UNIVERSIDADE DE LISBOA

Faculdade de Psicologia

The Social Amplification of Bias

Cristina Soares Pacheco Mendonça

Orientador(es): Prof. Doutor André Otelo Paraíba Mata
Prof. Doutor Mário Augusto de Carvalho Boto Ferreira
Prof. Doutor Klaus Fiedler

Tese especialmente elaborada para obtenção do grau de Doutor em Psicologia,

especialidade de Psicologia Social

2020

UNIVERSIDADE DE LISBOA

Faculdade de Psicologia

The Social Amplification of Bias

Cristina Soares Pacheco Mendonça

Orientador(es): Prof. Doutor André Otelo Paraíba Mata
Prof. Doutor Mário Augusto de Carvalho Boto Ferreira
Prof. Doutor Klaus Fiedler

Tese especialmente elaborada para obtenção do grau de Doutor em Psicologia, especialidade de Psicologia Social

Júri:
Presidente: Doutora Isabel Maria de Santa Bárbara Teixeira Nunes Narciso Davide, Professora Associada e Vice-Presidente do Conselho Científico da Faculdade de Psicologia da Universidade de Lisboa;
Vogais:

– Doutora Margarida e Sá de Vaz Garrido, Professora Associada da Escola de Ciências Sociais e Humanas do ISCTE-Instituto Universitário de Lisboa;

– Doutora Teresa Maria Freitas Teixeira de Morais Garcia Marques, Professora Catedrática da Unidade de Investigação em Psicologia Cognitiva, do Desenvolvimento e da Educação do ISPA – Instituto Universitário de Ciências Psicológicas, Sociais e da Vida;

– Doutor Leonel Garcia Marques, Professor Catedrático da Faculdade de Psicologia da Universidade de Lisboa;

– Doutor André Otelo Paraíba Mata, Professor Auxiliar da Faculdade de Psicologia da Universidade de Lisboa, orientador.

Projeto apoiado e financiado pela Fundação para a Ciência e Tecnologia
(PD/BD/113491/2015)

2020

-Tie them together, Piglet, can you tie a knot? – *Rabbit*

-I cannot. – *Piglet*

-Ah, so you can knot? – *Rabbit*

-No, I cannot knot. – *Piglet*

-Not knot? – *Rabbit*

-Who's there? – *Pooh*

-Pooh! – *Rabbit*

-Pooh who? – *Pooh*

-No, Pooh! It's -, Piglet, you'll need more than two knots. – *Rabbit*

-Not possible. – *Piglet*

-Ah, so it *is* possible to knot those pieces! – *Owl*

-Not these pieces. – *Piglet*

-Yes, knot those pieces. – *Pooh*

-Why not? – *Piglet*

-Cause it's all for naught. – *Eeyore*

– Winnie-the-Pooh

Acknowledgments

First of all, I would like to thank my advisor André Mata who was the best mentor I could wish for. You not only inspired me to become the best researcher I could be, but greatly supported me in doing so.

To my co-advisor Mário Boto Ferreira, who, among many things, got me into Judgment and Decision Making research and pursuing a PhD.

To my co-advisor Prof. Klaus Fiedler, who is a great inspiration and gave me the opportunity to visit his excellent research group. My stays in Heidelberg were a time of intense academic growth.

To Prof. Leonel Garcia-Marques for his profound knowledge and enthusiasm in our Science (and craft beer!) that he is always willing to share with all of us.

To all the teachers I had in Faculdade de Psicologia, in particular to Prof. Frederico Marques, who shattered my dreams of becoming a psychoanalyst (yes, really) and replaced them with dreams of scientific knowledge.

To all the founding Professors of LiSP for having created this amazing project and giving me the opportunity to be a part of it.

To Ana Lapa, for infinitely more than I can put down in words but, for the current purposes, for the long hours discussing theory, statistics, writing, and the state of our Science.

To Sara Hagá, for creating the Write On group and for being such a beautiful friend with a beautiful family. To the Write On colleagues for their feedback and motivation.

To Jerônimo: there isn't enough sushi in the world to repay you for taking care of the bureaucracy of this dissertation, but I sure am going to try. To my other close friends that shared most of the joys and sorrows of this PhD: Mara, Joana, and Filipa. I can't wait to celebrate with some Bailey's, Aperol Spritz, and handmade Sangria.

To my A321 colleagues who turned an empty room into one filled with great discussions, coffee, and good times. To all my LiSP colleagues, especially those of my cohort who shared with me the confusion and pride of being the 1st generation of LiSP students.

To Nuno, my good friend who, come what may, had the patience to keep our weekly card or boardgame session. Dark forces may stir against us, but I know we will always try our best to stop the sinister plots from being fulfilled.

To Maria Inês, for being such a good flat mate. I forgive you for having a bigger game collection than me only because you were such a good sport when we unnecessarily risked Ana killing you in Arkham Horror.

To all my vegetarian dinner friends: André, Andreia, Sílvia, Sónia, Conceição, Jorge, Bia, Mariana, among others over the years. Can we finally celebrate with the Mushroom Risotto contest?

To my mother and father who have always loved me unconditionally and gave me full support, even when I chose a high-unemployment University course, - I love you. To my brothers who, despite being a pain in the neck during my childhood years, I have come to love unconditionally. To my little beautiful nephew, to my sisters-in-law, to my godmother and godfather, to my grandparents, and to all my many uncles and aunts, cousins, and grand cousins. People say it is quality over quantity, but our family is evidence you can have both.

Finally, Isaac Newton once said, "If I have seen further it is by standing on the shoulders of Giants", but unfortunately, many of the giants we stand on are like Moai statues on Easter Island, - at least partially buried and forgotten. Because this dissertation is about the way previous generations influence new generations, many times quite indirectly, I cannot end without thanking all the people who I do not know personally (many of whom are actually long gone) but whose creations I benefited from. So, I want to thank those who, instead of simply going through life and doing enough just surviving, made the extra effort develop our collective knowledge. I can only dream that I may, like you, one day have made a small but significant contribution that, in the future, accumulated with all of yours and many others', ends up forming the giants that will support the Newtons of the future.

The research reported on this thesis was funded by

Fundação para a Ciência e a Tecnologia,

Doctoral Grant PD/BD/113491/2015.

LISBON SP
Lisbon PhD in Social Psychology
an FCT Programme
PD F PROGRAMAS DE DOUTORAMENTO FCT

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

POPH
PROGRAMA OPERACIONAL POTENCIAL HUMANO

U LISBOA | UNIVERSIDADE DE LISBOA

Faculdade de Psicologia
UNIVERSIDADE DE LISBOA

CENTRO DE INVESTIGAÇÃO
EM CIÊNCIA PSICOLÓGICA
Research Center for Psychological Science
UID/PSI/04527/2016

Abstract

The area of judgment and decision-making became very influential under the heuristics-and-biases research program by revealing that the prevailing economic theories lacked realism: The human being is not, as these theories had assumed, perfectly rational. Instead, people frequently rely on simplified, mostly automatic and unconscious strategies (i.e., heuristics) that enable the human mind, with its limitations, to provide acceptable judgments and decisions, but that can also lead to systematic errors (i.e., biases) under certain conditions. Yet, the judgment and decision-making research itself still neglects essential aspects of reality that can have an important influence on the way people judge and make decisions. Among these aspects is the fact that, as people live in societies, they receive information from, and transmit it to, other people. The main hypothesis explored in the current dissertation is that this social dynamic will lead to a social amplification of bias: As information travels from one person to the next, the message will aggregate individual biases leading to messages that are progressively more biased the further they travel from their source. In four experimental chapters, the social amplification of bias hypothesis was tested using the serial reproduction paradigm. In this paradigm, communication chains are formed using the responses of one participant (e.g., the recall of a text) as the stimuli to be presented to the next participant, thus recreating the social dynamic of receiving and transmitting information in the laboratory. These studies supported the social amplification of bias hypothesis, and did so covering different judgment and decision-making domains (risk perception, illusory correlations, denominator neglect, and cognitive reflection), using different types of response formats (frequency estimates, forced recognition, free recall), and including samples from Europe and the US, online and in the lab. The dissertation ends by discussing implications, future research, and potential modelling and debiasing techniques.

Keywords: Serial reproduction; social amplification; judgment and decision-making; heuristics and biases.

Resumo

A área de julgamento e tomada de decisão tornou-se muito influente com o programa de investigação das heurísticas e enviesamentos que revelou que as teorias económicas ignoravam aspetos importantes da realidade: o ser humano não é, como essas teorias presumiam, perfeitamente racional. Em vez disso, utiliza frequentemente estratégias simples, largamente automáticas e inconscientes (i.e., heurísticas) que permitem à mente humana, com as suas limitações, chegar a julgamentos e decisões aceitáveis, mas que também levam a erros sistemáticos (i.e., enviesamentos) em algumas condições. Apesar disso, a própria investigação em julgamento e tomada de decisão ignora ainda aspetos essenciais da realidade que podem ter uma influência importante na forma como formamos julgamentos e tomamos decisões. Entre estes aspetos encontra-se o facto de que as pessoas recebem informação de, e transmitem informação a, outras pessoas. A principal hipótese explorada na presente dissertação é a de que esta dinâmica social conduz à amplificação de enviesamentos: à medida que a informação viaja de uma pessoa para a outra, a mensagem vai agregando enviesamentos individuais, levando a mensagens progressivamente mais enviesadas quanto mais longe estiverem da sua fonte. Nos quatro capítulos experimentais desta dissertação, a hipótese da amplificação social de enviesamentos foi testada utilizando o paradigma da reprodução serial. Neste paradigma, constroem-se cadeias de comunicação ao utilizar a resposta de um participante (e.g., a sua recordação de um texto) como o material a ser apresentado ao participante seguinte, recriando assim, em laboratório, a dinâmica social de receber e transmitir informação. Os estudos apresentados apoiaram a hipótese de amplificação social de enviesamentos e fizeram-no cobrindo vários domínios diferentes de julgamento e tomada de decisão (perceção de risco, correlações ilusórias e reflexão cognitiva), usando diferentes formatos de resposta (estimativas de frequência, reconhecimento forçado, recordação livre), e utilizando dados de participantes da Europa e dos Estados Unidos, em laboratório e online. A dissertação termina discutindo implicações, investigação futura e potenciais técnicas de modelação e redução de enviesamento.

Palavras-chave: Reprodução serial; amplificação social; julgamento e tomada de decisão; heurísticas e enviesamentos.

Resumo Alargado

Uma parte significativa da investigação na área de julgamento e tomada de decisão tem sido feita sob uma abordagem que considera que as pessoas lidam com o mundo complexo em que vivem recorrendo a heurísticas (i.e., estratégias simples, que tendem a ser usadas de forma automática e inconsciente) que permitem à mente humana, com as suas limitações, resolver de forma satisfatória os problemas com que se depara, mas que por vezes levam também a enviesamentos (i.e., erros sistemáticos). Este foco nas heurísticas e enviesamentos permitiu que a área de julgamento e tomada de decisão tivesse uma influência muito grande dentro e fora da psicologia (e.g., na investigação em economia, direito e saúde), mas levou-a também a ser alvo de algumas críticas. Uma dessas críticas foca-se no facto de esta abordagem ter ignorando que as pessoas não vivem isoladas, mas sim em sociedade. Desde então, alguns estudos têm incluído contextos sociais e tarefas em grupo, revelando o impacto que este tipo de fatores tem no desempenho das pessoas. Ainda assim, há um aspeto social importante que tem sido negligenciado: nem sempre a forma como partilhamos informação com outros é simultânea (estando com eles em grupo), mas sim sequencial (recebendo informação de uma pessoa e transmitindo-a a outra).

A presente dissertação avança uma hipótese de amplificação social de enviesamentos onde, à medida que as pessoas transmitem informação de umas para outras, sequencialmente, elas vão adicionando os seus enviesamentos pessoais às mensagens que transmitem. Desta forma, à medida que os enviesamentos se vão acumulando, as mensagens ficam progressivamente mais enviesadas, tanto mais quanto mais distantes estiverem da sua origem. Esta hipótese pressupõe a existência de um enviesamento (i.e., uma tendência sistemática de distorcer o material num determinado sentido) único ou dominante na população, capaz de levar a distorções, filtragens ou adições sistemáticas de elementos da mensagem. A hipótese alternativa à amplificação social de enviesamentos é a hipótese de amplificação social de

ruído, onde as propriedades originais das mensagens vão também sendo distorcidas à medida que se afastam da sua origem, mas sem que haja algo de sistemático (i.e., um enviesamento) nessa distorção.

Para testar a amplificação social de enviesamentos, esta dissertação aplica o método de reprodução serial (Bartlett, 1932) onde um *output* da tarefa de um dado participante é utilizado como *input* da tarefa do participante seguinte (i.e., em vez de ler o problema na sua versão original, o participante lê o problema da forma que foi recordado por um participante anterior). A dissertação contém um total de quatro capítulos experimentais (II a V) que relatam estudos que utilizam este método e apoiam a hipótese de amplificação social de enviesamentos.

No Capítulo II, o método de reprodução serial é aplicado a uma área onde os efeitos de amplificação social já foram postulados: na perceção de risco (Kasperson et al., 1988). Estudos na área de perceção de risco revelam que as pessoas sobrestimam a frequência de causas de morte mais dramáticas (e.g., tornados ou ataques de tubarões) e subestimam a frequência de causas de morte menos dramáticas (e.g., doenças cardiovasculares ou cancro; Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978). A primeira experiência do Capítulo II testa se este enviesamento na estimativa de frequências pode ser socialmente amplificado e, em conjunto com a segunda experiência do capítulo, demonstra que o facto de algumas causas de morte serem imaginadas de forma mais ou menos vívida é um fator importante na explicação desta diferença na estimativa de frequências de causas mais e menos dramáticas.

No Capítulo III, testamos a hipótese da amplificação social com um efeito clássico da área de julgamento e tomada de decisão: as correlações ilusórias. Este capítulo apresenta duas experiências, uma onde as correlações emergem devido a fatores *bottom-up* (e.g., características do estímulo; Hamilton & Gifford, 1976) e outra onde as correlações emergem devido a factores *top-down* (e.g., guiadas pelas expectativas prévias dos participantes; Hamilton & Rose, 1980). Para além disso, a Experiência 2 deste capítulo testa se a amplificação social do enviesamento da memória de frequências (i.e., a correlação ilusória) pode levar a atitudes progressivamente mais polarizadas face a grupos. Mais ainda, a Experiência 2 compara também a técnica de reprodução serial com a técnica de reprodução repetida, - o método utilizado por Bartlett (1932) para estudar a memória individual ao longo do tempo, para testar qual destes dois fatores (social ou individual) leva a maiores níveis de enviesamento (na linha dos estudos de Roediger, Meade, Gallo, & Olson, 2014).

O Capítulo IV testa a hipótese de amplificação social de enviesamentos com outro enviesamento de julgamento: o enviesamento de rácios (e.g., Kirkpatrick & Epstein, 1992), também conhecido como negligência do denominador. Neste enviesamento, quando as pessoas são confrontadas com uma escolha entre duas opções, elas preferem opções com rácios com numeradores maiores (e.g., 9/100) do que opções com rácios com numeadores menores, mas probabilidades superiores (e.g., 1/10). Ao contrário dos dois capítulos anteriores, os participantes do Capítulo IV (e do V) não comunicaram as suas repostas a um problema, mas sim o problema em si. A experiência deste capítulo revela que 1) o enviesamento da negligência do denominador pode levar a distorções na maneira como as pessoas recordam problemas que descrevam duas escolhas, 2) estas distorções são socialmente amplificadas, e 3) estas distorções depois têm um impacto nas preferências que as pessoas têm face às opções descritas no problema.

O último capítulo experimental, o Capítulo V, estende o método de reprodução serial para ainda outro tipo de problemas, conhecidos conjuntamente como o Teste de Reflexão Cognitiva (Frederick, 2005). Nestes problemas, as pessoas tendem a dar uma resposta intuitiva que facilmente lhes surge na mente. Estudos anteriores (Mata, Ferreira, Voss, & Kollei, 2017; Mata, Schubert, & Ferreira, 2014) revelam que isto acontece porque as pessoas

não detetam um elemento crítico dos problemas e que leva à diferença entre a resposta intuitiva (mas incorreta) e a resposta correta. Com duas experiências, este capítulo testa a hipótese da amplificação social de enviesamentos, esperando que as pessoas filtrem o elemento crítico presente nos problemas, levando assim a um número progressivamente crescente de mensagens transmitidas que não contêm o elemento crítico, o que leva, por sua vez, a que as gerações futuras tenham uma progressiva erosão de deliberação.

No seu total, os quatro capítulos experimentais desta dissertação testaram, e apoiaram, a hipótese da amplificação social de enviesamentos. Fizeram-no variando vários aspetos, como diferentes domínios de julgamento e decisão (perceção de risco, correlações ilusórias, negligência do denominador e o Teste de Reflexão Cognitiva), utilizando diferentes formatos de resposta (estimativas de frequência, reconhecimento forçado, recordação livre), utilizando diferentes amostras de diferentes países (Portugal e EUA), quer em laboratório quer online, e desencadeando diferentes mecanismos de alteração das mensagens comunicadas (distorção, adição e filtragem). Estes estudos revelaram também que a amplificação social de enviesamentos pode ter efeitos indiretos em atitudes, escolhas e raciocínio subsequente.

A dissertação termina resumindo as principais implicações do conjunto de estudos empíricos apresentados, discutindo as suas limitações, considerando estudos futuros promissores que podem lidar com as limitações apresentadas, refletindo sobre a possibilidade de recuperar a mensagem original comunicada por uma população onde alguns dos seus elementos adicionam enviesamentos às mensagens comunicadas, sugerindo uma abordagem de modelação para guiar testes futuros à hipótese da amplificação social de enviesamentos e considerando, por fim, estratégias que podem ser úteis para atenuar a amplificação social de enviesamentos.

Table of Contents

List of Figures

List of Tables

1 – Chapter I – Introduction

In everyday life, people often use information that others provide them to form judgments and make decisions, such as what party to vote for, what to buy, or what to do when experiencing a given symptom. In addition to using this information, people also frequently transmit it to others, thus forming communication chains through which information travels. Being part of a communication chain allows people to leverage information obtained or synthetized by others, thus reducing the time and resources necessary to decide which option or course of action is better. However, if systematic biases distort information as it travels through the communication chain, then using this information may lead to worse outcomes than simply relying on first-hand knowledge. A large and influential research program in Psychology, known as the heuristics-and-biases research program, focuses on discovering, explaining, and predicting biases in human judgment and decisions. However, this literature has yet to consider this social dynamic of receiving and transmitting information. Doing so may shed light into the types of information that can lead to bias and the consequences of those biases in society.

1.1 – Heuristics and biases

Two central concepts of the judgment and decision-making literature are heuristics and biases. The importance of these concepts can be traced back to 1974, when Tversky and Kahneman published an impactful article that described three heuristics (representativeness, availability, and adjustment and anchoring) and 13 associated biases. While the literature (including Tversky and Kahneman's 1974 paper) often presents the concept of heuristics in a somewhat ill-defined manner, the common interpretation is that heuristics are simplified, mostly automatic and unconscious strategies that allow the human mind (with its limitations) to provide acceptable judgments and decisions, but that lead to biases in certain conditions (Keren & Teigen, 2004). Biases, in turn, are defined as systematic deviations between people's judgments and decisions and those prescribed by a given normative theory (e.g., Bayes' rule or regression towards the mean; Tversky & Kahneman, 1974). This definition of biases positions them as an effect or outcome of judgment and decision-making, but, as Keren and Teigen (2004) point out, some authors position biases as a cause or process that leads to systematic errors in a set of different tasks (e.g., Evans, 1989).

As a concrete example of the heuristics-and-biases approach, consider the phenomenon of illusory correlations, particularly as evidenced in the studies of Chapman and Chapman (1967). These authors set out to explain a "puzzling and distressing" (p. 193) problem in clinical psychology: that many psychologists report clinical observations that are not corroborated by research. For example, in the Draw-a-Person test, clinicians often report that patients with paranoid behavior usually produce drawings with more elaboration of the eye, an association which Chapman and Chapman's review shows not to be supported by research. These authors thus aimed to study this phenomenon in the laboratory and found that, indeed, naïve participants reported similar associations between certain symptoms (e.g., "suspiciousness") and traits with semantic associations (e.g., drawings with atypical eyes), despite no such association existing in the data that was presented to participants.

