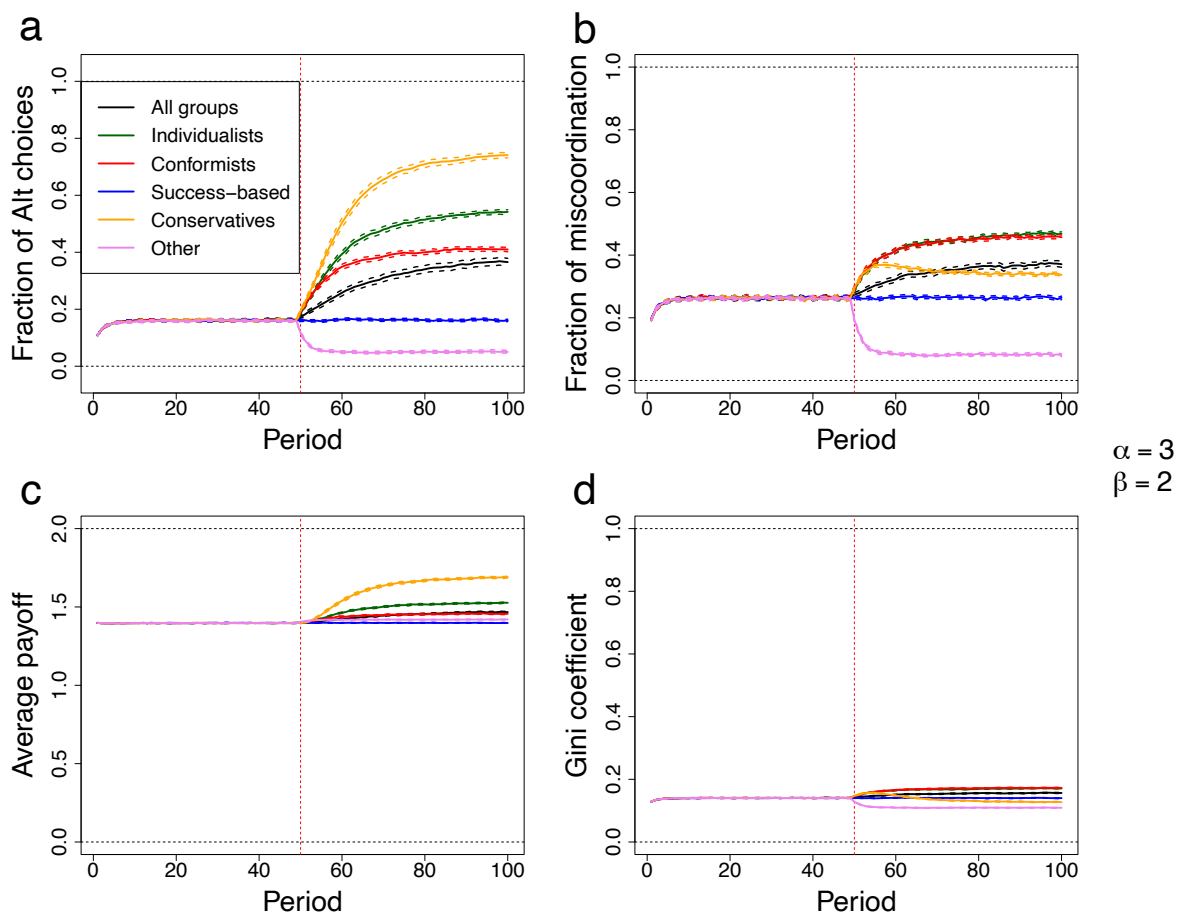


Figure 3.4: The figure shows the dynamics in a fully connected network for an intervention that turns a fraction of ϕ of agents into success based learners without changing their x_i values. Targeting different types of learners is shown in different colours. The black line shows the results for targeting a fraction of ϕ agents across all learning types. The solid lines show values averaged over 1000 simulations. The dashed lines show 95% bootstrapped confidence intervals. Graphs show the fraction of agents choosing *Alt* (a), the fraction of miscoordination (b), average payoff (c) and the Gini coefficient (d). Aside from α and β , parameter values are $N = 100$, $n = 10$, and $\phi = 0.2$.