Tversky and Kahneman (1974) posited that illusory correlations were a side effect of the availability heuristic, whereby people use the ease with which instances come to mind to evaluate the frequency or probability of an event (e.g., thinking that a given neighborhood is particularly dangerous because one can easily call to mind occurrences of crime in that neighborhood). Tversky and Kahneman (1974) explained the illusory correlation as follows: Because the (semantic) association between symptom (e.g., atypical eyes) and pathology (e.g., suspiciousness) is strong, people confuse this semantic strength of association with the frequency of association of the stimuli in the material presented to them, thus reporting increased frequencies in line with the strength of the semantic association.

In the five decades since its inception, the heuristics-and-biases research program has greatly expanded. Many more than the three original heuristics and their 13 associated biases have been uncovered or integrated into this literature (a handbook by Baron in 2008 listed over 30 biases, an average of one new heuristic per two years of research), and it has also contributed to theories such as prospect theory (Kahneman & Tversky, 1979) and dual-process theories of judgment and decision-making (e.g., Epstein, 1994; Evans & Stanovich, 2013; Reyna, 1991).

Furthermore, the impact of this research has spread far and wide: Judgment and decision-making heuristics and their respective biases have been studied or considered in areas as diverse as medicine (e.g., Bornstein, Emler, 2001; Detmer, Fryback, & Gassner, 1978), law (e.g., Korobkin, 2011; Rachlinski, 2000), business (e.g., Busenitz & Barney, 1997; Ucbasaran, Westhead, Wright, & Flores, 2010), developmental psychology (e.g., J. E. Jacobs & Klaczynski, 2010; Strough, Karns, & Schlosnagle, 2011), psychopathology (e.g., Bennett & Corcoran, 2010; Huq, Garety, & Hemsley, 1988), and social psychology (e.g., Pyszczynski & Greenberg, 1987; Strack & Deutsch, 2004). This research program also led to the rise of behavioral economics (Laibson & Zeckhauser, 1998) and to two Nobel Prizes in Economic Sciences: Daniel Kahneman’s in 2002 (Kahneman, 2003) and Richard Thaler’s in 2017 (Thaler, 2017).

Despite this widespread impact, this research program has not been without criticism. For example, some scholars questioned the criteria used to classify judgments and decisions as biased (e.g., Cohen, 1979; Gigerenzer, 1991), accused the theoretical development of the heuristics-and-biases approach as being poor (e.g., vagueness in the definition of heuristics, lack of overarching theory; Gigerenzer, 1996; Wallsten, 1983), defended that people’s reasoning is well adapted to real-life outside the lab (e.g., Gigerenzer & Brighton, 2009; McKenzie, 2003), posited that participants in experiments, at least sometimes, answer in a non-normative manner because they interpret the task in ways different than those intended by the experimenters (e.g., Mandel, 2014; Schwarz, Strack, Hilton, & Naderer, 1991), criticized that this research frequently ignores the continuous nature of real-life judgment and decisions, relying on one-shot paradigms instead (e.g., Hogarth, 1981, 2005), and argued that human reasoning is not well accessed with the use of asocial, abstract, and context-free problems, as is sometimes done in this literature (Ortmann & Gigerenzer, 1997). In line with this last criticism, we now turn to an exploration of the way in which social factors, ubiquitous in everyday life, influence judgment and decision-making.

1.2 – Social factors in the study of heuristics and biases

Most of the heuristics-and-biases research has been carried out in a social vacuum. Typical studies give participants standardized, experimenter-selected materials, isolate participants from others' opinions, and individually measure participants' judgments and decisions in reaction to these materials. While this is an essential starting point to investigate human judgment and decision-making, this procedure removes social factors that are omnipresent in everyday life and that may have a significant impact in real world judgments and decisions.

One way social factors influence judgment and decision-making is by providing a context. Many evolutionary arguments posit that human cognition developed to deal with social situations, such as the enforcement of social contracts (Cosmides, 1985) or argumentation in social interactions (Mercier & Sperber, 2011), and so reasoning should improve when there is a functional social context to the reasoning problem. Indeed, some studies show that people reason better when problems are imbedded within certain social frames, for instance when trying to detect cheaters (e.g., Cosmides, 1989; Gigerenzer & Hug,

1992) or when anticipating the need to justify one's judgment or decision to others (i.e., accountability; see Lerner & Tetlock, 1999, for a review). As an example, Simonson and Staw (1992) studied interventions meant to reduce people's commitment to losing courses of action (also known as entrapment and the sunk cost fallacy or bias), one of which was accountability. The authors found that when participants were made accountable for the decision strategies that they used (and not the outcomes of their decisions), they allocated less resources to a failing product than the control group, thus reducing the escalation of commitment. Still, one should note that accountability does not always lead to better decision making or less bias. Important factors of the decision maker (e.g., whether the participant knows the algorithm required to arrive at the correct solution), the task (e.g., whether the problem explicitly mentions aspects that should be ignored), and the audience (e.g., perceived expertise level) moderate the effect (Lerner & Tetlock, 1999).

Another way in which adding a functional social context may influence judgment and decision-making is by triggering certain motivations. For example, assuming social roles such as playing the role of a lawyer (Ginossar & Trope, 1987; Mata, Ferreira, & Sherman, 2013) or interpreting data that contrasts one's favored in-group against a despised out-group (Mata, Ferreira, et al., 2013) leads people to reason strategically, using simpler or more sophisticated reasoning depending on what is necessary to reach the desired conclusion (e.g., that one's client is innocent or that one's in-group is better than the out-group).

Other than providing a context, people can also be influenced by others when they are exposed to their judgment and decisions (e.g., when observing the choices others make in a restaurant) or when receiving others' advice. Research on advice-taking has shown an effect known as egocentric discounting, in which people do follow others' advice, but less so than what would be beneficial, instead overweighting their own initial opinions (e.g., Bonaccio & Dalal, 2006; Yaniv & Kleinberger, 2000). This literature has also shown that people are sensitive to the reliability or expertise of the advice-giver (Bonaccio & Dalal, 2006; see Pornpitakpan, 2004, for a review of similar effects in the persuasion literature), an effect also documented when people are given others' solutions to problems of the heuristics-and-biases type (Ginossar & Trope, 1987). Furthermore, evidence shows that people who expect other people to be biased (i.e., the bias blind spot; Pronin, Gilovich, & Ross, 2002) can perform better by detecting others' biases and then avoiding those biases themselves (Mata, Fiedler, Ferreira, & Almeida, 2013). Thus, in general people are influenced by others' judgments and decisions, albeit less than may be optimal, and make adjustments or metacognitive inferences depending on perceptions of competence or bias.

Beyond simply receiving information from others, people also discuss information in groups. In one of the oldest examples of the combination of the heuristics-and-biases approach with group judgment and decision-making, Nagao, Tindale, Hinsz, and Davis (1985, as cited in Tindale, 1993) used the base rate fallacy to compare the performance of isolated individuals with individuals who formed groups to discuss and reach a consensus. A classical problem used to study the base rate fallacy is known as the cab problem:

> A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:
>
> (a) 85% of the cabs in the city are Green and 15% are Blue.
>
> (b) a witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.
>
> What is the probability that the cab involved in the accident was Blue rather than Green? (Tversky & Kahneman, 1982, p. 156)

The base rate fallacy occurs when people estimate the probability of the car being blue mostly as a function of the witness having identified the car as blue and in so doing neglect the base rates of blue cars and green cars, which in this case suggest that the car is probably green (Tversky & Kahneman, 1982). Nagao and collaborators (as cited in Tindale, 1993) compared the individual answers of participants who were never part of a group with others that were first part of a group tasked with reaching a consensus and then answered individually. The authors used five problems similar to the cab problem, manipulating the level of witness accuracy. They found that, in low levels of accuracy (50% or below), individuals committed more base rate fallacy than groups, while in high levels of accuracy (80%), groups committed more base rate fallacy than individuals. Thus, discussion in groups does not necessarily lead to higher or lower levels of bias.[1]

An important determinant of the outcomes of group discussion appears to be the demonstrability of the correct answer (Laughlin & Ellis, 1986; Trouche, Sander, & Mercier, 2014): When the correct solution to the problem is easy to explain (i.e., high demonstrability), then the correct response becomes dominant in groups (i.e., "truth-wins") as long as at least one group member knows the correct response. When that is not the case (i.e., low demonstrability), then the most frequent individual response becomes dominant (i.e., "majority-wins") and so performance depends on whether the majority's response aligns with the correct option.[2]

Under the demonstrability perspective, the results of Nagao and collaborators (as cited in Tindale, 1993) can be interpreted as follows: The base rate fallacy is a low demonstrability problem, as it is hard for group members to explain to others how to adequately combine the

[1] Interestingly, Kerr and Tindale (2004) point out that after group discussion individuals tend to regress towards their initial, individual opinions, by comparison with the judgment or decision reached as a group, revealing an effect similar to egocentric discounting.

[2] For example, if the studies described here had used tasks typical of fast-and-frugal heuristics research (e.g., D. G. Goldstein & Gigerenzer, 2002), group discussion could have increased adhesion to the heuristic, leading groups to outperform individuals.

witness' report with the prior probabilities (i.e., to explain the Bayesian solution). Thus, a majority-win outcome is observed and so groups exacerbate the bias when the witness has high accuracy (as then most members commit the base rate fallacy), and attenuate the bias when the witness has low accuracy (as then most members discard the unreliable witness report; Tindale, 1993).

Similar effects have been documented in Social Psychology. A classical effect in this field is group polarization of attitudes, which is also an example of an individual tendency becoming more extreme after group discussion. For example, in a study by Moscovici and Zavalloni (1969), French participants, measured individually before group discussion, held a somewhat negative attitude towards Americans and a somewhat positive attitude towards French General de Gaulle. After group discussion (and reaching a consensus, - which, in this research area, groups are usually required to reach), the attitude of the participants shifted, becoming more negative towards Americans and more positive towards de Gaulle (i.e., attitudes polarized).

Reviewing the group polarization research (in attitudes, but also jury decisions, ethical decisions, judgments of facts, person perception, negotiation, and risk taking), Myers and Lamm (1976) found three factors that lead to polarization: 1) a cognitive foundation that consists of people rehearsing personal arguments and listening to others' arguments, 2) a social motivation to perceive and present oneself in a positive light and to express and communicate socially desired arguments, and 3) a tendency to commit to one's publicly expressed opinions. These variables resonate with the concept of groupthink, which is a tendency, by some groups, to seek concurrence. Groupthink symptoms include overconfidence in the group's abilities, tunnel vision focused on the desired judgment or decision, and pressure to conform (Janis, 1991).

The social factors found in group polarization and groupthink research probably have less effect in high-demonstrability tasks. Indeed, Myers and Lamm (1976) posit that the interplay between cognitive and motivational factors should vary depending on the situation, with cognitive factors being more important in more fact-oriented judgment tasks and motivational factors more important in matters of social desirability. This would explain the contrast of polarization research with the results of Laughlin and Ellis (1986) and Trouche and collaborators (2014) of a significant change towards the correct response, - not polarization or groupthink.[3]

An infrequent type of outcome in group-research is “minority-wins”. When this happens, minorities, in the absence of a correct solution to the problem, still manage to sway a majority in the direction of their opinion. For example, a study by Smith, Tindale, and Steiner (1998) compared participants who were part of a group with participants who only decided individually on a sunk cost problem. The task was to decide whether to buy or not to buy a bankrupt competitor’s printing press that worked three times as fast as one’s current press for $10,000 after having just invested $200,000 on a new printing press that worked only twice as fast. Smith and collaborators (1998) found that groups of five with two members (i.e., 40%) who listed the past investment (i.e., the sunk cost) as a reason not to buy the competitors’ machine could win over the majority 56% of the time, although truth-wins effects were also observed if a minority argued in favor of the efficiency gains. Nonetheless, this study does show that it is possible for minorities to win over majorities if they make use of shared representations (e.g., the intuition that one should not abandon a given course of

[3] The groupthink classical case study of the Challenger disaster may be thought of as evidence that groups that exhibit groupthink will not lead to truth-wins in high-demonstrability scenarios. Yet, another possibility is that the Challenger disaster was not a high-demonstrability situation. One of the conditions for high demonstrability is that individuals in the group have sufficient understanding of the problem to accept the correct answer (Laughlin & Ellis, 1986). Yet, many involved in the Challenger disaster thought that the issue that led to the disaster would be protected against by a secondary, backup system (Janis, 1991), and so could not understand the gravity of the issue.

action once one has started), at least when the group does not have the conditions for high demonstrability (see also Kerr and Tindale, 2004).

If one considers demonstrability to be just the existence of shared representations (such as certain concepts of mathematics and logic) that allow one group member to convince others to support his or her view, then a more parsimonious summary of the group judgment and decision-making literature is that majorities win by default unless minorities can make use of arguments that gain leverage in representations shared by the majority.

While studies using groups are an important contribution to the exploration of social factors in judgment and decision-making, groups are not the only way people share information with one another. Another (arguably as frequent, if not more) form of information transmission is one-to-one communication, which often originates communication chains. In communication chains, information travels sequentially: For example, Sarah talks about her personal experience with a given product to John, who then talks about it to Hannah. In contrast with the group structure, in which all elements of the group receive the same information and must typically reach a consensus, in the communication chain structure, information travels from one person to the next and elements often decide individually. This change in the structure of information transmission may have important consequences for bias formation and propagation, as we will see next.

1.3 – Social dynamics of information transmission

In a seminal book entitled *Remembering: A study in experimental and social psychology*, Bartlett (1932) studied the impact of others on memory by using a method he named serial reproduction. In this method, participants are exposed to stimuli, such as a picture or a text, and are later asked to reproduce the material, typically through free recall. This reproduced material is then provided to another participant as the stimuli to be recalled. Thus, this procedure mimics one of the major social dynamics of society in which information travels from one person to another. Bartlett used the serial reproduction method to complement the repeated reproduction method, in which the same participant recalled the same material over time, and which served to study individual memory processes.

Using the serial reproduction method, Bartlett's (1932) main findings with verbal material were that the texts tended to 1) become more concrete (e.g., through the omission of general opinions or arguments), 2) lose individual characteristics (e.g., specific writing style, names of characters or places), 3) become significantly shorter, and 4) become rationalized, as a result of an effort after meaning, often through the transformation of the unfamiliar into the conventional (e.g., interpreting and reproducing something black coming out of a character's mouth as his dying breath or his soul leaving his body). These main results were in line with what Bartlett found with individual participants recalling the material multiple times (i.e., repeated reproduction), but he noted that distortion and omissions were stronger in serial reproduction (an issue that we shall revisit in Chapter III).

While some characteristics of Bartlett's research would appear quaint to many modern experimental psychologists (e.g., he made no use of statistical inference and freely varied his procedure during the course of the experiment if he considered that it would enable him to better explore his hypotheses[4]), his ideas had a profound influence in psychology, particularly cognitive psychology (Roediger et al., 2014). Before Bartlett, memory was thought of as a storage of traces left behind by stimuli, and remembering was seen as simply the process of attempting to retrieve these traces (Roediger, 1997). After Bartlett, and the researchers who picked up and expanded on his ideas (Roediger, 1997), memory came to be thought of as a reconstructive process in which people's previous knowledge structures (i.e., schemata) help guide what is remembered (Roediger et al., 2014).

[4] Bartlett (1932) defends his approach in the first chapter of the book. For a recent defense of his approach, see Wagoner (2015).

As for the serial reproduction method, it first experienced a period of expansion: Some studies further explored some of the themes set out by Bartlett (1932), such as the serial reproduction of images (Ward, 1949), the role of motivation in remembering (Alper & Korchin, 1952), individual differences (Paul, 1959), or the impact, on remembering, of belonging to different cultures (Talland, 1956) or social groups (Maxwell, 1936). Other studies used the method for new purposes, such as to study rumor (Allport & Postman, 1947) or children's remembering (Northway, 1936).

Yet, after this initial period of expansion, a study by Gauld and Stephenson (1967) aimed to show that Bartlett's results were mostly a product of Bartlett's participants not trying their best to accurately recall the material. These authors varied the extent to which the instructions promoted accuracy, correlated participants' confidence with their accuracy in remembering, and tested differences in performance between people high and low in conscientiousness (with the rationale that high conscientious people follow experimental instructions more closely). They then concluded that Bartlett's findings were probably a result of his participants not trying their best to remember, as 1) accuracy instructions lowered memory errors,[5] 2) participants' confidence in the elements of the story correlated with their accuracy, and 3) conscientious participants committed fewer errors.

Gauld and Stephenson's (1967) study has been identified as the cause leading the serial reproduction method to become dormant until the second millennium (Wagoner, 2015), at which point it started experiencing a revival. In this new millennium, Bartlett's (1932) main findings were successfully replicated using inferential statistics and modern notions of experimental rigor (Roediger et al., 2014), and the serial reproduction method spread to many new areas, such as stereotypes (e.g., Kashima, 2000; Thompson, Judd, & Park, 2000), social

[5] As Wagoner (2015) notes, the criteria for measuring performance used by Gauld and Stephenson (1967) are a significant flaw of the set of studies. For example, omissions and conventionalization distortions (e.g., "boat" instead of "canoe") were not used to penalize performance.

representations (Bangerter, 2000), counter-intuitive concepts (Barrett & Nyhof, 2001; Claidière, Trouche, and Mercier (2017), intergroup conflict (Lee, Gelfand, & Kashima, 2014), risk perception (Jagiello & Hills, 2018; Moussaïd, Brighton, & Gaissmaier, 2015), evolution of language (e.g., Kalish, Griffiths, & Lewandowsky, 2007; Reali & Griffiths, 2009), sociology (Hunzaker, 2014, 2016), evolutionary psychology (e.g., Bebbington, MacLeod, Ellison, & Fay, 2017; Mesoudi, Whiten, & Dunbar, 2006), and cultural evolution (e.g., Mesoudi & Whiten, 2004; Mesoudi et al., 2006). The method is known in some of these new fields by other names, such as iterated learning, transmission chains, or diffusion chains.

Despite the variety of applications, many serial reproduction studies reveal a convergence of the transmitted material towards some sort of shared cognition, for example, schemata (Bartlett, 1932), stereotypes (Kashima, 2000), social representations (Bangerter, 2000), inductive biases (Kalish et al., 2007), and evolutionarily-selected biases (Mesoudi et al., 2006). This result is in line with the majority-wins outcomes of research using small groups, in which (when the minority cannot make use of shared representations to persuade the majority) the opinion of the majority determines the judgment of the group.

While less frequent in the literature, other studies suggest outcomes beyond majority-wins in serial reproduction. A first example reveals a truth-wins effect similar to those found in group research. Claidière and collaborators (2017) used a variety of methods (a modified serial reproduction, small groups, and simulations) to investigate how counter-intuitive concepts may spread in populations. They found that argumentation (in contrast with prestige and conformity) led to increased performance and the spread of counter-intuitive solutions. Furthermore, they found that the probability of these truth-wins were influenced by two factors: 1) the efficiency of arguments in leading people to adopt and comprehend the reasoning behind the correct answer (i.e., transmission effectiveness), and 2) the quantity of opinions that people were exposed to, as when people are exposed to more opinions, then lower levels of transmission effectiveness are permissible, because the probability of being exposed to a convincing argument increases.

In another example of a majority-wins exception in serial reproduction, a study by Navarro, Perfors, Kary, Brown, and Donkin (2018) revealed a minority-wins outcome. These authors, using simulations and an empirical study, demonstrated that in serial reproduction with varied biases (i.e., heterogeneous populations) there is no guarantee that the bias observed in the population over time is the distribution of biases in the population or the average of these biases. Instead, those with more extreme views (especially when no opposing extremists exist) tend to have more weight in the outcome than would be expected by a weighted population average (including a weighted average that takes into account not only the frequency, but also the strength of beliefs). This happens because those with weak biases are more easily swayed by those with strong biases than vice-versa in a way that makes it difficult to precisely predict the biases that the population will reveal (Navarro et al., 2018).

The results of Navarro and collaborators (2018), - a minority-wins outcome, may appear inconsistent with those of Claidière and collaborators (2017), - a truth-wins outcome, but are probably the result of two factors: 1) the participants of Navarro and collaborators' (2018) study communicated only answers, not arguments in support of answers (as in Claidière et al., 2017), and so argumentation-effects were precluded; and 2) the task used by Navarro and collaborators (2018) had participants predict who would win an Australian Federal election, which is a (low-demonstrability) task in which persuasive arguments are probably difficult to come by (while Claidière et al., 2017, used problems high in demonstrability; see Trouche et al., 2014).

Regardless, the study by Navarro and collaborators (2018) is particularly interesting because it shows how serial reproduction studies and group research are not redundant: A minority was able to sway a majority without making use of strong, valid arguments or of any shared representation. Instead, the population heterogeneity and extremity itself led to this effect, while group composition (which is the group-level equivalent) has not been documented as having a similar impact. Probably this happened due to differences in the structures of the two methods: Majority and minority members are well aware of their status in groups, but minority and majority members in serial reproduction only know the opinion of the person immediately preceding them.

In fact, one may think of serial reproduction and small group methodologies as atomic elements of a broader taxonomy of information diffusion paradigms (based on the list of paradigms by Mesoudi, 2007). Serial reproduction (see Figure 1, panel A) would then be unidirectional one-to-one transmission of information (Sarah tells John, who in turn tells Hannah about the event), while groups (Figure 1, panel C) represent simultaneous transmission of information with the possibility of feedback (Sarah, John, and Hannah together discuss the event). Some mixed methodologies may be thought of, namely, 1) the replacement method (e.g., R. C. Jacobs & Campbell, 1961; see Figure 1, panel B1) in which the composition of the group changes over time, such that older members only impact newer generations through their influence on intermediate generations, and 2) networks (Figure 1, panel B2) in which communication in the population may, at the same time, include portions that act as groups and portions that act as transmission chains.

As such, research into the basic paradigms of information diffusion (i.e., group research and serial reproduction) may shed light on more complex structures. The work presented in this dissertation focuses on the less researched of these two paradigms in the judgment and decision-making literature - serial reproduction - with the goal of exploring the dynamics of heuristics and biases when information is transmitted from one person to the next.

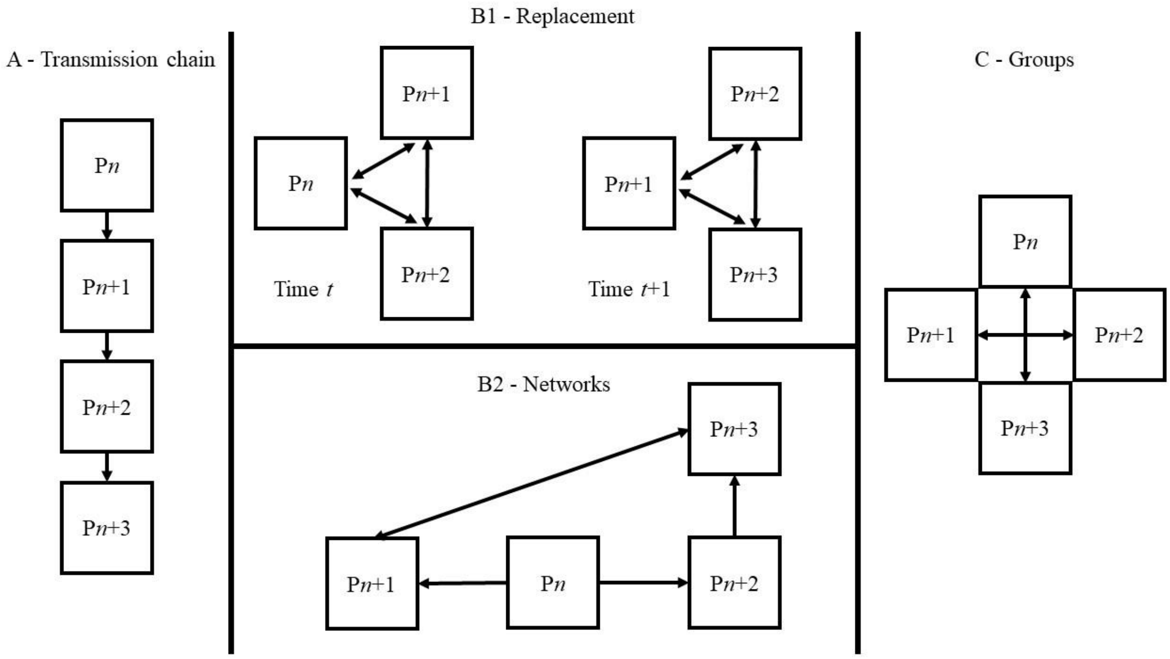


Figure 1. Some examples of information diffusion paradigms (P = participant, *n* = number, *t* = time).

1.4 – The social amplification of bias

The key hypothesis presented in this dissertation is that, as information travels in society from one person to the next, bias will aggregate in a systematic fashion in the message, such that what starts out as a small bias can grow larger and more consequential. We call this process the social amplification of bias: As each person in the communication chain adds his or her own bias to the message received, the message will become more biased the further away it is from the original source of the message.

To conceptualize the social amplification of bias hypothesis, it is useful to consider a few basic concepts of communication theory. A classical model in this research field (Shoemaker, 1987) is Shannon and Weaver's (1949) model of communication. In this model, an information source (e.g., a person's brain) selects a given message, then uses a transmitter (e.g., a person's vocal chords) to transform this message into a signal (e.g., speech) that then travels through a communication channel (e.g., air) to the receiver (e.g., another person’s auditory system), and finally to its destination (e.g., another person’s brain). Furthermore, noise in many steps of the process can alter the information. Thus, one may express the state of the message at the destination in the simplified first equation:

$$\text{Message at destination} = \text{Original information} + \text{Noise}.^{6} \qquad (1)$$

For example, if we had Sarah providing John an estimate of the distance, by road, between Lisbon and Madrid, and Sarah tries to estimate the real number (629km), but ends up overestimating (e.g., +21km), the message John receives is 650km. To expand this idea to the situation of a message being transmitted over a chain of people, it is enough to transform the noise parameter into a vector that reflects the impact of various sources of noise that have aggregated in the message up to a particular point in time (*t*), as seen in the second equation:

$$\text{Message at destination}_t = \text{Original information} + \text{Noise}_t. \qquad (2)$$

So, continuing the example, John communicates the message to Hannah, and in so doing adds more error (e.g., -25km), such that when the message reaches Hannah it is equal to 629 + (21 – 25) = 625km. Thus, the message at time *t* is equal to the value or elements of the original message, plus the combination of all noise aggregated up to time *t*. In contrast, the social amplification of bias hypothesis can be expressed by the third equation:

$$\text{Message at destination}_t = \text{Original information} + \text{Bias}_t + \text{Noise}_t. \qquad (3)$$

Thus, the message may accumulate not only non-systematic distortion (i.e., noise), but also systematic distortion (i.e., bias; see Yaniv, 2004b, for a non-vectorial version of this equation in research into how people aggregate others' opinions). Using the same example, if both Sarah and John have a bias to overestimate distances (e.g., by adding 10km), then the message that reaches Hannah is equal to 629 + (10 + 10) + (21 – 25) = 645km.

In general terms, then, a social amplification of bias is found whenever the bias term of the equation above is significantly different from zero. In the studies presented in this

[6] The mathematical theory of communication is, of course, more complex than described here: information is measured in a logarithmic scale, the process of composing a message is described as a Markoff process, noise causes entropy in the message to increase over time, etc. The equation presented here was simplified for the purposes of this dissertation.

dissertation, we expect the direction of bias to be in line with the evidence provided by the heuristics-and-biases literature using single individuals. For example, if people were to communicate base rate problems, we would expect distortion in the message to favor or preserve individuating information (e.g., the witness' report) in detriment of the probabilistic information (e.g., the base rate of blue and green cars), in line with the base-rate neglect bias.

The two main conditions for the social amplification of bias hypothesis are:

1. The message can suffer qualitative (i.e., filtering or adding elements) or quantitative (i.e., attenuation or intensification of elements) distortions. In the case of qualitative distortions, while *the messages* being communicated may not become more biased as generations progress (e.g., a dichotomous variable measuring distortion in the message is only either distorted or intact, so it cannot become more biased as time progresses), the bias *in the population* should increase as generations progress and more versions of the distorted message are present in the population. In the case of quantitative distortions, the bias both in the message itself and in the population may increase over time.[7]
2. There is either a) only one bias present in the population or b) the interplay between different biases results in one dominating over the others. As the study by Navarro et al. (2018) reveals, the conditions for dominance in the presence of many biases may be difficult to establish a priori, as minorities can have a stronger impact than is warranted by both their quantity and the extremity of their biases. In the present work we will assume, based on the previous research, that each problem leads to a dominant or single bias in the population.

[7] This criterion may sound similar to Allport and Postman's (1947) ambiguity variable in rumor diffusion (of the equation: rumor = importance of the rumor to the person x ambiguity of the facts associated with the rumor). But, as Chapter IV reveals, ambiguity in general is not a requirement of the social amplification of bias as even short (2-3 sentences) texts recalled in short (less than 5 minutes) intervals can lead to a social amplification of bias. Ambiguity may boost biases in some conditions, but it is, thus, not a necessary condition for their social amplification.

Given the above, the social amplification of bias hypothesis is formulated in opposition to a social amplification of noise hypothesis, in which either no bias is present, or different biases cancel each other out. Whether one observes amplification of noise or of bias has significant practical implications, chief among them the possibility to recover or approximate the original information. For example, two techniques, one based on abundance distributions of imperfect transmission and another based on the wisdom of the crowd, should perform well under a social amplification of noise, but should underperform under a social amplification of bias, as we will see next.

First, one may assume the distribution of variants of a message follows an abundance distribution (a “lazy-J-curve”), in which one value is by far the most frequent (the high point of the J), while all other values have significantly lower frequencies (the lower points of the J, see Figure 2, panel A1). Because the correct, original information has always the same value, but noise can assume many values, when looking at a distribution of reproductions, the most frequent result should be the original information, not noise. Dunn and Kirsner (2011) used this logic in a successful attempt to recover the lost World War II ship HMAS Sydney II. Indeed, the most common informational variant in the population of witnesses (enemy ship survivors) was found to be only 5km from the actual location, by comparison with the presumable best witness (the captain of the enemy ship), who missed the correct location by 50km. While Dunn and Kirsner's (2011) approach should work under a social amplification of noise, it should not work under a social amplification of bias, depending on the proportion of biased witnesses. For example, if there is the same amount of unbiased and biased witnesses, it is not possible to distinguish bias from the original information (see Figure 2, panel A2).

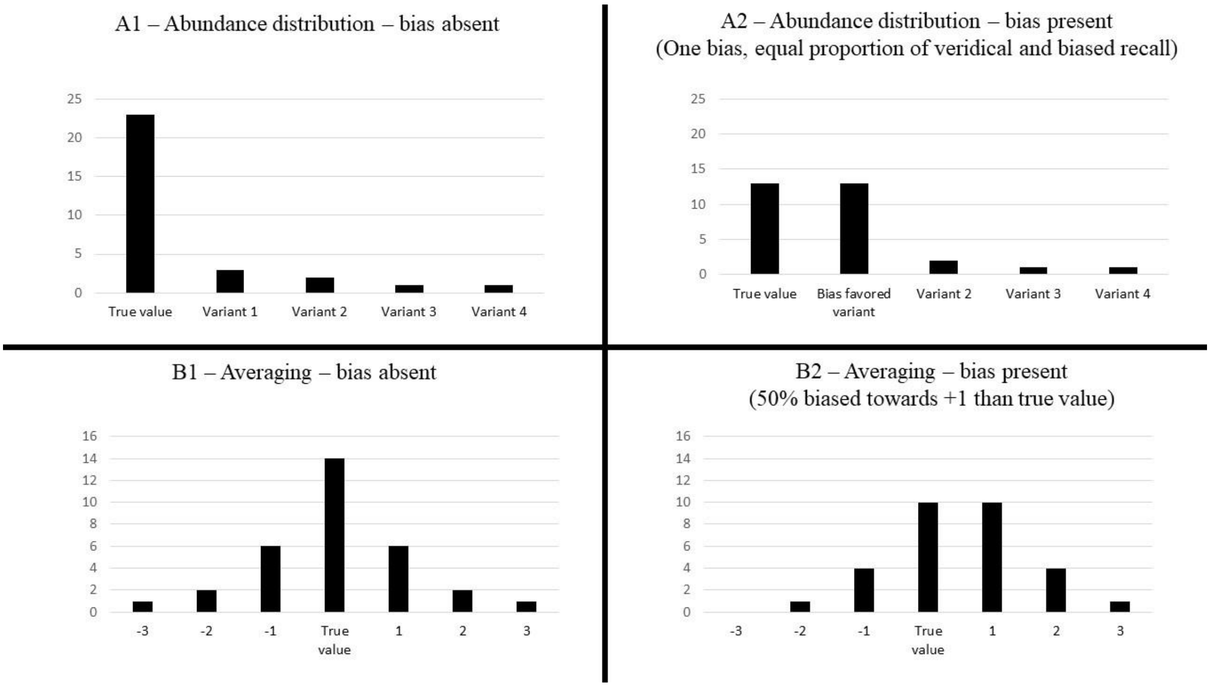


Figure 2. Conceptual examples of noise reduction technics in the absence (A1 and B1) and presence (A2 and B2) of populational bias.

Second, one may use wisdom of the crowd techniques, such as averaging independent estimates of various people, to cancel out the random variation present in each estimate (Galton, 1907; Surowiecki, 2004; see Figure 2, panel B1) and thus recover the actual value or to obtain a good approximation of it. While the message that is transmitted within a chain includes dependence (what a given person remembers is dependent on what the previous person of the chain remembered), there is no between-chains communication and, so, no dependence. As such, averaging the results of different chains under a social amplification of noise (see Equation 2) should allow the different sources of noise in each chain to cancel each other out, leading to an estimate that is very close to the original material, or, even, that performs better than most chains (depending on the amount of error present in the estimates). Averaging should not work under a social amplification of bias because, while the chains are still in practice independent (i.e., they cannot influence each other), the dominant bias present in the chains pushes estimates away from its original value in a systematic direction (see Figure 2, panel B2).

Thus, a social amplification of noise allows the application of certain techniques that improve performance or even recover the original message, while the social amplification of bias precludes the benefits of these techniques, without the application of some correction for bias. It is the social amplification of noise hypothesis that will serve as the null hypothesis in all experiments presented in the current dissertation.

1.5 – Overview

The main body of the current dissertation consists of four empirical chapters in the form of research articles (Chapters II–V). Chapter II starts by showing how the method of serial reproduction can be used to study the social amplification of bias in an area of research in which social amplification effects have already been theorized: risk perception (Kasperson et al., 1988).[8] Research has shown that people overestimate the frequency of dramatic causes of death (e.g., shark attacks) and underestimate the frequency of ordinary causes of death (e.g., cancer; Lichtenstein et al., 1978). The first experiment in Chapter II tests whether this frequency estimation bias can be socially amplified, and, together with Experiment 2, demonstrates that the vividness of the causes of death is an important factor behind this dramatic versus ordinary difference.

In Chapter III, we extend the serial reproduction paradigm to a traditional judgment and decision-making phenomenon by testing the social amplification hypothesis with illusory correlations. This chapter presents two experiments, one in which people report frequencies that reveal illusory correlations due to bottom-up processes (e.g., characteristics of the stimuli; Hamilton & Gifford, 1976), and one in which they emerge due to top-down processes (e.g., guided by participants' expectations; Hamilton & Rose, 1980). Experiment 2 also evaluates whether the social amplification of biased frequency recall (i.e., illusory

[8] At the time the studies in Chapter II were completed, we were unaware of the studies by Moussaïd et al. (2015) and Jagiello and Hills (2018). These studies reveal results convergent with the social amplification of bias and Chapter II mentions the differences between these studies and the serial reproduction experiment of Chapter II.

correlation) can lead to increasingly polarized attitudes towards groups, and in addition it compares serial reproduction with serial repetition, - the method used by Bartlett (1932) to investigate individual memory over time, to test which of these factors (social or individual) leads to greater bias (following Roediger et al., 2014).

Chapter IV then expands the social amplification of bias to another judgmental bias: the ratio bias (e.g., Kirkpatrick & Epstein, 1992). In this bias, people prefer options with larger numerators (e.g., 9/100) over options with smaller numerators, but greater probability (e.g., 1/10). Unlike the two preceding chapters, participants in Chapter IV (and V) did not communicate their answers to a problem, but instead communicated the problem itself. The experiment in this chapter shows 1) that the bias can distort people's recall of vignettes describing the two choices, 2) that these distortions in recall are socially amplified, and 3) that these distortions then impact people's preferences towards the situations described in the vignettes.

The last empirical chapter, Chapter V, expands the method to yet another type of problems, those known as the Cognitive Reflection Test (Frederick, 2005). In these problems, people usually answer in line with a strong intuitive answer that is incorrect. Previous studies (Mata et al., 2017, 2014) have revealed that this happens because people fail to detect a critical verbal element in the problems that leads to a difference between the (incorrect) intuitive answer and the correct one. In two experiments, a social amplification of bias hypothesis is tested, such that we expected people to filter the critical element present in the problems, thus leading the following generations to a progressive erosion of deliberation.

Finally, Chapter VI summarizes the contributions of the empirical chapters, discusses their limitations, considers promising future extensions of this work, offers a modeling approach that might prove useful to guide future research, and considers possibilities for debiasing.

2 – Chapter II – The social amplification of risk: Why dramatic risks are more likely to be remembered and communicated than ordinary ones. Mata, A., Mendonça, C., Schott, M., Ihmels, M., & Simão, C. (*submitted*)

In their seminal work, Lichtenstein and collaborators (1978) showed that people tend to overestimate the number of deaths related to dramatic, non-natural causes of death (e.g., homicides, tornadoes, accidents), while they underestimate the number of deaths associated with ordinary, natural causes of death (e.g., diabetes, asthma, stroke). They attributed this to the greater availability in memory of the dramatic events (Tversky & Kahneman, 1973). Moreover, dramatic causes of death are overrepresented in the media (Combs & Slovic, 1979; Frost, Frank, & Maibach, 1997). Thus, people fear more, and are more exposed to, information about dramatic vs. ordinary risks, despite the fact that the latter are much more deadly than the former (Sunstein & Zeckhauser, 2011). This paper provides an account for this phenomenon, based on intra-psychic biases of memory and interpersonal mechanisms of communication.

This account starts with the assumption that information about dramatic risks is more vivid and distinctive than information about ordinary risks, where vividness is operationalized as ease to imagine in a concrete, detail-rich manner, and distinctiveness is operationalized as the degree to which information is represented in a differentiated and more unique way. For instance, imagine someone dying in a car crash, flood, homicide, etc. Presumably it is easier to form a clear and distinct mental image of someone dying from those causes than it is to form a vivid and distinct mental picture of someone dying from cancer, hepatitis, diabetes, etc.

The second hypothesis is that, because dramatic risks are more vivid and distinctive, they are more memorable. Indeed, a large body of research shows that vividness leads to better recall (Collins, Taylor, Wood, & Thompson, 1988; Nisbett & Ross, 1980; Reyes, Thompson, & Bower, 1980; Shedler & Manis, 1986; S. M. Smith & Shaffer, 2000), and the same holds for distinctiveness (Eysenck & Eysenck, 1980; Hunt & Mitchell, 1982; Rawson & Van Overschelde, 2008; Schmidt & Saari, 2007; Von Restorff, 1933).

Lichtenstein et al. (1978) only speculated on the link between vividness and recall in thinking about risks, but they tested neither vividness nor recall. The present studies do so. Moreover, Shanteau (1978) noted that Lichtenstein et al.'s research had the limitation that participants were not exposed to the stimuli whose frequency they had to estimate (i.e., participants had to estimate how many people died of a certain cause of death, but they did not receive actual information about it), which he argued made it difficult to trace the observed biases in frequency estimation to psychological mechanisms. We manipulated the frequency of the different types of deaths and tested how sensitive participants are to this information. To the best of our knowledge, no study has manipulated the frequency of different causes of death and tested how accurate people are in keeping track of those numbers.