3.3.4 Turning agents into success-based learners in a polarised population

The last analysis discusses a polarised network structure where the most amenable agents with the lowest x_i values are in group 1 and the most resistant agents with the highest x_i values are in group 2. Again, a fraction of ϕ agents is turned into success-oriented learners without changing their preferences.

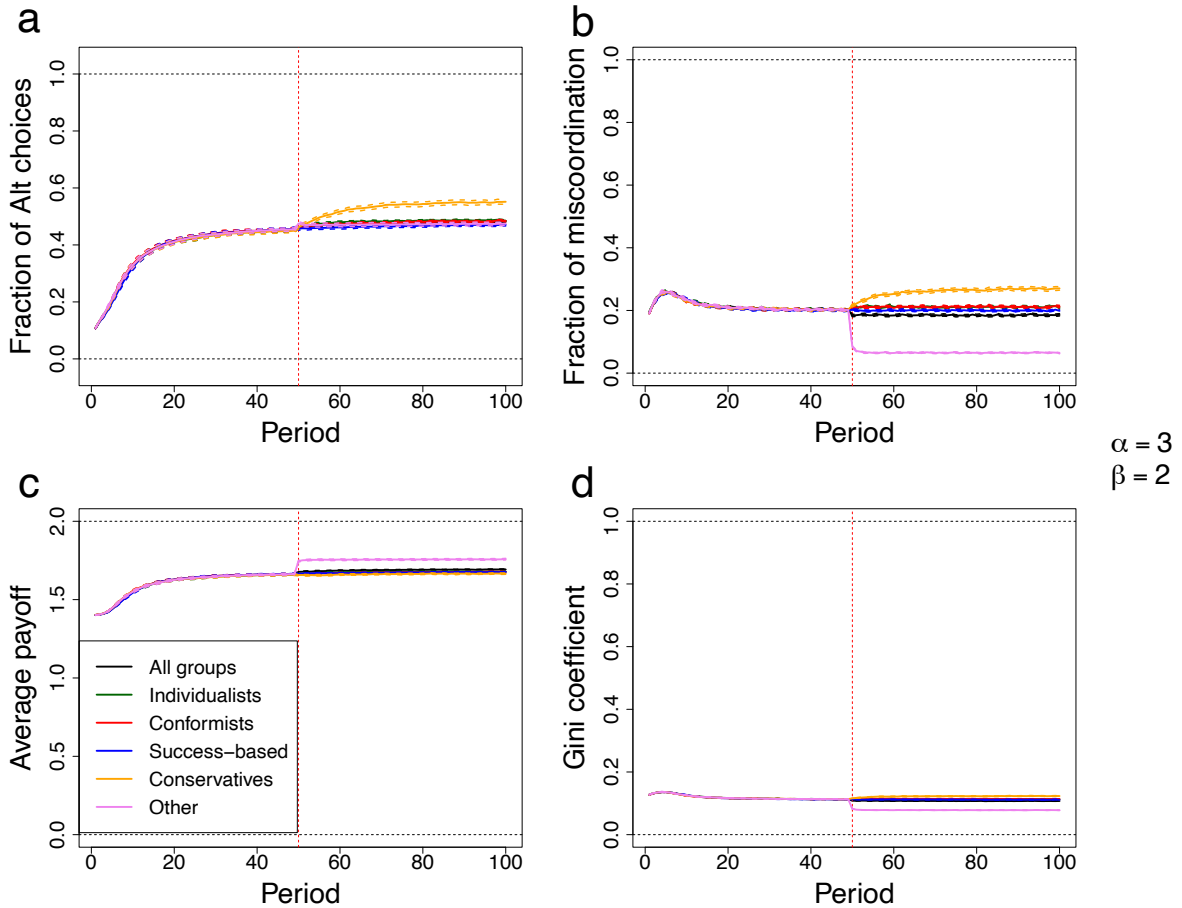


Figure 3.5: The figure shows the dynamics in a polarised network with two groups for an intervention that turns a fraction of ϕ of agents into success based learners without changing their x_i values. Targeting different types of learners is shown in different colours. The black line shows the results for targeting a fraction of ϕ agents across all learning types. The solid lines show values averaged over 1000 simulations. The dashed lines show 95% bootstrapped confidence intervals. Graphs show the fraction of agents choosing *Alt* (a), the fraction of miscoordination (b), average payoff (c) and the Gini coefficient (d). Aside from α and β , parameter values are $N = 100$, $n = 10$, and $\phi = 0.2$.

Figure 3.5 shows two key findings. First, in a polarised network of agents, the pre-intervention equilibrium fraction of *Alt* choices is higher than in a fully connected network. Second, except for the target of other learners, the intervention is generally less effective in a polarised network than in the fully connected network shown in figure 3.4 in terms of relative changes in the fraction of *Alt* choices.

The reason for a higher pre-intervention fraction of *Alt* here is that group 1 with the most amenable agents almost fully tips to the *Alt* equilibrium, as can be seen in figure 3.6. The fact that the intervention is less effective in a polarised network can also be explained with the dynamics in the separate groups. Group 1 reaches a very high fraction of *Alt* already pre-intervention, making an intervention less effective in this group. In group 2, the fraction of *Alt* choices stays very low pre-intervention because the agents in this group are very resistant to change. Crucially, the *Alt* fraction stays below 10% in group 2, whereas the pre-intervention fraction of *Alt* choices in the fully connected network stays above 15% (c.f. figure 3.4). The lower fraction of *Alt* choices decreases the number of agents coordinating on *Alt*, making an intervention that turns agents into success-oriented learners less effective.

3.4 Discussion

This study shows that the effectiveness of the different types of intervention varies greatly depending on the target group and the nature of the intervention. Interventions promoting individualistic learning and shifting preferences prove to be effective. This approach capitalises on the individual learner's focus on expected payoff. Conversely, interventions that merely transition learners to individualistic strategies without modifying their underlying preferences demonstrate minimal impact, underscoring the critical role of preference alignment when making private information about expected payoff more salient.