The third hypothesis in our account pertains to social amplification: This research tests whether the hypothesized memory advantage of information about dramatic vs. ordinary risks is amplified as this information is communicated through a chain of people. As this memory bias (i.e., that information about dramatic risks is more likely to be recalled than information about ordinary risks) accumulates across a communication chain, the information about dramatic risks is more likely to be preserved and passed on, whereas the information about dramatic risks is more likely to be forgotten and filtered out. This hypothesis is also consistent with research showing that arousal (which is presumably greater when thinking about dramatic risks) increases social transmission of information (Berger, 2011).

To explore this social amplification hypothesis, our research adapts the serial reproduction paradigm (Bartlett, 1932; Kashima, 2000), which organizes participants in communication chains where they pass on information from one person to another. Thus, whatever memory bias is built into the memory of one participant influences the information that the next participant receives. For instance, imagine that a first participant sees information about several causes of death, including dramatic and ordinary ones. Later on, when asked to recall that information, if this participant only remembers information about the dramatic causes of death and not the ordinary ones, then that is the information that the next participant will receive.

The few existing studies trying to capture the social amplification of risk have either tracked people's attitudes about a certain risk (e.g., genetically modified food; Frewer, Miles, & Marsh, 2002), or examined how messages about a certain risk get transmitted from person to person (Jagiello & Hills, 2018; Moussaïd et al., 2015). However, people often learn about several different risks, and not just one. When they do so, do they selectively remember the dramatic ones, and forget the ordinary ones? That is, from the universe of potential causes of death that people learn about, are the ones that people remember and talk about those that are most dramatic? This research tests a new dimension of the social amplification of risk framework: not just how people come to collectively represent a certain risk as they share information about it, but how different types of risk can be asymmetrically amplified as people remember some more than others, and give them greater emphasis in the information that they convey to others.

The outline of the studies is as follows: A first pre-test assessed whether the vividness of dramatic causes of death is greater than that of ordinary ones. Experiment 1 tested the recall advantage of dramatic versus ordinary causes of death, and whether that advantage gets amplified in the process of social communication. Experiment 2 further explored the cognitive basis of the memory advantage of dramatic versus ordinary risks (by assessing both vividness and distinctiveness), and whether that advantage generalizes across processing goals (memory versus communication).

2.1 – Pretest: Vividness

2.1.1 – Method

Participants. Twenty-five undergraduates from the University of Heidelberg participated in the pretest.

Procedure. Participants were asked to rate the vividness of 6 dramatic causes of death and 6 ordinary causes of death (taken from national statistics of Germany and the US; Hertwig, Pachur, & Kurzenhäuser, 2005; Hoyert & Xu, 2012). The dramatic causes of death were: tornado; car crash; flood; fire; homicide; accidental fall. The ordinary causes of death were: asthma; stomach cancer; tuberculosis; hepatitis; diabetes; leukemia.

For each of these, participants were asked: 1) "How easy is it for you to form a concrete mental image of someone dying from X?" (1 – *very hard*; 9 – *very easy*); and 2) "When you imagine someone dying from X, can you picture the event clearly and vividly in your mind, as if it was a photo or a movie?" (1 – *not at all*; 9 – *totally*).

2.1.2 – Results

On both items, dramatic causes of death were considered more vivid than ordinary ones: 1) $M = 6.57$, $SD = 1.37$ vs. $M = 4.13$, $SD = 1.30$, $t(24) = 6.95$, $p < .001$; 2) $M = 7.12$, $SD = 1.49$ vs. $M = 3.83$, $SD = 1.18$, $t(24) = 8.50$, $p < .001$.

2.2 – Experiment 1: Serial Reproduction

2.2.1 – Method

Participants. Fifty participants were recruited at the University of Heidelberg. These participants were grouped into 10 five-person chains. This sample size is comparable to that used in previous studies on serial reproduction (e.g., Kashima, 2000, had 12 five-person chains; Lyons & Kashima, 2003, had 8 four-person chains).

Procedure. Participants were presented information about how several people died in a certain town. Information about the name and cause of death of each deceased person was presented one-by-one, in random order (e.g., Mrs. Goldstone died of asthma[9]). Half of the cases pertained to ordinary, natural causes of death; the other half to dramatic, non-natural causes of death. The frequencies of the different causes of death varied between 1 and 7. Specifically, the dramatic causes of death and respective frequencies were: tornado (1), car crash (1), flood (4), fire (4), homicide (7), and accidental fall (7). The ordinary causes of death were: asthma (1), stomach cancer (1), tuberculosis (4), hepatitis (4), diabetes (7), and leukemia (7). Both type of cause of death (ordinary vs. dramatic) and frequency were manipulated within-subjects.

After seeing all the instances, participants were asked to indicate all the different causes of death that they could recall, and to estimate the frequency of people who had died of each of them (these responses were open-ended, not selected from a list).

This procedure was repeated across five generations of participants in a serial reproduction paradigm, whereby the recall and frequency estimates made by a participant in a certain generation determined the information that a participant in the next generation received. For instance, if a participant in generation 1 recalled tornado but not asthma, and estimated having seen 2 cases of deaths by tornado (instead of the 1 that was actually presented), then a subsequent participant in generation 2 would be presented 2 cases of tornado-related deaths and no case of an asthma-related death.

Thus, the design was a 2 (type of causes of death: dramatic vs. ordinary) x 3 (frequency in the first generation: 1, 4, or 7) x 5 (generation), with the first two factors

[9] This was not one of the names used in the experiment. Typical German names were used.

manipulated within-subjects, and the third one between-subjects.

2.2.2 – Results

An ANOVA assessing the proportion of recalled items for the 2 types of death by the 3 frequency levels over the 5 generations revealed: a main effect of type of death, such that dramatic causes of death were better recalled than ordinary ones, $F(1, 45) = 53.46$, $p < .001$, $\eta_p^2 = .54$; a main effect of frequency, such that more frequent causes of death were better recalled than less frequent ones, $F(2, 44) = 59.74$, $p < .001$, $\eta_p^2 = .57$; a main effect of generation, such that recall scores were higher for earlier generations than later ones, $F(4, 45) = 22.04$, $p < .001$, $\eta_p^2 = .66$; an interaction of type of death by frequency, such that the difference in recalling dramatic versus ordinary causes of death was significant for higher frequencies (for $n = 4$ and $n = 7$, ps $\leq .008$) but not for the lowest frequency (for $n = 1$, $p = .128$), $F(2, 90) = 7.75$, $p = .001$, $\eta_p^2 = .15$; and more importantly, the predicted interaction effect of type of death by generation, such that the difference in recalling dramatic versus ordinary causes of death increased over generations, $F(4, 45) = 3.09$, $p = .025$, $\eta_p^2 = .22$; other Fs < 1.

Another way to quantify this interaction (one that respects the ordinal value of the generation variable) is to compute the difference between the probability of recalling dramatic versus ordinary risks, and correlate that with generation. This correlational analysis shows that the advantage of dramatic over ordinary risks increases across generations: $r = .37$, $p = .009$. In yet another analysis, comparing the proportion of recall for the different types of risk within each generation, the difference in recall for dramatic versus ordinary risks is not significant in the first generation ($p = .193$), it grows to marginally significant levels in the second generation ($p = .066$), and it becomes significant from the third generation on (ps $\leq .005$). The mean recall values per generation and type of risk are shown in Figure 3.

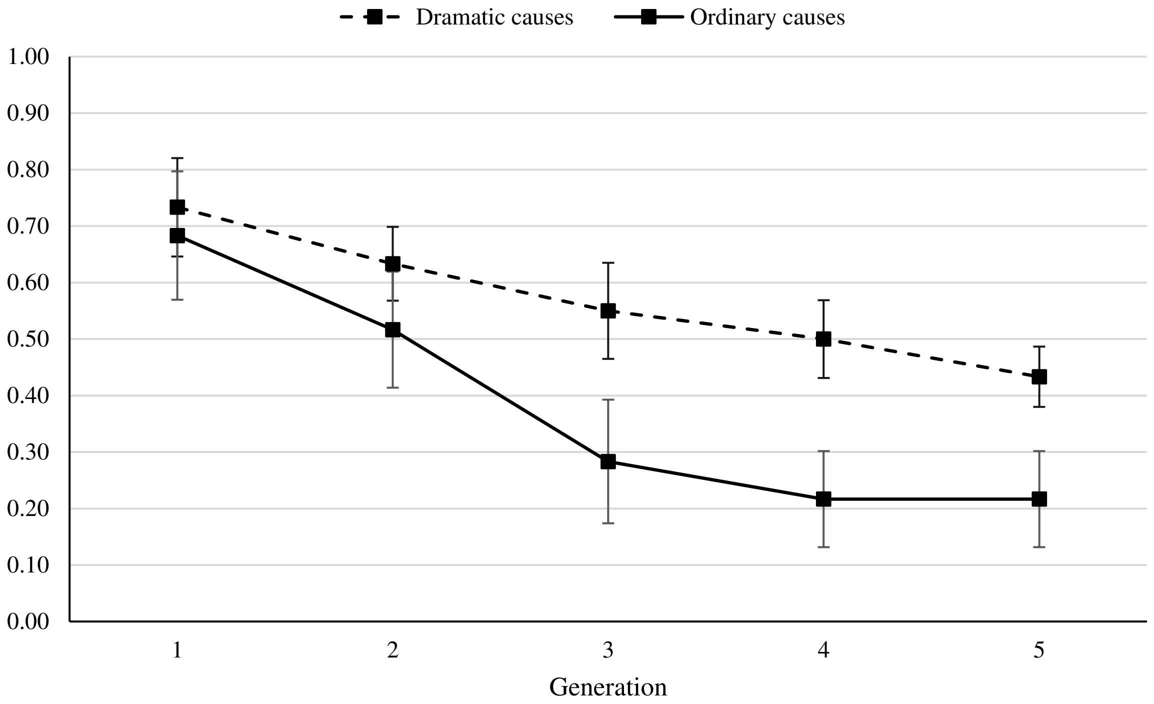


Figure 3. Mean probability of recall, per generation and type of causes of death (error bars represent 95% between-subjects CIs).

Some participants misrecalled new causes of death that they had not seen. Some of these were transformations of causes that they had seen (e.g., storm, suicide, fall in stairway), others were totally new ones (e.g., poisoning). The probability of misrecalling new causes of death was larger for dramatic vs. ordinary risks: respectively, $M = 0.28$, $SD = 0.45$ vs. $M = 0.08$, $SD = 0.27$, $t(49) = 2.65$, $p = .011$.

While the main hypothesis pertains to recall, an analysis similar to that which was done for recall data was performed for frequency estimates, after excluding two extreme outliers. There was a significant effect of frequency ($F(2, 86) = 66.09$, $p < .001$, $\eta_p^2 = .61$), such that high frequency deaths were estimated as higher than low frequency deaths ($n = 7 > 4 > 1$, $ps \leq .006$), and a significant effect of type of death ($F(1, 43) = 4.34$, $p = .043$, $\eta_p^2 = .09$), such that, on average, estimated deaths were more frequent for dramatic vs. ordinary risks (this does not include the intrusions, i.e., the misrecalled causes of death), $M = 2.95$, *SD*

= 1.35 vs. *M* = 2.66, *SD* = 2.51. However, contrary to the recall results, this main effect was not qualified by an interaction with generation, *F* < 1. The main effect of generation was also not significant ($F(4, 43) = 1.63$, $p = .185$, $\eta_p^2 = .13$).

2.2.3 – Discussion

The results of Experiment 1 showed that dramatic are more easily recalled than ordinary ones, and that, as function of that memory advantage, they are more likely to survive in a communication chain where what one person recalls is transmitted to another person. Over generations, this resulted in a different forgetting curve for the two types of death (i.e., an interaction of generation-by-type of death, with greater forgetting of ordinary risks). Finally, dramatic causes of death also had an advantage for false memories.

2.3 – Experiment 2: Vividness, Distinctiveness and Memorability

Experiment 2 had several goals: First, it sought to replicate the results of Experiment 1 (i.e., dramatic risks are more memorable than ordinary ones) with a larger and more diverse sample (American Mturk participants).

Second, it sought to replicate the results of the pre-test (i.e., dramatic risks are more vivid than ordinary ones) and tie vividness to memorability, in order to provide evidence for the mechanism underlying the memory advantage of dramatic risks. To assess vividness, a measure similar to that used in the pre-test was evaluated how easy it is to picture the risk mentally. Moreover, an additional memory attribute was measured: distinctiveness. In particular, a sorting measure was borrowed from research on categorization (e.g., Isen & Daubman, 1984; Johnston & Hewstone, 1992; Rawson & Van Overschelde, 2008). This measure asked participants to group the different risks into categories that are as inclusive or exclusive as they decide. The rationale underlying this measure is that a risk is more distinctive to the extent that it is grouped with fewer risks and represented more uniquely in the semantic space. Thus, this experiment tested whether dramatic risks are more vivid and distinctive than ordinary ones, and whether any of these differences can account for the memory advantage of dramatic risks.

Finally, serial reproduction research has sometimes found that the recall varies as a function of processing goals. In particular, it can depend on whether people know that what they recall will be passed on to other people (a communication goal) or not (a mere memorization goal; Lyons & Kashima, 2001). Therefore, we wished to test whether the memory advantage of dramatic risks holds across different information processing contexts. Some participants were told that they were to memorize the information, whereas others were told that what they recalled would be communicated it to others.

2.3.1 – Method

Participants. One hundred participants were recruited on Mturk (a total of 101 ended up participating in the experiment). A sensitivity analysis with power set at 80% revealed that this sample size is sufficient to detect effect sizes of Cohen's $f \geq 0.14$, which is adequate considering the effect size for the main effect of type of death observed in Experiment 1 (f = 1.08).

Procedure. The procedure for presenting the information was the same as in Experiment 1, except that there was no serial reproduction: All participants saw the same information, varying in type of risk and frequency, as in Experiment 1.

In the communication-goal condition, participants read the following instructions: "Often, people tell us information about all kinds of events. Sometimes, we retell this information to other people, perhaps at a party or in conversation with a friend. The following task involves seeing some information and later passing it on to another participant in this experiment. That person will see the information that you provided and then he or she will pass it on to yet another participant, and so on. In sum, participants in this experiment receive information from some people and then pass it on to other people. The information in question pertains to the way different people died in a small village last year. This information was already seen by the participant before you, and what you will see next is what that person reported having seen." Participants in the memorization condition were given another set of instructions: "Often in life, we are required to remember various kinds of information. Sometimes it is necessary to remember this information in a very precise and accurate way, whether it is for work or some other activity. The following task is a test of how accurately you can remember the information that will be presented next. After memorizing the information, you will be required to recall it. The information in question pertains to the way different people died in a small village last year." And before they completed the recall and frequency estimation tasks, participants in the communication condition were told: "Remember that the information that you provide now will be passed on to another participant." Participants in the memorization condition were not given this information.

Thus, the design was a 2 (type of causes of death: dramatic vs. ordinary) x 3 (frequency: 1, 4, or 7) x 2 (instruction: communication vs. memorization), with the first two factors manipulated within-subjects, and the third one between-subjects.

After the recall and frequency estimation tasks, participants completed the two tasks assessing how vivid and distinctive the different risks are (the order of these tasks was counterbalanced between participants). One task was very similar to the one used in the pre-test: "For each of the causes of death below, please indicate how easy it is for you to form a concrete mental image of someone dying from it. This is not about whether it is emotionally easy or hard for you to think about it, but simply whether you can picture the event clearly and vividly in your mind, as if it was a photo or a movie." The different causes of death were then presented, and participants had to rate each of them on a scale from 1 – *very hard* to 9 – *very easy*.

The distinctiveness measure was provided by a sorting task where participants were told: “Now we ask you to sort the different causes of death into categories. To do so, please use numbers to place the items in similar categories. For example, if you had three items: apple, orange and astronaut, you should place the same number (for example, 1) in front of orange and apple and place another number (2) in front of astronaut. You can use as many categories as you need.” The different causes of death were again presented and participants had to assign a number/category to each. The results below are coded not for this number, but rather for the number of causes of death that share the same category (i.e., that share the same number), such that the higher this score, the less distinctive a certain risk is.

2.3.2 – Results

We used Linear Mixed Models, so that for each cause of death we could test for the effect of type of death, frequency, instruction, vividness and distinctiveness on recall.

Recall. A first Linear Mixed Model analysis tested whether the probability of recalling an event (0 = no, 1 = yes) differed as a function of: frequency (1 vs. 4 vs. 7), type of death (dramatic vs. ordinary) and instruction (memorization vs. communication). The intercept of both subject and event (i.e., the specific cause of death) were entered as random effects to control for possible variation. Both type of death ($F(1, 1199) = 27.53$, $p < .001$) and frequency ($F(2, 1199) = 98.32$, $p < .001$) significantly contributed to the model; all other effects were not significant (Fs < 1.32, ps > .268). Specifically, dramatic (vs. ordinary) deaths were more likely to be recalled, *estimate* = 1.50, *SE* = 0.43, $t = 3.47$, $p < .001$, 95% CI = [0.65; 2.34]. Additionally, low frequency (1), when compared to high (7) frequency of events, were less likely to be recalled (*estimate* = -1.96, *SE* = 0.31, $t = -5.86$, $p < .001$, 95% CI = [-2.61; -1.30]). The difference between medium (4) and high frequency (7) of events was not significant (*estimate* = -0.27, *SE* = 0.33, $t = -0.82$, $p = .414$, 95% CI = [-0.91; 0.38]).

Vividness. We first tested whether vividness was affected by type of death, instruction, and frequency, after controlling for the intercept of both subject and event. Type of death (*F*(1, 948.16) = 539.40, *p* < .001), frequency (*F*(2, 641.39) = 15.14, *p* < .001), and the interaction term of type of death by frequency (*F*(2, 641.39) = 8.33, *p* < .001) significantly contributed to the model. All other effects were not significant (*F*s < 2.98, *p*s > .085). Dramatic (vs. ordinary) deaths were more likely to be perceived as vivid, *estimate* = 2.45, *SE* = 0.27, *t*(294.56) = 9.02, *p* < .001, 95% CI = [1.92; 2.99]. Additionally, moderate frequency (4), when compared to high (7) frequency of events, were perceived as less vivid (*estimate* = -1.00, *SE* = 0.33, *t*(294.20) = -3.03, *p* = .003, 95% CI = [-1.65; -0.35]). No differences were found between low (1) and high (7) frequency of events for vividness (*estimate* = 0.17, *SE* = 0.32, *t* < 1, *p* = .599, 95% CI = [-0.46; 0.79]). Although the interaction term did contribute significantly to the model, the difference between dramatic deaths with low (*estimate* = -0.58, *SE* = 0.40, *t*(656.66) = -1.46, *p* = .145, 95% CI = [-1.35; 0.20]) and moderate frequency did not significantly differ from ordinary deaths with high frequency (*estimate* = 0.57, *SE* = 0.39, *t*(586.82) = 1.47, *p* = .143, 95% CI = [-0.19; 1.34]).

We then tested the direct effect of vividness on recall**.** Specifically, we tested the hypothesis that the more vividly one can imagine a certain type of death the higher is the likelihood that one will recall it. We ran a generalized linear mixed model, with recall as the dependent variable, vividness as the fixed effect, and both subject and event as random effects to adjust for possible variation. Results yielded a significant main effect of vividness on recall (*estimate* = 0.09, *SE* = 0.03, *Z* = 3.31, *p* < .001, 95% CI = [0.03; 0.14]).

Finally, we tested the potential mediation effect of perceived vividness in three steps (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014). First, we tested the effect of the type of death on the vividness of each event using a linear mixed model analysis (vividness was the dependent variable, type of death was entered as a fixed effect, and as a random effect we had the intercept for subjects to adjust for possible variation). The results yielded a significant main effect of type of death (*estimate* = 2.62, *SE* = 0.12, *t* = 22.69, *p* < .001, *d* = 1.12, 95% CI [2.39; 2.84]) on vividness. Dramatic deaths are more vivid (*M* = 7.39, *SD* = 2.09) than ordinary ones (*M* = 4.77, *SD* = 2.58). Next, we tested a second model entering type of death, vividness, and their interaction term as fixed effects, and recall as the dependent variable. As a random effect, we entered the intercept for type of event. Neither effects were significant (type of death: *estimate* = 0.09, *SE* = 0.79, *d* = 0.02, *Z* = 0.11, *p* = .914, 95% CI [-1.47; 1.64]; vividness: *estimate* = 0.05, *SE* = 0.04, *Z* = 1.17, *p* = .241, *d* = 0.24, 95% CI [-0.04; 0.14]). Lastly, we tested the mediation effect of vividness, grouped by subject, which was also not significant (*indirect effect* = 0.01, *p* = .520, 95% CI [-0.03; 0.06]).

Distinctiveness. We tested whether distinctiveness was affected by type of death, instruction, and frequency, after controlling for the intercept of both subject and event (entered as random effects). Type of death (*F*(1, 1164) = 350.18, *p* < .001), frequency (*F*(2, 1164) = 4.44, *p* = .012), and the interaction term of type of death by frequency (*F*(2, 1164) = 3.66, *p* = .026) contributed significantly to the model. All other effects were not significant (*F*s < 1.11, *p*s > .294). Dramatic (vs. ordinary) deaths were more likely to be perceived as more distinctive, *estimate* = -2.49, *SE* = 0.26, *t* = -9.732, *p* < .001, 95% CI = [-3.00; -1.99]. When analyzing the contribution of frequency, low frequency (1) did not significantly differ from high (7) frequency of events (*estimate* = -0.21, *SE* = 0.30, *t* < 1, *p* = .481, 95% CI = [-0.81; 0.38]). Moderate (4) frequency did not also differ from high frequency (7) of events (*estimate* = 0.10, *SE* = 0.31, *t* < 1, *p* = .764, 95% CI = [-0.50; 0.71]). Although the interaction term did contribute significantly to the model, the difference between dramatic deaths with low (*estimate* = 0.63, *SE* = 0.37, *t* = 1.73, *p* = .084, 95% CI = [-0.08; 1.35]) and moderate frequency did not significantly differ from ordinary deaths with high frequency (*estimate* = 0.53, *SE* = 0.37, *t* = 1.41, *p* = .158, 95% CI = [-0.21; 1.26]).

We then tested the direct effect of distinctiveness on recall. Specifically, we tested the hypothesis that distinctiveness (as measured by the sorting task) increases the probability of recall through a generalized linear mixed model analysis. Recall (0 = no, 1= yes) was entered as the dependent variable, distinctiveness was entered as fixed effect and the intercept of both subject and event were entered as random effects to control for possible variation. Results revealed that the more a certain cause of death was perceived as distinctive, the higher was the likelihood that it was recalled (*estimate* = -0.17, *SE* = 0.04, *Z* = -4.54, *p* < .001, *d* = 1.02, 95% CI = [-0.24; -0.01]).

We also tested a potential mediation effect by distinctiveness. Specifically, we tested the hypothesis that type of death influenced recall through distinctiveness, again in a three-step analysis using generalized linear mixed model analyses. For the first model, distinctiveness was entered as the dependent variable and type of death was entered as a fixed effect. The intercept for subjects was entered as a random effect. Results yielded a significant effect (*estimate* = -0.62, *SE* = 0.03, *t* = -19.48, *d* = 1.32, *p* < .001, 95% CI = [-0.68; -0.55]), such that dramatic deaths are more distinctive (*M* = 2.60, *SD* = 1.40) than ordinary deaths (*M* = 4.82, *SD* = 1.92). For the second model, we entered recall as the dependent variable, both main effects of type of death and distinctiveness as fixed effects, and we controlled for possible variation of event as a random effect. The effect of type of death was not significant (*estimate* = 0.56, *SE* = 0.55, *Z* = 1.02, *d* = 0.21, *p* < .310, 95% CI = [-0.52; 1.64]), but the effect of distinctiveness was (*estimate* = -0.09, *SE* = 0.04, *d* = 0.45, *Z* = -2.21, *p* < .027, 95% CI = [-0.17; -0.01]), such that deaths that were perceived as more distinctive had a higher chance of being recalled than less distinctive causes of deaths. Next, we tested whether the relation between type of death and probability of recall was mediated by distinctiveness, grouped by subject (Tingley et al., 2014). The mediation analysis was significant (*indirect* *effect* = 0.04, $p < .001$, 95% CI [0.01; 0.07]), suggesting that dramatic types of death were perceived as more distinctive and consequently had a higher chance of being recalled.

As in Experiment 1, an analysis similar to that which was done for recall data was performed for frequency estimates. The effects of type of death (F(1, 419.01) = 52.43, $p <$ .001), frequency (F(2, 233.81) = 6.73, p = .001), and their respective interaction term (F(2, 233.82) = 3.16, p = .044) were significant. Neither of the other fixed effects was significant (Fs < 1, ps > .433).

When analyzing the estimates of the fixed model effects, results revealed a main effect of type of death (*estimate* = 2.38, *SE* = 0.46, t(94.81) = 5.22, $p < .001$, 95% CI = [1.47; 3.28]), a significant main effect of low vs. high frequency (*estimate* = 1.50, *SE* = 0.56, t(109.89) = 2.69, p = .008, 95% CI = [0.40; 2.61]), a significant main effect of medium vs. high frequency (*estimate* = 1.43, *SE* = 0.53, t(97.43) = 2.71, p = .008, 95% CI = [0.38; 2.48]), qualified by a significant interaction effect of dramatic and low frequency deaths vs. ordinary and high frequency deaths (*estimate* = -1.55, *SE* = 0.69, t(234.63) = -2.25, p = .026, 95% CI = [-2.91; -0.19]). All other estimates did not reach significance (ts < 1.61, ps > .108).

Simple pairwise comparison tests were used to decompose the interaction of type of death by frequency. Participants made higher estimates for dramatic deaths regardless of the frequency by which these were presented. However, the mean difference between dramatic and ordinary death was higher for the high frequency deaths (M^{diff} = 2.15, *SE* = 0.32, $p <$ .001), followed by the moderate frequency (M^{diff} = 1.36, *SE* = 0.33, $p < .001$) and by the low frequency deaths (M^{diff} = 0.99, *SE* = 0.38, p = .010), similar to Experiment 1.

2.3.3 – Discussion

The results of this experiment replicate those of the previous ones: 1) dramatic risks were more vivid than ordinary ones (replicating the pre-test), and also more distinctive; 2) dramatic risks were more memorable than ordinary ones (replicating Experiment 1); 3) the fact that dramatic risks are more vivid and distinctive than ordinary ones seems to account for why they are more memorable (even though the mediation test only produced significant results for distinctiveness, not vividness); 4) indeed, the effects of vividness and distinctiveness are such that, if a certain risk is vivid or distinctive, be it ordinary or dramatic, then it is more likely to be remembered. Finally, processing goals had no effect on the results: regardless of whether participants were told that the information that they were asked to recall was for memory-testing purposes or whether it would be retold to other people.

2.4 – General Discussion

Previous studies on the social amplification of risk have examined how people transmit information about a certain risk. For instance, Moussaïd et al. (2015) studied how people talked about information that they had read on media articles about an antibacterial agent (see also Jagiello & Hills, 2018). However, in the flood of information that people receive from the media or from other people, they are likely to learn about several risks, not just one. This experiment tested not how information about specific risks gets distorted in the course of social transmission of information, but rather how entire risk factors can get filtered out from the public discourse, while others are selectively propagated in communication networks.

This experiment showed that people are more likely to remember, and therefore transmit, information about dramatic risks (non-natural causes of death that are easy to imagine vividly and distinctively) than information about ordinary risks (natural causes of death that are more difficult to picture in a vivid and distinctive manner). While the former are actually rarer than the latter, this experiment shows that they can have a memory advantage and be more likely to survive in the public discourse. This might also account for availability cascades (Kuran & Sunstein, 1999) whereby the public fear of certain risks can get overblown when comparing the attention devoted to them with their actual risk potential (see also Sunstein & Zeckhauser, 2011).

Studies on serial reproduction, including recent studies on the social amplification of risk, typically focus on the content of single case stories. But the content of single cases is not the only thing that people communicate. They also talk about the frequency of different kinds of cases. The method used in Experiment 2 reveals how social transmission can create biased samples with regard to the frequency of different kinds of cases. This paradigm enables the study of how frequency biases, such as illusory correlations (Hamilton & Sherman, 1989), or the ratio bias (Denes-Raj & Epstein, 1994), can be amplified in society.

One basic premise in the present approach is that participants need to remember a certain piece of information in order to communicate it to others. In previous studies (e.g., Moussaïd et al., 2015), it is not clear whether participants did not communicate certain elements of a risk message because they did not remember them, or because they did not find them relevant. In the present experiment, the memory-dependent nature of interpersonal communication about risk became clear: Certain risks were simply forgotten completely, and therefore could not be communicated to others, even if people wished to do so.

On the other hand, there were distortions that were introduced in the process of social communication. Consistent with the pattern of forgetting, whereby ordinary causes of death were more likely to be filtered out from the communication chains, these results suggest an advantage for dramatic risks in terms of false memory as well. Indeed, when new causes of death were introduced, they were more likely to be dramatic than ordinary.

Is the bias in the mind or in the environment?

Despite the role of memory errors in the social amplification process documented here, participants were nevertheless quite sensitive to the actual frequency of stimuli. That is, they showed a good overall sense of whether certain causes of death occurred frequently or not. Thus, it takes more than the shortcomings of memory to explain the present results.

We believe that this study provides a fitting demonstration of the principles of the ecological perspective on judgment and decision making, whereby the reasons for judgment biases are not only to be found in intrapsychic mechanisms, but also in the information environment in which those mechanisms operate (Fiedler, 2000a; Fiedler & Wänke, 2009). Indeed, the social amplification process described here has both intrapsychic and ecological bases: There is a memory bias, which may be small when considering a single generation of participants (in Experiment 1, the difference in recalling dramatic vs. ordinary deaths in the first generations was not significant). However, that small bias will feed into the information environment in which the next generation of participants have to form their impressions and make their judgments (i.e., a biased ecology). In other words, an individual's biased processing will generate biased information that will be fed to others. It is this interplay of intrapsychic and ecological biases that accrue over generations to form a social amplification pattern.

3 – Chapter III – The social amplification of illusory correlations. Mendonça, C., Mata, A.,[10] & Ferreira, M. B. (*submitted*)

The perception of correlations is pervasive in everyday life and underlies many of the choices and judgments that we make. For example, people may perceive a correlation between social groups (e.g., immigrants) and behavior (e.g., crime), or between events (e.g., air pollution) and consequences (climate change), or even between clinical signs (e.g., drawings with salient eyes) and mental health (e.g., paranoia). When these perceived correlations do not correspond to reality, they are said to be illusory. They can be illusory because the correlation does not exist, or because it has a different intensity or direction than the one perceived (L. J. Chapman, 1967).

Because correlation judgments are involved in various phenomena, such as stereotypes (e.g., Hamilton & Gifford, 1976; Hamilton & Rose, 1980), phobias (e.g., Tomarken, Mineka, & Cook, 1989; Wiemer & Pauli, 2016), and perception of diagnosticity (e.g., L. J. Chapman & Chapman, 1967; King & Koehler, 2000), illusory correlations have been used to explain judgments and behaviors in many areas, such as social psychology (e.g., Hamilton & Gifford, 1976; Schaller & Maass, 1989), clinical psychology (e.g., L. J. Chapman & Chapman, 1969; Wiemer & Pauli, 2016), organizational psychology (e.g., Pryor & Stoller, 1994; Smither, Collins, & Buda, 1989), political psychology (Castelli & Carraro, 2011; Rodríguez-Ferreiro & Barberia, 2017), and forensic psychology (e.g., Carter, 2018; Smith & Alpert, 2007).

Research on illusory correlations has focused mostly on the cognitive factors that lead to the emergence of the bias, with several potential mechanisms and models being proposed,

[10] C. Mendonça and A. Mata share co-first authorship of this work.

such as distinctiveness (Hamilton & Gifford, 1976), accentuation of inter-category differences (McGarty, Haslam, Turner, & Oakes, 1993), regression towards the mean (Fiedler, 1991), learning mechanisms such as those proposed in the Rescorla-Wagner model (Murphy, Schmeer, Vallée-Tourangeau, Mondragón, & Hilton, 2011), the Brunswikian induction algorithm for social inference model (Fiedler, 2000b), and the twofold retrieval by associative pathways model (Garcia-Marques & Hamilton, 1996).

In this paper, our focus is not so much on the cognitive factors and how illusory correlations emerge, but rather on social factors and how illusory correlations can spread and intensify as information is communicated throughout society. Indeed, an important social factor that has been neglected in illusory correlations research is the social dynamics of information transmission.

The various ways in which the social transmission of information can bias the outcome of basic cognitive processes has been well established since Bartlett's (1932) seminal work. In Bartlett's serial reproduction paradigm, participants form communication chains in which the output of one participant (usually the result of a free recall task) is provided as the input for the next participant. In this way, participants form generations (participants who receive the original message are the first generation, those who receive information from a participant of the first generation are the second generation, and so on), with the collective output of each generation becoming increasingly more distant from the original material. Indeed, Bartlett found that the material transmitted in these chains became progressively conventionalized, such that the material was simplified and made more consistent with pre-existing cultural expectations.

Recent studies have replicated Bartlett’s work (Roediger et al., 2014) and expanded this paradigm to other areas, such as stereotype maintenance (Kashima, 2000), cultural evolution (Mesoudi, 2007), inter-group conflict (Lee et al., 2014), transmission of pseudoscientific beliefs (Mercier, Majima, & Miton, 2018), and the advantage of negative information (Bebbington et al., 2017). In these various areas, the serial reproduction paradigm has shed light on how information changes and evolves as it travels through society. With the present studies, we aim to use the serial reproduction paradigm to gain insight on how illusory correlations can propagate in society.

In two experiments, we seek to study the social amplification of illusory correlations. Experiment 1 uses the original paradigm that gave rise to research on illusory correlations in social psychology (Hamilton & Gifford, 1976). In this paradigm, illusory correlations emerge in a bottom-up fashion due to asymmetries in the frequencies of stimuli. The rationale is as follows: Information is presented about the frequency of cases where members of two groups (groups with no particular meaning a priori; e.g., "Group A" vs. "Group B") perform positive and negative behaviors. The proportion of positive vs. negative behavior is the same across groups, but different frequencies are used to convey the information about each group (e.g., 18 positive and 8 negative behaviors for Group A, 9 positive and 4 negative behaviors for Group B). Typically, participants' estimates of how many behaviors of each type were performed by each group reveal an illusory correlation, such that they associate the majority group (in this case, Group A) with the more frequent behaviors (in this case, positive behaviors), and the minority group (Group B) with the infrequent (negative) behaviors (for a meta-analysis, see Mullen & Johnson, 1990). By serially reproducing this material using the frequencies reported by one participant to determine the stimuli seen by the next participant, we expect each participant to add distortion to an increasingly distorted message. Because we expect most participants' memory bias to be systematic (i.e., we expect them to distort the frequencies in the same way, consistent with the original results), we expect the illusory correlation to be socially amplified.

In Experiment 2, we turn from bottom-up processes (based on stimuli frequency) to top-down processes (based on expectations). Borrowing from a different paradigm of research on illusory correlations (Hamilton & Rose, 1980), this experiment explores the fact that social perceivers have expectations about the associations between certain groups and certain attributes (e.g., accountants are boring). These expectations then lead perceivers to notice an illusory correlation, by overestimating the frequency of stereotype-consistent traits. As in Experiment 1, we expect the illusory correlation to be socially amplified, such that the overestimation of stereotype-consistent traits increases as information about certain groups travels through different people.

Ultimately, these studies seek to demonstrate the importance of considering the social dynamics of information transmission in assessing the true impact of illusory correlations. Outside the lab, people often do not receive information from its original source but rather from others, and in this process, information changes and evolves. Thus, the real extent and impact of illusory correlations in society might have been underestimated in previous research using single generations and carefully controlled experimenter-provided information. By focusing on the dynamic aspect of information transmission in society, the present studies align with recent research that brought new insights into classical social psychology phenomena, such as social influence (e.g., Mason, Conrey, & Smith, 2007; Salganik, Dodds, & Watts, 2006).

3.1 – Experiment 1

In Experiment 1, we fused Hamilton and Gifford's (1976) paradigm of illusory correlations driven by frequency asymmetries with the serial reproduction method (Bartlett, 1932). In a first phase, participants read sentences describing desirable and undesirable behaviors performed by members of two fictitious groups. In a second phase, participants were asked to estimate how many members of either of the two groups committed desirable and undesirable behaviors, and to express their attitudes towards the two groups. The frequency estimates were then used to determine the stimuli that the next participants would see. We expected to find that, as information travelled through chains of participants, the illusory correlation would become socially amplified, with the bias of each individual accumulating across generations. We also expected that this amplification effect would have an indirect impact on the way participants perceived the groups, such that, as the illusory correlation increases, so too should the preference for the majority group over the minority group.

3.1.1 – Method

Participants. A total of 118 undergraduate psychology students at a large European university participated in this experiment. Given the available sample, we aimed to collect 35 participants per generation, for a total of three generations (105 participants). Mullen and Johnson's (1990) meta-analysis reports an effect size of $r = .46$ for a paradigm similar to the one implemented in this experiment (where the distinctive behaviors are negative, and the dependent variable is frequency estimates). Our sample size afforded 81.48% power to detect this effect in the first generation. We did not know which effect size to expect for the social amplification effect, but sensitivity analysis using G*Power 3.1.9.2 (Faul, Erdfelder, Buchner, & Lang, 2009) revealed 80% power to detect an $r = .27$ with our full sample, which is adequate to detect effects of intermediate size or even smaller (the conventional criterion for intermediate effect sizes is .3; Cohen, 1988).

Each generation was collected serially (i.e., we only began collecting data for the next generation after completing data collection from the previous generation). Five participants were removed and replaced because they gave non-numeric responses to the frequency estimates. Eight extra participants were collected in the last generation due to a miscommunication with the laboratory assistants but were not included in the analysis. Thus, a total of 105 participants were included in the analysis (no demographic information was collected in this experiment, but the population from which the sample was drawn is comprised mostly of female undergraduates).

Materials and procedure. Participants started the experiment by reading a brief description of its goals and procedure. They were informed that participants in this experiment both received information from, and transmitted information to, other participants. Then, they were asked to read attentively all the sentences and to press a key each time that they wished to read the next sentence (i.e., the information intake was self-paced).

Participants who formed the first generation received the same frequencies used in Hamilton and Gifford's (1976) first study: 18 positive behaviors performed by members of Group A, 8 negative behaviors by Group A members, 9 positive behaviors by Group B members, and 4 negative behaviors by Group B members. Therefore, there was no difference in the proportion of the two types of behaviors across groups. Participants in the second and third generation were assigned to one of the 35 chains. The frequencies that they saw depended on what the previous participant in that chain had estimated. Thus, for instance, if a given participant reported only 20 positive behaviors by Group A and 10 negative behaviors by Group B, the next participant in the chain would read just 30 sentences: 20 sentences describing positive behaviors by Group A, 10 sentences describing negative behaviors by Group B, and no sentence describing positive behaviors by Group B or negative behaviors by Group A.

To avoid experimental sessions that were excessively long, we imposed an upper limit on the total number of sentences that participants could receive. Specifically, when the participant from the previous generation had recalled absolute frequencies over 40, the next participant in the chain received a total of 40 sentences, preserving the proportion of the four types of sentences. For example, if a participant reported having seen 20 (33%) positive behaviors by Group A, 10 (17%) positive behaviors by Group B, 10 (17%) negative behaviors by Group A and 20 (33%) negative behaviors by Group B (a total frequency estimate of 60), the next participant would see 13 (33%) positive behaviors by Group A, seven (17%) positive behaviors by Group B, seven (17%) negative behaviors by Group A, and 13 (33%) negative behaviors by Group B.