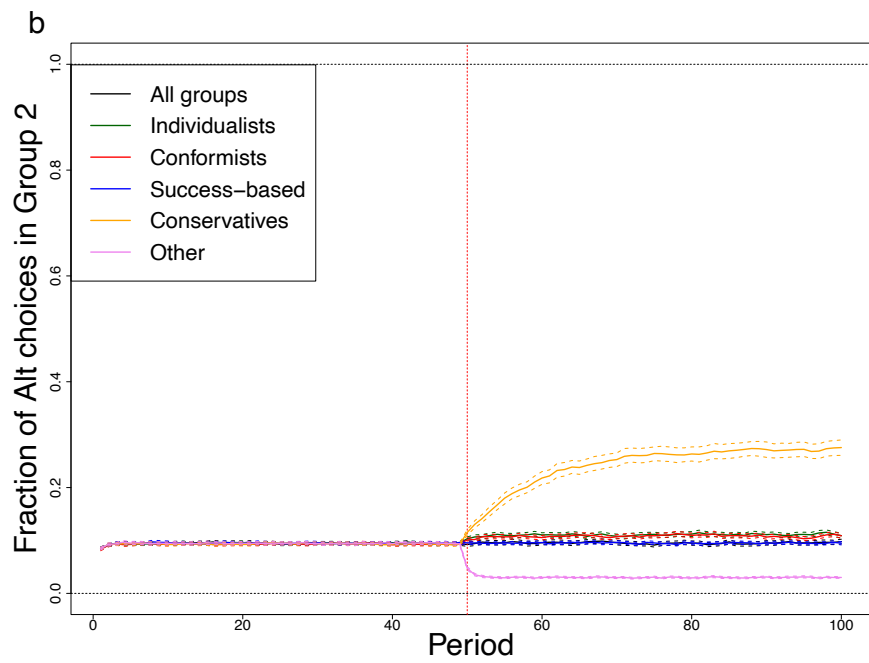
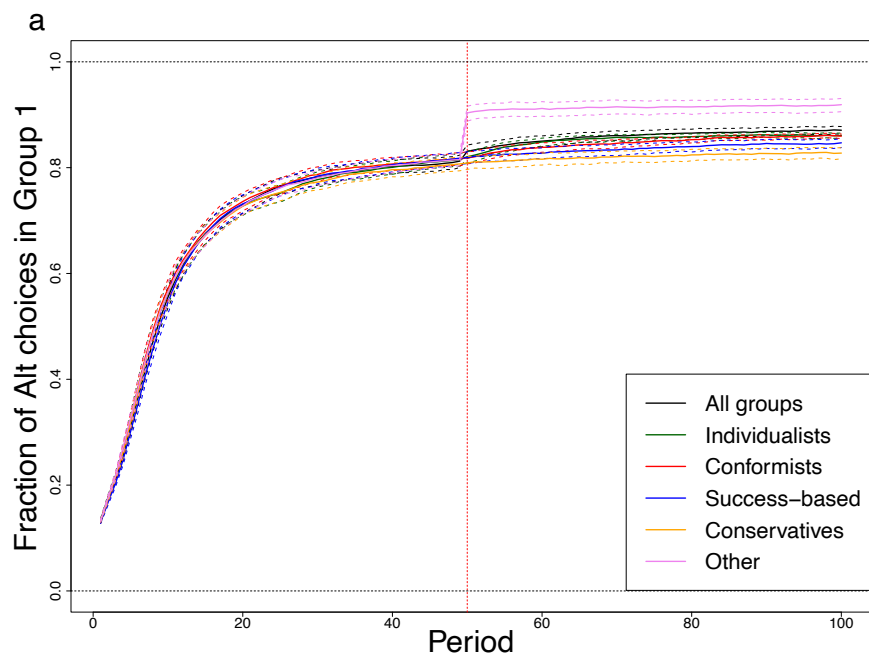


Figure 3.6: Separate group dynamics corresponding to the population shown in figure 3.5.

A key innovation of this model is the inclusion of success-based learning. Interventions that increase the visibility of successful agents' behaviour, i.e. those achieving the highest payoff from adopting *Alt*, are promising for promoting norm change. This approach shows substantial real-world applicability, indicating that promoting the success stories of norm adopters can effectively drive cultural evolution. However, the effectiveness of such interventions depends critically on the target group. Targeting other learners, for example, can lead to adverse outcomes. The risk of backlash is something that policy makers have to be very mindful about if they want to initiate cultural change (Efferson et al., 2020). This underscores the need for detailed knowledge about the distribution of social learning strategies within a population, especially given the significant heterogeneity observed in recent studies (Mesoudi et al., 2016; Kendal et al., 2018). Interestingly, however, if a social planner manages to identify specific groups of learners in the population, a success based intervention might be very promising. In particular, changing preferences might not be possible for a policy maker. Instead, it might be feasible to amplify certain learning biases, such as success bias (Brady et al., 2023). If such an intervention is successful in identifying certain learner types, the results show that making success based information more salient might be a promising substitute for a preference changing intervention.

The discussion on interventions that promote conformist behaviour, detailed in the supplementary material (section 3.4), reveals that such approaches are generally ineffective in populations predominantly adhering to the status quo. However, targeting conservatives with this type of intervention can moderately increase adoption of *Alt*. Therefore, encouraging conformism among conservatives might be one type of intervention to consider if there is a small proportion of people who already adhere to the desired norm. This could be a viable approach if a large proportion of individuals are conservatives and if promoting conformism is easier to implement than other measures for certain reasons. However, such a policy should probably be combined with other forms of intervention to achieve a complete tipping.