Sentences were formed by a male name, a group membership clause, and a behavior description (e.g., "Filipe, of Group A, makes frequent donations to charity"). Behaviors were taken from norms for desirable and undesirable behaviors (Garrido, 2003) and formed two pools (desirable and undesirable). For each participant, the program randomly, and without replacement, selected a behavior from the undesirable or desirable pool depending on the previously defined frequencies (see previous paragraph) of each type of behavior for each group.

When participants finished reading all the sentences, they were asked to take part in a different, unrelated task, which served, for the purposes of the current study, as a distractor task. As participants finished the unrelated task, they were informed that they would continue the task about Groups A and B. Afterwards, they were asked to estimate the number of people in each group that had performed either desirable or undesirable behaviors by inserting a number after each prompt (e.g., "People of Group A who performed desirable behaviors:"). They were also reminded that a participant in the next generation would receive their frequency estimates. Participants in the first generation, as in Hamilton and Gifford's (1976) experiment, were told the total number of sentences and members of each group (though they were not informed about the frequencies of desirable vs. undesirable behaviors), but such information was omitted in Generation 2 and 3, as these varied for each individual participant.

Finally, participants were asked to express their attitudes towards each group, using 7-point scales from 1 (*Very negative*) to 7 (*Very positive*).

3.1.2 – Results

Descriptive statistics for all dependent variables can be seen in Table 1.

Table 1. Means (and standard deviations) for all dependent variables of Experiment 1 per generation.

Variable	Generation 1	Generation 2	Generation 3
Frequency of Group A desirable behaviors	11.29 (5.91)	9.53 (6.86)	8.69 (3.59)
Frequency of Group A undesirable behaviors	8.14 (3.92)	8.18 (9.73)	6.11 (2.93)
Frequency of Group B desirable behaviors	9.11 (4.50)	8.94 (11.17)	6.34 (2.46)
Frequency of Group B undesirable behaviors	8.37 (4.05)	8.81 (6.02)	8.06 (4.39)
Phi coefficient	0.05 (0.17)	0.07 (0.19)	0.14 (0.19)
Liking of Group A	4.66 (0.94)	4.40 (0.91)	4.69 (0.96)
Liking of Group B	4.26 (0.98)	3.86 (0.91)	3.63 (1.14)

To measure the association inherent in participants' frequency estimates, we calculated a phi coefficient for each participant (see Hamilton & Gifford, 1976) using the following formula:

$$((A^{+} \cdot B^{-}) - (A^{-} \cdot B^{+})) / \sqrt{}((A^{+} + B^{+}) \cdot (A^{-} + B^{-}) \cdot (A^{+} + A^{-}) \cdot (B^{+} + B^{-}))$$

Thus, as the phi coefficient becomes more positive, the association between Group A and desirable behaviors and between Group B and undesirable behaviors becomes stronger.

Considering only the first generation, we obtained a marginal difference between the mean phi coefficient and zero, $t(34) = 1.71$, $p = .097$, $d = 0.29$, 95% CI [-0.04, 0.63] (Hamilton and Gifford also report a marginal difference: $t(32) = 1.92$, $p < .10$), and a significantly greater overestimation of undesirable behaviors by Group B (*M* overestimation = 4.37, *SD* = 4.05) than of undesirable behaviors by Group A (*M* overestimation = 0.14, *SD* = 3.92), *t*(34) = -7.35, *p* < .001, *d* = -1.22, CI [-1.74, -0.71].

The next set of analyses tested social amplification effects. Observing the effect of generation on each type of sentence separately, there is no reliable linear relation: for desirable behaviors by Group A, *r*(103) = -.19, *p* = .055, 95% CI [-.37, .00]; undesirable behaviors by Group A, *r*(103) = -.13, *p* = .180, CI [-.31, .06]; desirable behaviors by group B, *r*(103) = -.16, *p* = .105, CI [-.34, .03]; undesirable behaviors by Group B, *r*(103) = -.03, *p* = .788, CI [-.22, .16]. However, and more importantly, the main test of our social amplification hypothesis shows that there is a linear relationship between generation and the phi coefficient, *r*(103) = .20, *p* = .044, 95% CI [.01, .38], such that, overall, the association of desirable behaviors with Group A and negative behaviors with Group B increased over generations.

Moreover, we tested whether the effect of generation had an indirect impact on the difference between liking Group A versus B, mediated by the phi coefficient (see this model depicted in Figure 4). Using model 4 of the PROCESS macro (Hayes, 2013) with 10,000 bootstrap samples for bias-corrected confidence intervals, a significant indirect effect emerged, *b* = 0.29, *SE* = 0.13, 95% CI [0.02, 0.54]. As predicted, as generations progressed, the phi coefficient became more positive (in favor of Group A vs. B), and in turn this promoted more liking of Group A versus B.

3.1.3 – Discussion

Replicating the classical findings of Hamilton and Gifford (1976), the results of the first generation revealed an illusory correlation in participants' frequency estimates, such that there was a perceived association of the majority group (Group A) with the most common behavior (desirable behaviors) and of the minority group (Group B) with the least common behavior (undesirable behaviors), as well as a significantly stronger overestimation of Group B's undesirable behaviors by comparison with the overestimation of Group A's undesirable behaviors.

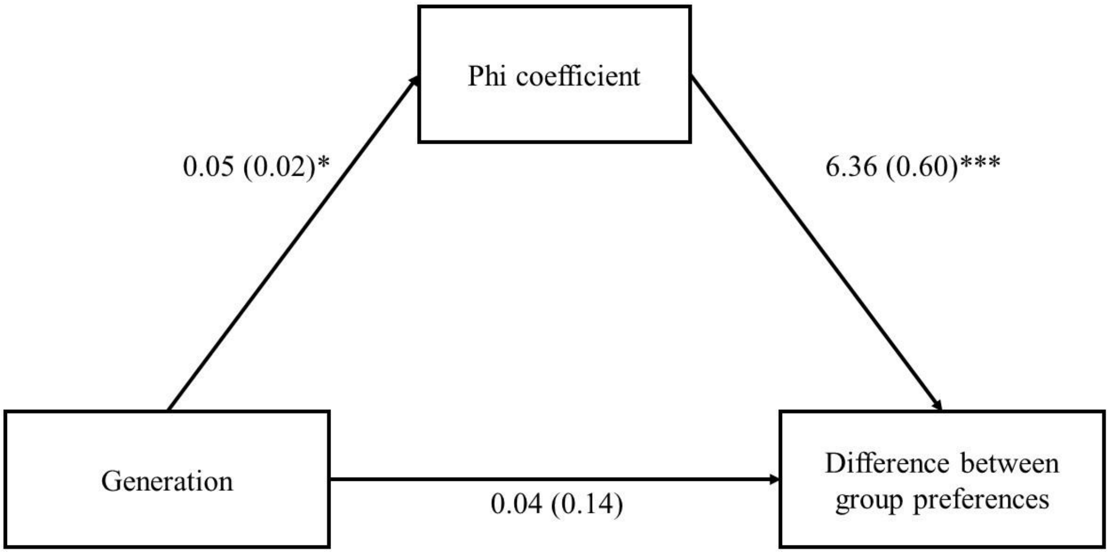


Figure 4. Mediation model of the indirect effect of generation in difference between group preference through the perceived correlation (measured by the phi coefficient) with unstandardized regression coefficients (outside brackets) and standard errors (inside brackets); * = p < .05; *** = p < .001.

More importantly, the illusory correlation was amplified through social communication: As information was transmitted from one participant to the next, what started out as a mere trend in the direction of illusory correlation (i.e., the perceived contingency was only marginally significant) then grew larger, such that, as generations passed, there was an increase in the perceived association of the majority group with the most frequent behavior and of the minority group with the least frequent behavior. This social amplification of the illusory correlation in turn led participants to prefer Group A over B.

3.2 – Experiment 2

In the first experiment, we saw that illusory correlations driven by bottom-up processes can be socially amplified, and that this amplification then has an impact on participants' attitudes towards abstract and artificial groups (Group A, Group B). In the second experiment, our first goal was to generalize this finding to top-down processes using Hamilton and Rose's (1980) paradigm. In this paradigm, the groups are familiar to participants, and the illusory correlation is driven not by bottom-up factors, such as the asymmetries in the frequencies of each group's behaviors, but by top-down factors: namely, people's expectations about the groups.

The second goal of Experiment 2 was to test whether an illusory correlation can be amplified not only interpersonally, as an individual passes on information to other individuals, but also intrapersonally, as a single individual remembers the same information repeatedly and in the process of doing so may add a systematic bias to the remembered information. Indeed, people do not just communicate information to others, but they also retell or recall information in multiple occasions, and their memories or impression of data may change as a result. These two processes – interpersonal communication, and intrapersonal repeated retrieval – are central to the shaping of collective memories (Roediger, Zaromb, & Butler, 2009). In order to test their joint influence on the amplification of illusory correlations, we turned to a paradigm that merges the serial reproduction and the repeated reproduction paradigms (Roediger et al., 2014). Specifically, in this experiment, participants were not only asked to recall information once, in order to pass it on to others; they also recalled the original information repeatedly, at different times. With this paradigm, we hoped to capture the different ways in which collective memories form, and how information changes or crystalizes as people communicate and recall it.

3.2.1 – Method

Participants. A total of 154 participants at a large European university, were recruited through an external participant pool. We aimed to collect 50 participants per generation, until we reached three generations (for a total of 150 participants). As in Experiment 1, each generation was collected serially. Four participants in the third generation were removed due to internet-access problems in the laboratory that led to timeout errors in accessing the Qualtrics server. These participants were replaced with four new participants, such that the originally planned 150 participants remained even after exclusions (103 females, 47 males, $M_{\text{age}} = 25.28$, $SD_{\text{age}} = 5.84$).

In this experiment, we used an expectancy-based illusory correlation paradigm with the proportion of each type of attribute in a forced recognition task as the dependent variable. We could not find a similar previous experiment (including Hamilton & Rose's, 1980) that enabled us to calculate an effect-size estimate. A sensitivity analysis using G*Power (Faul et al., 2009) revealed 80% power to find an effect size equal to or above an $r = .37$ in the first generation (i.e., an effect size within the intermediate effect size range, between .30 and .50; Cohen, 1988). As for the social amplification hypothesis, we obtained 80% power with our full sample to detect an $r = .22$, which is adequate considering the effect size in Experiment 1 ($r = .20$).

Materials. The stimuli were 24 sentences composed of a male name, a profession membership clause, and two attributes. The attributes were obtained in a pre-test in which 43 participants rated the degree to which 12 attributes (environmentalist, friendly to animals, creative, eccentric, gluttonous, fat, adventurous, well-travelled, hard-working, nice, rich, and upper class) were typical or not typical of five professions (doctor, biologist, cook, archeologist, designer).

In line with Hamilton and Rose's (1980) first experiment, we aimed to select three professions and eight attributes: one set of two attributes equally typical of each of the three professions, and three sets of two attributes that were more typical of one profession than of the others. We selected the professions doctor, designer, and biologist with the following attributes: *upper class* and *rich* as typical of doctor, *environmentalist* and *friendly to animals* as typical of biologist, *creative* and *eccentric* as typical of designer, and *well-travelled* and *nice* as equally typical of all three professions (means and standard deviations for each attribute and each profession can be consulted in Table 2). An example of one of the sample sentences is: "Bruno, a doctor, is rich and eccentric".

Table 2. Means (and standard deviations) of typicality (1 – *Not at all typical* to 7 – *Very typical*) in the pre-test (*N* = 43) of Experiment 2 for the chosen professions and attributes.

	Doctor	Biologist	Designer
High class	7.51 (1.22)	4.67 (1.57)	5.44 (1.83)
Rich	7.12 (1.33)	4.26 (1.75)	5.12 (1.68)
Environmentalist	5.00 (1.79)	8.16 (1.09)	5.37 (2.22)
Friendly to animals	5.56 (1.70)	8.19 (0.88)	4.91 (2.08)
Creative	4.72 (1.65)	4.72 (1.55)	8.42 (0.82)
Eccentric	5.44 (1.62)	4.53 (1.79)	7.33 (1.34)
Well-travelled	6.60 (1.56)	6.86 (1.68)	6.53 (1.52)
Nice	5.77 (1.23)	6.05 (1.75)	5.40 (1.59)

Following Hamilton and Rose (1980), in the original material provided to the first generation, each attribute could only be paired with an attribute of a different kind (e.g., two attributes typical of the same profession could not be paired in the same sentence); and each attribute was paired with each of the remaining six attributes only once. Thus, there was no correlation between any attribute or any type of attribute and any profession. To deal with potential order effects, the order of each attribute was randomized within a pair, and to deal with potential interactions between each specific attribute and a given profession, we created three sets of pairs of attributes with the same frequency of types of attributes and counterbalanced the attribution of each set to each profession.

Procedure. Participants started the experiment by providing demographic information and reading a brief description of the objectives and procedure of the experiment. Then, participants read the 24 sentences in randomized order, one-by-one, with an eight second interval automatic presentation.

After the stimuli were presented, participants completed the first forced recognition task under the following instructions:

> "The first part of the study has ended. Next, we will again present the sentences that you have just read. You are to select the two, and only those two, attributes that you remember having been used to describe the person described in the sentence.
>
> Try to be as precise as possible. If you cannot recall with precision, please select the two attributes that appear to you most consistent with what you read, so as to transmit information to the next participant that is as close as possible to what you read initially.
>
> The sentences, with the two attributes you select, will be presented to other participants in this study. Therefore, always select two attributes so that the next participants can read complete sentences."

After reading these instructions, participants were presented the 24 sentences again with the attributes omitted from the sentence (e.g., "Bruno, a doctor, was ____ and ____."), and they had to select two attributes out of the total eight attributes, in a multiple-choice format. This forced recognition task kept memory demands constant throughout the experiment (as all participants saw 24 sentences).

After the first forced recognition task, participants completed another task, which served as a distractor for the purposes of this study, then repeated the forced recognition task, after which they again completed a distractor task, and finally repeated the forced recognition task yet again. Thus, each participant did the forced recognition task of the 24 sentences a total of three times. This design mimics the combination of the serial and repeated reproduction paradigms implemented by Roediger et al. (2014), and it tested whether the hypothesized distortion effects occur not only inter-personally, as a function of the social transmission of information, but also intra-individually, as a function of how memory provides an advantage to certain pieces of information and neglects others over time.

As for the serial reproduction procedure, we also used the design by Roediger et al. (2014), whereby only the first of the three forced recognition tasks was used to produce the sentences that were transmitted to the next generation of participants. As in Experiment 1, the transmission chains were created such that the output of a previous participant in a given chain was used as stimuli for the next participant in the chain. The participant that was next in the chain thus saw each of the 24 sentences (including the name of the person described in the sentence) with the attributes selected by the previous participant. The order of the attributes in each pair was randomized, as it was not possible to register the order in which participants selected the attributes in the forced recognition task.

3.2.2 – Results

When analyzing results, we were interested in the frequencies of three types of attributes: a) neutral attributes (i.e., the attributes that were equally typical of the three professions); b) typical attributes (i.e., the attributes considered typical of each profession); and c) non-typical attributes (for each profession, the non-typical attributes were those that were typical of the other professions). Because, for any given profession, there were more non-typical attributes (four) than either typical (two) or neutral (two) attributes, we calculated the proportion of typical, non-typical, and neutral attributes, in relation to the original frequency. Thus, a proportion above one reveals an increase in frequency in relation to the original frequency, while a proportion below one reveals a decrease in frequency in comparison to the original material. Means and standard deviations of the absolute frequencies, per type of attributes and generation, are presented in Table 3.

Table 3. Mean (and standard deviation) of absolute frequencies per generation, recall, and attribute type in Experiment 2.

	Attribute type		
	Typical	Neutral	Non-typical
Original material	12	12	24
		Generation 1	
Recall 1	16.06 (5.09)	12.04 (3.45)	19.90 (5.50)
Recall 2	17.04 (4.83)	12.24 (3.43)	18.72 (5.79)
Recall 3	17.36 (6.24)	12.26 (3.52)	18.38 (6.37)
		Generation 2	
Recall 1	18.34 (5.81)	12.22 (3.86)	17.44 (5.88)
Recall 2	19.84 (7.13)	12.48 (3.78)	15.68 (7.27)
Recall 3	20.98 (7.02)	11.96 (4.26)	15.06 (6.96)
		Generation 3	
Recall 1	20.90 (5.75)	11.82 (4.08)	15.28 (5.81)
Recall 2	21.46 (6.29)	11.98 (4.44)	14.56 (7.51)
Recall 3	21.70 (6.48)	11.76 (4.55)	14.54 (7.27)

To test whether results are similar to those observed by Hamilton and Rose (1980), we first conducted a 3-way (attribute type: typical, non-typical, neutral) repeated-measures ANOVA for the data of the first generation, first recall only. A significant attribute type effect was obtained, $F(1.74, 85.46) = 22.50$, $p < .001$, such that the proportion of typical attributes ($M = 1.34$, $SD = 0.42$) was higher than that of neutral attributes ($M = 1.00$, $SD = 0.29$), $t(49) = 4.22$, $p < .001$, $d = 0.52$, 95% CI [0.12, 0.92], and the proportion of non-typical attributes ($M = 0.83$, $SD = 0.23$) was lower that of than neutral attributes $t(49) = -2.82$, $p = .007$, $d = -0.44$, CI [-0.83, -0.04]. Thus, we replicated the classical findings (Hamilton & Rose, 1980).

In terms of the serial reproduction, we found that, as generations progressed, the proportion of typical attributes increased, $r(148) = .34$, $p < .001$, 95% CI [.20, .46], the proportion of non-typical attributes decreased, $r(148) = -.32$, $p < .001$, CI [-.45, -.17], and the proportion of neutral attributes remained constant, $r(148) = -.02$, $p = .772$, CI [-.18, .14]. To allow an analysis similar to the one done in Experiment 1 with the phi coefficient, we calculated the following index:

(Typical attributes – Non-typical attributes) / (Typical attributes + Non-typical attributes)

This index, thus, varies from 1 (predominance of typical attributes) to -1 (predominance of non-typical attributes). The analysis with this index revealed an increase, as generations progressed, of the predominance of typical attributes over non-typical attributes, $r(148) = .34$, $p < .001$, 95% CI [.19, .48], also supporting the hypothesis of social amplification.

The same pattern of results of the serial reproduction was found in the repeated reproduction, such that, as participants did more repetitions of the forced recognition task, the proportion of typical attributes increased, $r(448) = .10$, $p = .032$, CI [.01, .19], the proportion of non-typical attributes decreased, $r(448) = -.09$, $p = .047$, CI [-.18, -.01], and the proportion of neutral attributes remained constant, *r*(448) = .00, *p* = .941, CI [-.10, .09]. Comparing the two methods of reproduction as independent,[11] the serial reproduction method lead to a stronger increase of typical attributes, *Z* = 2.67, *p* = .008, and a stronger decrease of non-typical attributes, *Z* = -2.54, *p* = .011, but no difference in the neutral attributes, *Z* = -0.21, *p* = .833, as compared to the repeated reproduction.

3.2.3 – Discussion

In this second experiment, we replicated the classical, expectancy-guided illusory correlation (Hamilton & Rose, 1980). Furthermore, as the information was transmitted from one participant to the next, the illusory correlation was intensified, such that the proportion of typical attributes increased, while the proportion of non-typical attributes decreased, and the proportion of neutral attributes remained constant. This social amplification of bias was also found to be stronger than the increase in bias observed when the same participants repeated the same task over time, in line with Bartlett's (1932) findings of stronger distortion in the serial reproduction than in the repeated reproduction method. Thus, the findings of Experiment 1 were generalized in Experiment 2 to a different type of illusory correlation mechanism (previous expectations instead of frequency asymmetries), and a different type of dependent variable (forced recognition instead of frequency estimates).

3.3 – General Discussion

The results of these experiments suggest that illusory correlations can grow stronger as information travels across different people. In Experiment 1, we used Hamilton and Gifford's (1976) illusory correlations paradigm, whereby participants read sentences about members of two fictitious groups who perform desirable or undesirable behaviors. As generations progressed, when participants estimated the frequency of each behavior type

[11] The two correlations are, in fact, not independent, as the first recall of each generation was used both in the serial and repeated reproductions. Yet, because there was no simple way to account for this dependence, we treated the correlations as independent.

performed by each group, they showed an illusory correlation that increased in intensity. This social amplification of the illusory correlation in turn had an indirect impact on the way participants evaluated the two groups, such that the majority group was progressively seen in a more positive light than the minority group. In Experiment 2, we used Hamilton and Rose's (1980) paradigm, in which participants read sentences about three professions and their typical, neutral, or non-typical characteristics. As generations progressed, participants progressively attributed more typical attributes and less non-typical attributes to each profession.

Thus, the two experiments reported in this paper demonstrate a social amplification effect using the two most popular illusory correlation paradigms (Mullen & Johnson, 1990), with both top-down and bottom-up factors driving the effect, and with frequency estimates and forced recognition as dependent measures.

Implications. A key implication of this research is that the size and impact of illusory correlations in society might be underestimated in research that neglects the social dynamics of information transmission. In the present studies, even when the results of the first generation suggested small and marginally significant biases (as was the case in Experiment 1), these biases grew larger when information was transmitted from one person to the next. This paradigm serves as a magnifying lens that allows for even small and apparently inconsequential biases to grow and reach significance.

We expect these results to generalize to any domain where there is a dominant bias that shapes the information people communicate to others. Indeed, social amplification-like effects have been shown with varied phenomena, such as cultural expectations (Bartlett, 1932), stereotypes (Kashima, 2000), and the negativity bias (Bebbington et al., 2017), and recent research also shows the importance of transmission chain homogeneity (i.e., the existence of a dominant bias) in the predictability of social amplification-like effects (Navarro et al., 2018).

These results also have implications for the understanding of collective memories. In particular, Experiment 2 captured the two mechanisms that are thought to underlie the way in which collective memories emerge and grow in society: intra-individual repeated reproduction and inter-individual serial reproduction (Roediger et al., 2009). Even though the effect was larger for the latter than for the former, repeated reproduction was also shown to amplify illusory correlations. Indeed, Roediger et al. posited that repeated retrieval of erroneous information strengthens the trace it leaves in memory, which is consistent with our finding that the repeated retrieval of the same event (without feedback) led to an increase in illusory correlations over time. The combination of this individual bias over time as the same individual tells the story again and again, with the social amplification of bias as information travels from one person to the other, should lead to collective memories that are strongly biased, at least by comparison with individuals recalling the event for the first time after encountering the data first-hand.

Another implication is that even perfectly rational individuals may become biased when receiving information from others in society. If individuals include the opinions of others in their judgments, which is of course rational in some situations (e.g., Yaniv, 2004b), they may become biased. Thus, attempts to decrease the nefarious downstream effects of illusory correlations, such as negative attitudes towards minorities, require “bias inoculation” approaches (Turk & Salovey, 1986) that can decrease the emergence and spread of bias at the population level.

Our results may also have implications for metacognition. Although our studies did not measure participants’ confidence in their judgments, we expect that judgment confidence should increase with the progression of generations, particularly in the case of illusory correlations guided by previous expectations. Much like the increase in confidence that is observed when people receive advice that is more consistent with their initial opinions (Schultze, Rakotoarisoa, & Schulz-Hardt, 2015), we expect confidence to increase as generations progress and the illusory correlation becomes closer to people's a priori expectations. Future studies could test this hypothesis.

To be sure, these experiments were not developed to test specific mechanisms and accounts of illusory correlations. Yet, one may, at this point, speculate on the meaning of the results of social amplification paradigms for evaluating different illusory correlations accounts. For example, according to the distinctiveness hypothesis (Hamilton & Gifford, 1976), the illusory correlations are the result of enhanced memory for infrequent behaviors. In line with this, mean estimates of Experiment 1 show an initial overestimation of infrequent behaviors done by the minority group that is kept largely unchanged across generations.

The information loss account (Fiedler, 1991), in turn, assumes more information loss or impaired memory for infrequent than for frequent behaviors (due to the small sample size). More specifically, illusory correlations have been shown to stem from the failure to detect the predominance of positive behaviors in the minority group, but not enhanced memory for negative minority behaviors (Fiedler, Russer, & Gramm, 1993). Although the present studies were, as previously mentioned, not designed to test the contribution of specific accounts for the reported illusory correlations, Experiment 1's results do show what seems to be a failure in detecting the preponderance of Group B positive behaviors. However, the third generation shows higher mean estimated frequency of undesirable behaviors than desirable behaviors for Group B. Such reversal seems hard to explain by a mere phenomenon of information loss.

In sum, the reported pattern of socially amplified illusory correlation is consistent with the idea that the subjective salience of undesirable behavior done by minority groups may be strong enough to counteract information loss effects for these observations across generations. The observed information loss in the social transmission of the frequency estimates concerning the remaining three types of behaviors is such that the relative frequency (and possibly psychological salience) of undesirable behaviors done by minority groups keeps increasing across generations. Future research should further explore this and other potential explanations for the social amplification of illusory correlations.

Limitations. The first limitation that we would like to discuss pertains to the way in which information was transmitted in our studies. When communicating information about associations, people probably tend to summarize their perception (e.g., “Teachers tend to be nice.”), and not provide precise frequency estimates based on exhaustively naming all the cases that they remember. Our methodological choices were meant to replicate classical studies and maximize the standardization of output (all participants either recalled frequencies, in Experiment 1; or recognized each individual instance in a forced recognition paradigm, in Experiment 2). Still, we would expect that more naturalistic methodologies would also reveal a social amplification of illusory correlations. While the verbal descriptors that people probably use in describing differences are coarser and easier to memorize (e.g., “Group A members are nice people.”), the fact that some people form biased opinions and then communicate them to others should lead to an overall increase in populational bias as even people who would be resistant to biases when receiving the original evidence would have a hard time avoiding bias when receiving second hand, distorted information. Future studies can investigate whether the spread and intensity of the social amplification changes as a function of the format of communication.

The second limitation has to do with the minimal social situation created by our paradigm. When receiving and transmitting information across the different generations, participants in our studies were not aware of any characteristics of the source of the information nor any characteristic of the receptor of the information (although they could infer that the source generation and the receptor generation were composed of fellow undergraduate students of their university). While this was a deliberate choice meant to allow an isolated test of the social amplification hypothesis in a minimally social situation, future research can explore potential moderators. For example, people may distort the information in the direction of what they expect to be the recipient's preference (i.e., audience tuning; Higgins, 1992) or instead compensate for perceived sender or recipient bias (i.e., the bias blind spot; Pronin et al., 2002). We suspect that people's liking for certain groups over others, or their expectation about groups that they interact with, or talk about frequently, would further amplify this pattern of results.

Finally, another limitation has to do with the computer-mediated transmission of information. As participants received information passively and were not allowed to interact with the source of the information and ask clarifications, counterexamples, or other questions, the social amplification may result differently when this aspect is included in the paradigm. We expect this factor to only significantly change our results when those who are receiving information become skeptical (which may happen, e.g., when people high in internal motivation to control prejudice receive prejudice-confirming information; Plant & Devine, 1998, or when people have first-hand experience with the data), which, in the context of the current, studies would not be expected as participants only received information from a previous participant and we used arbitrary groups in Experiment 1 and mostly positive attributes of professional groups in Experiment 2. Our choice not to include this aspect was, again, deliberate, as doing so would trigger several factors inherent in social perception and dyadic communication that would add noise to the initial test of our social amplification hypothesis. Future studies can now build upon the current results to show how two-way communication may change our conclusions.

Conclusion. With two studies, we have shown that illusory correlations can be socially amplified, with the illusory correlation becoming stronger as information travels from one person to the next. This finding suggests that previous research has underestimated the size and impact of illusory correlations in society and thus also the importance of finding and implementing ways to unravel this illusion.

4 – Chapter IV – The social amplification of the ratio bias. Mendonça, C., Mata, A., & Ferreira, M. B. (*submitted*)

In everyday life, people are often confronted with information expressed in the form of ratios. This occurs, for instance, in risk-related contexts, such as when dealing with the frequency of side effects of a given medicine (e.g., 1 in 10,000 report feeling nauseous), infant mortality rates (e.g., 14 infant deaths per 1,000 births), or car accident ratios (e.g., 20 deaths per billion vehicle miles). Research has shown that people do not always process this information optimally. Specifically, their judgment and decision making tends to rely more on the numerator and neglect the denominator. In a classic example, Miller, Turnbull, and McFarland (1989) found that participants were more suspicious of cheating behavior when the probability of success involved a ratio with a smaller numerator (1 in 10) than when it involved a ratio with a larger numerator with the same exact probability (10 in 100). This is known as denominator neglect or ratio bias.

Theoretical explanations of this bias abound, a few notable examples being 1) norm theory (Miller et al., 1989), according to which denominator neglect arises because larger than smaller numerators imply a higher number of possibilities for the event to occur, 2) the cognitive-experiential self-theory (Kirkpatrick & Epstein, 1992), according to which the effect is driven by two principles of the experiential information processing system, the concreteness principle (e.g., numerators tend to be smaller numbers and smaller numbers are more concrete) and the experiential principle (e.g., people tend to learn that every 1-in-X ratio where X is a large number means a very unlikely event), and 3) fuzzy-trace theory (Reyna, 1991; Reyna & Brainerd, 2008), according to which the effect is driven by a difficulty in keeping part-whole relationships (such as the numerator and denominator within a ratio) in mind, leading to a focus on numerators instead.

Research on the factors that influence the ratio bias has focused mostly on properties of the problems, such as the magnitude of the difference between the two ratios (e.g., Bonner & Newell, 2010; Denes-Raj, Epstein, & Cole, 1995) or presentation format (e.g., Bonner & Newell, 2010; Garcia-Retamero, Galesic, & Gigerenzer, 2010), and on how people face the problems, such as motivation (e.g., Dale, Rudski, Schwarz, & Smith, 2007; Denes-Raj et al., 1995; Mata, Sherman, Ferreira, & Mendonça, 2015) and perspective taking (e.g., whether people answer how they would solve the problem themselves versus how they think most people would solve it; Alonso & Fernández-Berrocal, 2003; Bonner & Newell, 2010).

A main interest in this area of research is the development of insights that can be used in applied settings to enhance the public understanding of risks (e.g., Garcia-Retamero et al., 2010; Pinto-Prades, Martinez-Perez, & Abellán-Perpiñán, 2006). Yet, an important fact has been neglected in this research that may hinder debiasing efforts: In real-life settings (i.e., outside the lab), information is often transmitted from person to person, and in that process it can be transformed in meaningful ways. In the current paper, we used the serial reproduction paradigm (Bartlett, 1932) to mimic this social dynamic and gain insight into the way information about ratios changes as it is transmitted in society.

In the serial reproduction method, participants form communication chains where a participant's recall of the experimental materials is presented as stimuli to a subsequent participant. Multiple such chains are formed, with participants grouped in generations: first generation participants have direct access to the original materials, second generation participants have access to the information recalled by first-generation participants, etc. This method has provided important insights in domains such as rumor (Allport & Postman, 1947), cultural evolution (Whiten & Mesoudi, 2008), and stereotypes (Lyons & Kashima, 2003).

Our main hypothesis is that, when ratio-bias problems are fed into communication chains, a social amplification of bias will emerge, such that the bias will grow stronger as generations progress and individuals add their own small distortions to an already distorted representation. To test this hypothesis, participants in the current experiment read texts describing various everyday life decisions and indicated their relative preference with regard to two options. Afterwards, participants were asked to recall the texts so that subsequent participants could also read the texts and state their preference. Participants' recall outputs were then transmitted to other participants in linear, one-to-one, communication chains. With this experiment, we aimed to demonstrate that: 1) participants' recall outputs get distorted in a denominator neglect-consistent manner, 2) this distortion accumulates as information travels from one participant to the next, leading to a social amplification of denominator neglect, and 3) this then has an indirect impact on participants' decisions, leading to increasingly suboptimal choices, that is, a social amplification of the ratio bias.

4.1 – Method

Participants

We aimed to collect three generations with 40 participants each, for a total of 120 participants. Each generation was collected serially (i.e., the data for the second generation were only collected when the first generation was completed). In total, 159 participants took part in the experiment. Four participants in the first generation and 32 in the third generation were excluded due to miscommunication with the lab assistants or programming errors that resulted in these participants being assigned to a slot in a chain that was already filled. Another three participants were excluded for providing unusable recall protocols (e.g., the participant commented on the problem instead of recalling it). In every case, only the first participant to be assigned to a valid slot and to provide valid recalls was included in the analysis. The expected 120 valid participants remained after exclusions (105 female, M_{age} = 19.59, SD_{age} = 3.37).

Materials and Procedure

At the beginning of the experiment, participants provided written informed consent, read a brief description of the study's goals and procedure, including the dynamic of receiving information from, and transmitting information to, other participants.

After these initial instructions, the first phase of the experiment consisted of participants reading and giving their opinion about each of four problems. There were two critical ratio-bias problems and two filler problems (i.e., the latter were problems with a single ratio, which were presented in between the critical problems so that there could be variability in the format of the problems, making it harder for participants to realize their structure and memorize their premises in a systematic fashion). The problems were devised to appear to be everyday problems rather than analytical-thinking problems. Each problem was presented in a separate page and was followed by a question asking the participants' opinion, which was expressed on 7-point rating scales (see Table 4 for the problems, questions, and labels of the scales). Four orders of presentation of the problems were created and counterbalanced, with the only constraint being that the ratio-bias problems were always separated by the filler problems.

After participants responded to all four problems, the second phase of the experiment started. In this phase, participants performed two types of recall tasks: a free recall task and a guided number recall. In the free recall task, participants were asked to recall the text. Some minimal details of each story were provided as cues to allow participants to differentiate between the four problems (e.g., "Now, please recall the text in which Nuno tried to choose between two cars."; see Table 4 for all the exact cues given to participants). Instructions also asked participants to include as much information as possible so that the next participant could form his or her own opinion and make a decision. This free recall task provided the material for the serial reproduction.

Table 4. Experimental materials (translated).

Ratio bias 1	
Original text	Nuno is on the market for a new car. He spent a lot of time online searching for a car that would satisfy all his preferences and finally managed to narrow down his options to two cars: Car A and Car B. Both are very similar in terms of price and specifications, including design and performance, even though they have different brands. He managed to find surveys online of two consumer magazines about consumer satisfaction. Regarding Car A, 40 people out of 50 were satisfied with the car and would recommend it to other people. Regarding Car B, 78 out of 100 people were satisfied with the car and would recommend it to other people.
Choice question	If you were in this situation, which car would you prefer to buy?
Choice scale	1 (*Certainly Car A*) to 7 (*Certainly Car B*)
Free recall cue	Now, please recall the text in which Nuno tried to choose between two cars. Describe this story with as much detail as possible so that the next participant may form a judgment and come to a decision on their own.
Guided recall cue	In the text in which Nuno considered choosing between two cars, which number(s) did he find on the surveys about the satisfaction of each cars' owners? Relative to Car A: Relative to Car B:

Ratio bias 2	
Original text	Alexandre is married and tried to have a baby with his wife for a while, but with no success. They then decided to make an appointment in a fertility clinic. The doctor prescribed some analyses and after a few weeks the couple had another appointment where the doctor informed Alexandre that he was infertile, while his wife was fertile. Before they started considering using a donor or adopting, the doctor informed them about two new treatments that would allow Alexandre to conceive a child biologically with his wife. Both treatments were in an experimental phase and no side effects were known, but they were tested in different samples of couples. Treatment X worked in 1 couple out of 100, while treatment Z worked in 9 out of 1000.
Choice question	If you were in this situation, which treatment would you choose?
Choice scale	1 (*Certainly treatment X*) to 7 (*Certainly treatment Z*)
Free recall cue	Now, please recall the text in which Alexandre tried to choose between two fertility treatments. Describe this story with as much detail as possible so that the next participant may form a judgment and come to a decision on their own.
Guided recall cue	In the text in which Alexandre considered choosing between two fertility treatments, which number(s) did the doctor mention regarding the results of the experimental tests of each treatment? Relative to the results of treatment X: Relative to the results of treatment Z:
Filler 1	

Original text	João is the father of two pre-adolescent boys. He and his wife moved to a house very close to the boys' school and discussed allowing them to go to and from school alone. When João himself was a little boy, he frequently went to school alone, even when he was younger than they were, so he was initially in favor of allowing them to go. But his wife told him that 1 in 825 minors are kidnapped every year in Portugal.
Choice question	If you were in this situation, would you allow or not allow that your children went to school alone?
Choice scale	1 (*Certainly would not allow*) to 7 (*Certainly would allow*)
Free recall cue	Now, please recall the text in which João tried to decide whether he should allow his pre-adolescent children to go to school on their own or not. Describe this story with as much detail as possible so that the next participant may form a judgment and come to a decision on their own.
Guided recall cue	In the text in which João considered whether to allow his children to go to school on their own, which number(s) did his wife mention regarding the number of minors that are kidnapped every year in Portugal?

Filler 2

Original text	Luís' younger brother is considering being a police officer after an activity in his secondary school which had the aim to provide information to the students about various professions. Luís was initially very proud of his brother but was then worried when he started thinking about the possibility of his brother dying in service. He then decided to find out the number of police officers who die in service. He found out that 1 in 46,000 police officers die every year in Portugal.

Choice question	If you were in this situation, would you advise your brother to pursue a career as a police officer or not?
Choice scale	1 (*Certainly would advise him to not pursue this career*) to 7 (*Certainly would advise him to pursue this career*)
Free recall cue	Now, please recall the text in which Luís tried to decide whether he should advise his younger brother to pursue a police officer career or not. Describe this story with as much detail as possible so that the next participant may form a judgment and come to a decision on their own.
Guided recall cue	In the text in which Luís considered advising his brother to pursue a police officer career, which number(s) did he find online regarding the number of police officers that die in service in Portugal?

In the guided number recall task, participants first read: "Often, in everyday life we hear and transmit information in the form of concrete numbers. All four stories that you read included numbers. Try to recall these numbers while answering the following questions. If you do not recall the exact number, please try to estimate a number that is as close to the original number as possible." In each trial of this task, participants were again given some details of each story and they were specifically asked to recall the numbers contained in the text (e.g., "In the text in which Nuno considered choosing between two cars, which number(s) did he find on the surveys about the satisfaction of each cars' owners?"). Participants typed their answers into text boxes. Each text box was preceded by a prompt (e.g., in the car story, "Relative to Car A:___" and "Relative to Car B:___"). The guided number recall task was not used in the analysis.[12] Its purpose was to reduce the number of

[12] Using data from the guided number recall task instead of free recall does not change the significance level of any statistical test: Gist distortion in the first generation: $t(39) = -3.60$, $p = .001$; social amplification of gist

invalid free-recall protocols by allowing us to reintroduce the numeric information into protocols that did not include that information. Participants always performed each type of recall task for all four problems in sequence before starting the other type of recall task. The order of these two tasks was counterbalanced.

In terms of the serial reproduction method, participants in the second and third generations did not see the original problems, but instead saw each problem as it was recalled by a participant from the previous generation. Serial reproduction was implemented in a one-to-one manner, such that all protocols of a given participant were given to a single participant in the next generation. Spelling mistakes were corrected, and in the cases where participants did not spontaneously recall any numeric information, participants' answers to the guided number recall task were used to provide that missing information. This editing process was used for 13.75% of the first generation recall protocols and 18.75% for the second generation.

4.2 – Results

Ratio Bias

The frequency distribution of participants' preferences, per generation, can be found in Table 5. Considering only the participants who saw the original version of the problems, most participants revealed at least some preference for the normative option, with 18.75% of participants, on average, revealing a suboptimal preference. Previous studies (Ferreira, Mata, Donkin, Sherman, & Ihmels, 2016; Mevel et al., 2015) have found similarly low levels of suboptimal choices. Means and standard deviations for option preferences per generation can be seen in Table 6.

distortion: $r(118) = -.21$, $p = .021$; social amplification of the ratio bias effect: indirect $b = 0.10$, 95% CI [0.01, .026].

Table 5. Preference distribution in the ratio-bias problems, from 1 – complete preference for the normative option, to 7 – complete preference for the non-normative option.