More broadly, however, the results presented in this study suggest an important finding.

Separately, interventions targeting particular groups of agents or promoting specific types of content might often be insufficient to achieve full tipping, and by extension higher average payoffs, less miscoordination and lower inequality. However, different interventions could be combined to achieve that. For example, if an initial intervention succeeds in convincing a critical mass of people to adopt the alternative norm, a subsequent intervention that emphasises success-based information more strongly could have an impact. At a later stage, when the proportion of people following the alternative norm is even higher, a policy to encourage conformity could become effective. Such a combined approach could be particularly important if the population is structured in a polarised network, so that endogenous change in certain groups full of resistant people is unlikely. Indeed, a step-wise approach to initiating cultural change may be a particularly important concept in relation to harmful norms (Gulesci et al., 2021), and future iterations of this model could test how effective it is to combine different interventions at different points in time.

Like any model, the one presented here contains many simplifications. However, if this type of model is to be used in conjunction with empirical research as a tool to inform policy, some restrictions will need to be relaxed in future iterations. For instance, it is likely that people respond differently to interventions. One possibility is to link the response to interventions with an agent's inherent resistance to norm change in the form of x_i values. In other words, agents may not respond uniformly to an intervention. Prior research indicates that a larger intervention may be required in that case to achieve results comparable to those of an always effective intervention (Efferson et al., 2024). In addition, the likelihood of responding to interventions could also be related to the learning strategy. It seems plausible, for example, to assume that individualists respond less to interventions that promote social information than conformists and success-oriented learners.

Additionally, a social planner may not always succeed in altering individuals' social learning behaviours, and this aspect has not been explored in the current model. Future versions could investigate the effects of unsuccessful changes in social learning. Instead, this study has

focused on whether interventions that leverage specific types of information, without altering preferences, can be as effective as those that do manage to change preferences. Importantly, there is increasing evidence that companies are leveraging online data to enhance particular learning biases on social platforms (Brady et al., 2023). This approach could be valuable for policymakers to consider adopting.

Currently, the x_i values are not correlated to specific learner types. If future empirical studies indicate that certain preferences are more pronounced in people who are particularly receptive to certain types of information, x_i values could be correlated with specific learning strategies.

This study focuses on a very small set of learning strategies. Importantly, each well-defined type of learner focuses on only one type of information. Only 1/5 of the population exhibits random strategies that can potentially be influenced by multiple types of information simultaneously. However, it is very likely that most people are constantly influenced by many different types of information (Mesoudi et al., 2016; Kendal et al., 2018).

At the same time, starting with simpler versions of norm change models that incorporate multiple types of information could also facilitate the development of analytic models alongside simulation models. Such an approach is advantageous as analytic models enable the derivation of clear predictions for steady-state conditions for specific parameter combinations of the corresponding simulation model. This could be particularly insightful when simpler analytic models, which might consider only three types of learners instead of the five simulated here, are used to compare norm change dynamics with analytic models of threshold models that focus solely on agents responding to expected payoffs (Efferson et al., 2024).

And yet this first step of integrating more forms of learning into models of norm change is promising. The results presented here emphasise how sensitive the effectiveness of interventions is to the specific structure of learning and preferences in the population. This is in line with more general findings that emphasise the importance of first better understanding individual heterogeneity in order to capture the evolutionary dynamics of norm change (Young,

[2009; Vogt et al., 2016; Efferson et al., 2020; Constantino et al., 2022; Efferson et al., 2023, 2024]. The results on success-based interventions suggest that there are potentially effective interventions whose evolutionary dynamics we can only analyse if we take into account key empirical findings on social learning. Therefore, future empirical research is essential to increase the value of modelling by providing detailed information about specific network structures and the distribution of learning strategies and preferences relevant to the norm changes targeted by interventions. Given the increasing interest in promoting endogenous cultural change for the benefit of society, the pursuit of such a research agenda is crucial (Nyborg et al., 2016; Otto et al., 2020; Constantino et al., 2022; Efferson et al., 2023).

Acknowledgements and funding

The author acknowledges support from the Swiss National Science Foundation (Grant Nr. 100018_185417/1).

Author contributions

LvF designed and programmed the model. LvF analysed the data and wrote the article.

Research transparency and reproducibility

The code for the agent-based model can be found on <https://github.com/LukasVonFluee/StrategySpaceNormChange.git>. The supplementary material is available on <https://osf.io/jdq8a/>.

Bibliography

- Alvergne, A. and Stevens, R. (2021). Cultural change beyond adoption dynamics: Evolutionary approaches to the discontinuation of contraception, *Evolutionary Human Sciences* pp. 1–45.
- Andreoni, J., Nikiforakis, N. and Siegenthaler, S. (2021). Predicting social tipping and norm change in controlled experiments, *Proceedings of the National Academy of Sciences* **118**(16).
- Bellamy, A., McKay, R., Vogt, S. and Efferson, C. (2022). What is the extent of a frequency-dependent social learning strategy space?, *Evolutionary Human Sciences* **4**: e13.
- Boyd, R. and Richerson, P. J. (1985). *Culture and the Evolutionary Process*, University of Chicago press.
- Brady, W. J., Jackson, J. C., Lindström, B. and Crockett, M. (2023). Algorithm-mediated social learning in online social networks, *Trends in Cognitive Sciences* .

- Castilla-Rho, J. C., Rojas, R., Andersen, M. S., Holley, C. and Mariethoz, G. (2017). Social tipping points in global groundwater management, *Nature Human Behaviour* **1**(9): 640–649.
- Constantino, S. M., Sparkman, G., Kraft-Todd, G. T., Bicchieri, C., Centola, D., Shell-Duncan, B., Vogt, S. and Weber, E. U. (2022). Scaling up change: A critical review and practical guide to harnessing social norms for climate action, *Psychological Science in the Public Interest* **23**(2): 50–97.
- Cook, J., Ellerton, P. and Kinkead, D. (2018). Deconstructing climate misinformation to identify reasoning errors, *Environmental Research Letters* **13**(2): 024018.
- Cook, J., Lewandowsky, S. and Ecker, U. K. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence, *PloS One* **12**(5): e0175799.
- DellaVigna, S. and La Ferrara, E. (2015). Economic and social impacts of the media, *Handbook of media economics*, Vol. 1, Elsevier, pp. 723–768.
- Dufo, E., Dupas, P. and Kremer, M. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from kenyan primary schools, *Journal of Public Economics* **123**: 92–110.
- Efferson, C., Ehret, S., von Flüe, L. and Vogt, S. (2024). When norm change hurts, *Philosophical Transactions of the Royal Society B* **379**(1897): 20230039.
- Efferson, C., Lalive, R., Cacault, M. P. and Kistler, D. (2016). The evolution of facultative conformity based on similarity, *PLOS One* **11**(12): e0168551.
- Efferson, C., Lalive, R., Richerson, P. J., McElreath, R. and Lubell, M. (2008). Conformists and mavericks: The empirics of frequency-dependent cultural transmission, *Evolution and Human Behavior* **29**(1): 56–64.

- Efferson, C. and Vogt, S. (2018). Behavioural homogenization with spillovers in a normative domain, *Proceedings of the Royal Society B: Biological Sciences* **285**(1879): 20180492.
- Efferson, C., Vogt, S., Elhadi, A., Ahmed, H. E. F. and Fehr, E. (2015). Female genital cutting is not a social coordination norm, *Science* **349**(6255): 1446–1447.
- Efferson, C., Vogt, S. and Fehr, E. (2020). The promise and the peril of using social influence to reverse harmful traditions, *Nature Human Behaviour* **4**(1): 55–68.
- Efferson, C., Vogt, S. and von Flüe, L. (2023). Activating cultural evolution for good when people differ from each other, in J. Kendal, R. Kendal and J. Tehrani (eds), *Oxford Handbook of Cultural Evolution*, Oxford University Press, chapter TBD.
- Ehret, S., Constantino, S. M., Weber, E. U., Efferson, C. and Vogt, S. (2022). Group identities can undermine social tipping after intervention, *Nature Human Behaviour* pp. 1–11.
- Granovetter, M. (1978). Threshold models of collective behavior, *American Journal of Sociology* **83**(6): 1420–1443.
- Gulesci, S., Jindani, S., La Ferrara, E., Smerdon, D., Sulaiman, M. and Young, H. (2021). A stepping stone approach to understanding harmful norms.
- Harsanyi, J. C. and Selten, R. (1988). *A General Theory of Equilibrium Selection in Games*, The MIT Press.
- Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M. and Jones, P. L. (2018). Social learning strategies: Bridge-building between fields, *Trends in Cognitive Sciences* **22**(7): 651–665.
- La Ferrara, E. (2016). Mass media and social change: Can we use television to fight poverty?, *Journal of the European Economic Association* **14**(4): 791–827.
- La Ferrara, E., Chong, A. and Duryea, S. (2012). Soap operas and fertility: Evidence from brazil, *American Economic Journal: Applied Economics* **4**(4): 1–31.

- Mäs, M. and Nax, H. H. (2016). A behavioral study of “noise” in coordination games, *Journal of Economic Theory* **162**: 195–208.
- Mesoudi, A., Chang, L., Dall, S. R. and Thornton, A. (2016). The evolution of individual and cultural variation in social learning, *Trends in Ecology & Evolution* **31**(3): 215–225.
- Morgan, T. J., Rendell, L. E., Ehn, M., Hoppitt, W. and Laland, K. N. (2012). The evolutionary basis of human social learning, *Proceedings of the Royal Society B: Biological Sciences* **279**(1729): 653–662.
- Newton, J. (2021). Conventions under heterogeneous behavioural rules, *The Review of Economic Studies* **88**(4): 2094–2118.
- Nielsen, K. S., Cologna, V., Bauer, J. M., Berger, S., Brick, C., Dietz, T., Hahnel, U. J., Henn, L., Lange, F., Stern, P. C. et al. (2024). Realizing the full potential of behavioural science for climate change mitigation, *Nature Climate Change* **14**(4): 322–330.
- Nyborg, K., Anderies, J. M., Dannenberg, A., Lindahl, T., Schill, C., Schlüter, M., Adger, W. N., Arrow, K. J., Barrett, S. and Carpenter, S. (2016). Social norms as solutions, *Science* **354**(6308): 42–43.
- Otto, I. M., Donges, J. F., Cremades, R., Bhowmik, A., Hewitt, R. J., Lucht, W., Rockström, J., Allerberger, F., McCaffrey, M. and Doe, S. S. (2020). Social tipping dynamics for stabilizing Earth’s climate by 2050, *Proceedings of the National Academy of Sciences* **117**(5): 2354–2365.
- Thaler, R. H. and Sunstein, C. R. (2021). *Nudge: The Final Edition*, Penguin.
- Vivalt, E. (2015). Heterogeneous treatment effects in impact evaluation, *American Economic Review* **105**(5): 467–70.
- Vogt, S., Zaid, N. A. M., Ahmed, H. E. F., Fehr, E. and Efferson, C. (2016). Changing cultural attitudes towards female genital cutting, *Nature* **538**(7626): 506–509.

Young, H. P. (2009). Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning, *American Economic Review* **99**(5): 1899–1924.

Young, H. P. (2015). The evolution of social norms, *Annual Review of Economics* **7**(1): 359–387.