	Preference distribution						
Scale	1	2	3	4	5	6	7
	Ratio bias problem 1 – Two cars						
Generation 1	15	10	5	2	5	2	1
Generation 2	10	11	4	8	1	6	0
Generation 3	11	7	4	2	3	5	8
	Ratio bias problem 2 – Two treatments						
Generation 1	14	8	8	3	2	0	5
Generation 2	12	9	5	3	4	3	4
Generation 3	21	7	5	4	0	1	2

Table 6. Means (and standard deviations), per generation, for preferences and gist distortion (varying between -1 – distortion to 1 – no distortion).

	Preference		Gist distortion	
	M (*SD*)	*N*	*M* (*SD*)	*N*
Generation 1	2.66 (1.36)	40	0.69 (0.57)	39
Generation 2	3.00 (1.57)	40	0.68 (0.62)	38
Generation 3	2.90 (1.51)	40	0.33 (0.81)	39

Gist Distortion

To include as much data as possible, regardless of distortions in format (e.g., some participants recalled absolute frequencies instead of ratios, others did not include numeric information but only described which option was better), recall protocols were coded in terms of gist distortion. Specifically, recall protocols where the gist favored the originally normative option were coded as 1 (e.g., "[…] car A satisfies more people and car B satisfies fewer people." or "[…] 40 in each 50 people recommended car A, and 78 in 100 people recommended car B"); and recall protocols where the gist favored the originally non-normative option were coded as -1 (e.g., "[…] 40/50 people bought car A and 87/100 bought car B"). There were no free recall protocols where both options were deemed equally favorable. Mean and standard deviations for this gist score per generation can be seen in Table 6.

In the first generation, the average gist score already differed significantly from accuracy standards, *t*(38) = -3.38, *p* = .002, *d* = -0.54, 95% CI [-0.88, -0.21], with a significant number of participants representing the relationship between the two options incorrectly: 17.14% of all valid recall protocols in the first generation were distorted in favor of the non-normative option.

Social Amplification of Gist Distortion

Using the gist score, we then tested whether there was a social amplification of distortion, such that as information was transmitted from one participant to the next, the gist distortion became intensified. Indeed, there was an increase in gist distortion as generations progressed, *r*(114) = -.22, *p* = .021, 95% CI [-.39, -.03].

Social Amplification of the Ratio Bias

As for our second social amplification hypothesis, we expected gist distortions to have an impact on participants' preferences. Using model 4 of the PROCESS macro (Hayes, 2013) with 10,000 bootstrap samples, we tested whether, as generations progressed, there was an increase in gist distortions favoring the ratio bias, which would then lead to an increase in preferences consistent with the ratio bias. This indirect effect was significant, *b* = 0.16, *SE* = .08, 95% CI [0.02, 0.34] (see Figure 5 for a visual representation of the mediation model).

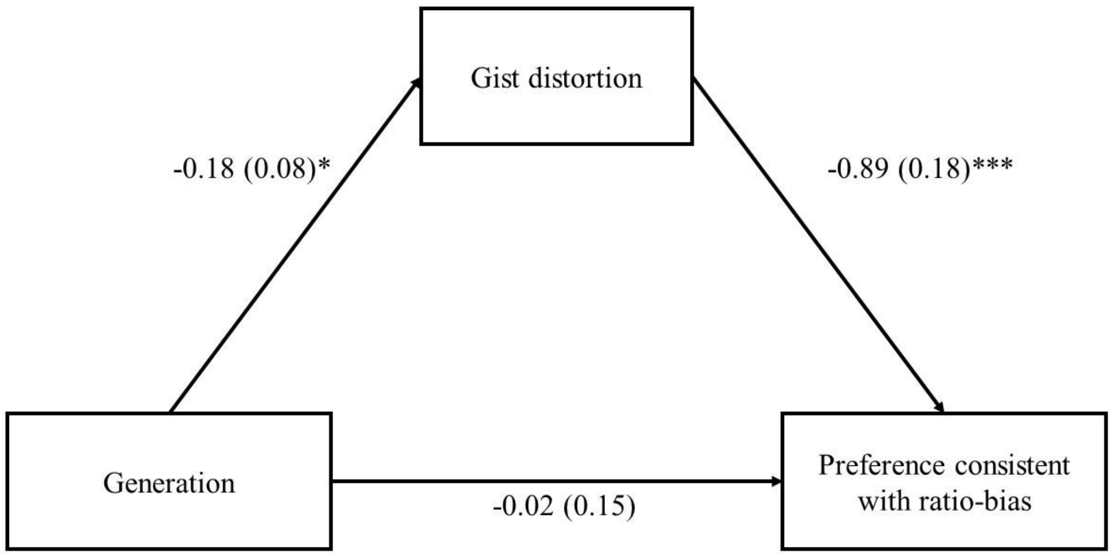


Figure 5. Mediation model for the indirect effect of generation on preferences, via gist distortion (varying between -1 – distortion to 1 – no distortion); * = p < .05; *** = p < .001.

4.3 – Discussion

In this experiment, ratio-bias problems were transmitted from one participant to another. We found that, as the problems were communicated through chains of participants, the ratios became increasingly distorted in a manner consistent with denominator neglect, favoring the suboptimal option with the initially larger numerator, but smaller probability. This amplification of gist distortion in turn led participants to increasingly prefer the suboptimal choices (i.e., the ratio bias).

Recent research has questioned whether the ratio bias, as documented in previous research (e.g., Epstein & Pacini, 2001; Kirkpatrick & Epstein, 1992), is a real judgment bias or merely an artifact of instructions that invite participants to express their intuitions and counteract their desire to appear rational. In fact, when participants are instructed to respond from their own perspective (as in the current experiment) the level of bias is often vestigial (e.g., Ferreira et al., 2016; Mevel et al., 2015), which has led some researchers to question the practical importance of this bias (e.g., Lefebvre, Vieider, & Villeval, 2011).

Our results, however, suggest that even if denominator neglect and ratio bias are low in the first generation of reasoners who are presented with a problem, they tend to be socially amplified in the following generations. In this way, even a few initial distorted representations and suboptimal judgments may have a significant impact as they grow across generations. Thus, our results suggest that the effects of denominator neglect in society may have been underestimated in previous research, as the fact that people receive and transmit information to others should lead to an intensification of the bias. This research shows that denominator neglect produces distorted representations of the original ratios. Once these biased ratios are socially transmitted, the seeds of social amplification of bias are sown and the next generation is destined to fail, as even fully rational people with perfect memory will transmit the bias as they (correctly) respond to the (incorrect) representations of the original ratio problems.

Different debiasing techniques have been suggested to eradicate or mitigate denominator neglect and the ratio bias: 1) Expressing ratios in a manner that favors the desired outcome: For example, if the population underestimates a certain risk, it can be communicated with a wide range (e.g., 1,000 deaths by suicide in the national population of ten millions); if instead the population overestimates a risk, it can be communicated with a narrow range (e.g., one death by suicide per ten thousand national residents; Pinto-Prades et al., 2006); 2) Using visual displays instead of verbal descriptions (e.g., Reyna, 1991); And 3) using a standardized format of risk communication, such as the micromort (Ahmad, Peterson, & Torella, 2015). Future research may investigate which method is better suited to attenuate the social amplification pattern observed in this study, keeping in mind that some strategies are ineffective as information becomes second-hand. For instance, visual displays may not be successful after the first generation, as second-generation participants receive verbal descriptions and no longer have access to the visual displays.

In conclusion, the present study suggests that, because information in society is often transmitted from one person to another, denominator neglect may have a greater impact in society than could be expected from one-generation studies. These results call for further research on ways to counteract this bias and its social amplification.

5 – Chapter V – The collective erosion of deliberation: Social transmission of reasoning problems creates biases. Mata, A., Mendonça, C.,[13] & Mascarenhas, M. (*invited for resubmission* – Journal of Experimental Psychology: Learning, Memory, and Cognition)

Research following the heuristics-and-biases tradition of judgment and reasoning often uses tricky problems that trigger a response that is highly intuitive, although incorrect (e.g., Kahneman, 2011; Stanovich, 1999). Take, for instance, the popular bat-and-ball problem: A bat and a ball together cost 110 cents. The bat costs 100 cents more than the ball. How much does the ball cost? (Frederick, 2005). Recent studies show that people often fail to solve such problems not necessarily because they lack the ability to deliberate carefully about them, but rather because they ignore or misrepresent information in the premises that is crucial to fully comprehend the problem (the critical premise is that the bat costs 100 cents *more than the ball*). The evidence for this comes from studies tracking visual attention (Mata et al., 2017), as well as change-detection studies assessing the ability to tell the difference between the original version of a reasoning problem, which poses a conflict between the intuitive but incorrect response and the deliberative correct response, and a slightly modified version of the same problem without the conflict (Mata et al., 2014). Thus, errors that are often attributed to faulty reasoning might emerge earlier. Indeed, even if responders are able to reason deliberatively, they will fail to produce the correct response if they do not represent a problem correctly.

As a first goal, this research aims to further test this relation between problem-solving performance and the accuracy of representation of the premises. Whereas previous research used change detection (Mata et al., 2014) and eye-tracking (Mata et al., 2017), this research

[13] A. Mata and C. Mendonça share co-first authorship of this work.

uses free recall to assess people's representation of the premises (De Neys & Glumicic, 2008). The use of memory as a tool to measure the accuracy of a representation has proven useful in other areas, such as developmental psychology, where it has also been shown that problem solving difficulties sometimes arise not when people are attempting to solve a problem but rather when they are initially representing it (Fazio, Kennedy, & Siegler, 2016; Laski & Siegler, 2014; McNeil & Alibali, 2004; Siegler, 1976). Here too we expect to observe memory dependence, whereby the ability to solve a problem depends in part on how reasoners interpret it, as measured by how accurately they recall its critical premises.

More importantly, this research explores the downstream effects of these (mis)representations when people communicate about the problems with others. People share information, and often they do not have first-hand information, but rather receive it from others. As demonstrated in classic memory research, when people transmit information in communication chains, they introduce systematic distortions (Bartlett, 1932). We adapted this approach to investigate how communicating about reasoning problems might distort them in systematic and consequential ways. Specifically, we predicted that problem-solvers (incorrect responders in particular) often fail to represent the critical conflict-relevant premises accurately. As a consequence, when they communicate the problem to other problem-solvers, they will reproduce it in a distorted fashion, without those premises. Eventually, the problem-solvers in the next generations will converge towards a more simplistic representation of the problem, such that what started out as a challenging problem where intuition was at odds with deliberation eventually turns into a much simpler problem where there is no longer a conflict to be resolved.

Let us again use the bat-and-ball problem to illustrate this: If reasoners do not pay attention to the "more than the ball" part of the second premise, then they will understand and in turn communicate to others a much simpler version of the problem: "A bat and a ball together cost 110 cents. The bat costs 100 cents. How much does the ball cost?" In this version, the intuitive solution (10 cents) is correct. In fact, it is the only possible response, but it no longer requires effortful deliberation. Thus, reasoning in later generations should become less elaborate than in earlier ones, not because problem-solvers are less able, but simply because they are working with distorted premises where the critical conflict-relevant information is missing.

Finally, we sought to test yet another hypothesis pertaining to the effect of this collective misrepresentation of problems on reasoning: We tested whether misrepresenting a problem makes it harder for reasoners to solve the problem when they see it again in its original version, because their previous misrepresentation biases their subsequent interpretation. If so, the incorrect representations that reasoners form about reasoning problems would have further downstream effects: not only would they hinder reasoning when they first see a problem for the first time (perhaps a distorted version of the problem, misrepresented by others before), but they would also compromise their ability to think properly about the problem when it is shown again in its original version with all the critical information available.

Overview. We adapted the serial reproduction paradigm (Bartlett, 1932; Kashima, 2000) to study the social communication of reasoning problems. Participants were distributed across three generations: The first generation received the original versions of reasoning problems, solved them, and then recalled their premises. Participants in the next generations received the problems as they were recalled by participants in the previous generations, and in turn solved and recalled them. Moreover, at the end of each trial, participants in each generation saw the problem again in its original version.

The hypotheses are the following: First, sound reasoning should hinge on the accurate representation of the critical premise, such that incorrect responders should be worse at recalling the conflict-relevant premises (we termed this: memory dependence). In addition, and to test how specific this effect is, we categorized the different memory errors that participants made as a function of their location (i.e., which premise they occurred in) and their type: distortion (i.e., transforming the information in a premise) or filtering (i.e., forgetting the information entirely). According to previous research (Mata et al., 2014, 2017), the error that should be most detrimental for deliberation should be the filtering of the critical premise. Second, we predict that this misrepresentation of the problems will be amplified as participants transmit the problems to other participants, that is, the misrecall of conflict-relevant premises will increase over generations (filtering of conflict). Third, this will have consequences for reasoning performance, such that, as conflict gets filtered in the process of communicating problems across generations, the amount of deliberative responses to the presented problems will decrease (erosion of deliberation). And fourth, the effects of this erosion might be so detrimental that, when reasoners are subsequently shown the original (conflict-preserved) version of a problem, they will more likely respond incorrectly if they have misrepresented it earlier (corruption of subsequent reasoning).

5.1 – Experiment 1

5.1.1 – Method

Participants. We created 30 chains of three participants each, for a total of 90 participants (30 women, 60 men, $M_{\text{age}} = 34.18$, $SD_{\text{age}} = 11.87$), recruited through the Prolific platform. Participants were native English speakers and had a Prolific score of 90/100 or higher.

A total of 105 participants completed the experiment. Of these, three were excluded and replaced by different participants as all their recalls were invalid. When participants only gave one or two invalid recalls, we recruited extra participants in order to always transmit three problems to the next generation. Twelve of such “donors” participated in the experiment to provide valid recalls but were not included in analysis. After excluding these 15 participants, 90 valid participants remained and were considered in the analysis.

Materials. The three problems (adapted from Mata et al., 2017) followed the same logic as the bat-and-ball problem (see Table 7).

Table 7. Problems used in Experiment 1 (conflict-relevant premises are highlighted in bold).

Problem	Problem description	Intuitive answer	Deliberative answer
1	"A TV and a DVD are on sale. Together they cost 110 cents. The TV costs 100 cents **more than the DVD**. How much does the DVD cost?"	10	5
2	"Anna and Sophie are collecting shells. Together they found 12 shells. Anna found 10 shells **more than Sophie**. How many shells did Sophie find?"	2	1
3	"A rock climber ties a long rope to a short rope. Together the two ropes measure 88 meters. The long rope is 80 meters **longer than the short rope**. How much does the short rope measure?"	8	4

Second- and third-generation participants received the problems as they were recalled by participants in the previous generation. We used a one-to-one transmission, whereby all three problems recalled by a given participant were passed on to another participant in the next generation. Recall protocols that were ambiguous or insufficiently specified (e.g., "Tv cost 100 more, both 110"), or that did not reproduce the problem at all (e.g., instead of recalling the problem, the participant explained the reasoning behind his or her response) were considered invalid. These invalid recalls were replaced by valid recalls taken from donor participants (i.e., participants who were not part of a chain but whose function was to generate recall protocols to replace invalid recalls of participants who did belong to chains). A total of 21 recalls (12 in the first generation, 11 in the second generation) out of the 180 used were considered invalid and were replaced by using donor protocols (as the recalls of the third generation were not transmitted to other participants, no invalid recalls were replaced there).

Editing of recall protocols (before they were used as premises in the next generation) was kept to a minimum: When the question (e.g., How much does X cost?) was not included in the protocol, it was reintroduced. Comments and responses to the problem were removed. Spelling or grammatical errors were corrected, as were small conflict-irrelevant inconsistencies (e.g., "Two girls are collecting shells. Together they found 12 shares").

Procedure. For each problem, first, participants read the instructions: "Please answer the following problems" and solved the problem. This was either the original problem (Generation 1), or the problem as it was recalled by another participant (Generations 2-3). Then, the problem was removed from the screen, and participants were asked to recall it. They were instructed that what they recalled would be transmitted to other participants. Finally, all participants read the following instructions:

> "Now you will see the same problem again or a slightly different version of this problem. We ask you to answer the problem. You may give the same answer as before, if you think that the problem is the same as before. Or you may give a different answer, if you think that the problem is different."

Participants then solved the original version of the problem (in Generation 1, this was the same as in the first step).

5.1.2 – Results

The means for all dependent variables per generation are shown in Figure 6.

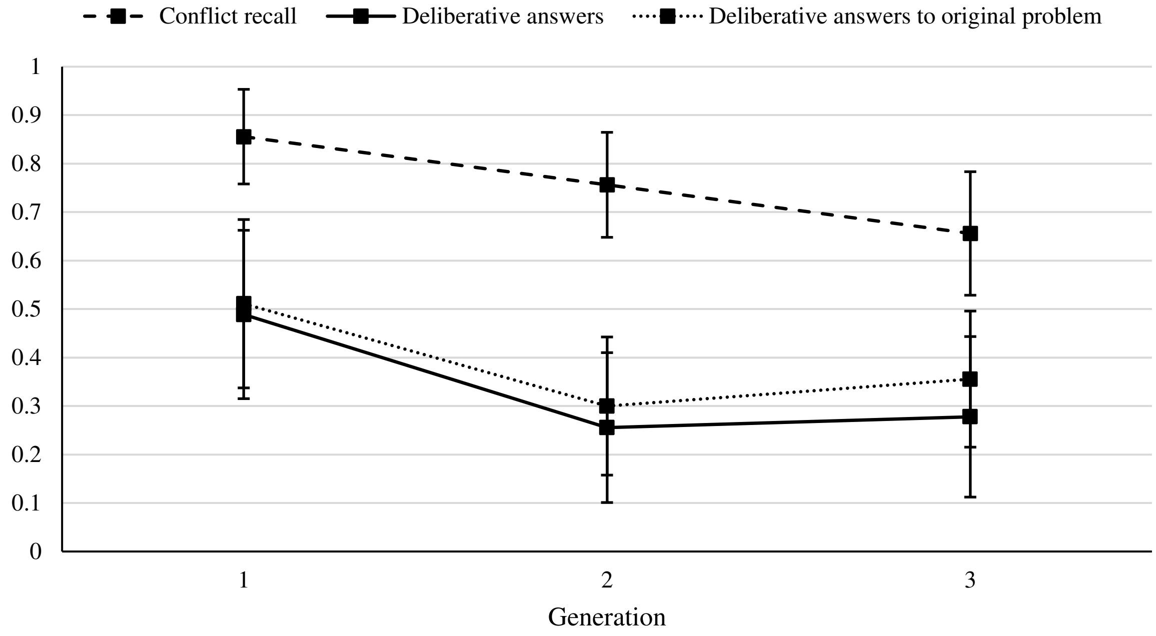


Figure 6. Proportion of conflict recall, deliberative answers (first trial), and deliberative answers given to the original version of problems (second trial) with 95% CI error bars (Experiment 1).

Memory Dependence. Responses to the problems were coded as 1 if they corresponded to the deliberative solution to the original version of the problem (see Table 7 for the deliberative answers), and 0 if not. Recall protocols were coded as 1 if they contained the critical conflict-relevant premise (see Table 7; an example from a participant: "A TV and DVD are on sale. In total, they cost 110 cents. The TV costs 100 cents more than the DVD. How much does the DVD cost?"), and 0 if not (e.g., "A TV and DVD is for sale for 110 cents. The TV costs 100 cents, how much does the DVD cost?"). The proportion of deliberative responses correlated with the number of recalled problems that preserved the critical premise, $r(88) = .44$, $p < . 001$.

Filtering Conflict. Consistent with the second hypothesis, recall performance (i.e., the proportion of recalled problems that preserved the critical premise) decreased across generations, $r(88) = -.26$, $p = .015$.

Erosion of Deliberation. We then tested whether this progressive filtering of conflict results in worse performance (when participants first encounter the problems) over generations. Specifically, using a percentile bootstrapping procedure (10000 samples, PROCESS v3.2 for SPSS; Hayes, 2018) we tested a mediation model whereby the proportion of deliberative responses (categorized as the correct responses to the original problems; see Table 7) decreases over generations because of how participants represented the problems (i.e., recall). The indirect effect was significant: *b* = -0.06, *SE* = 0.02, 95% CI [-0.11, -0.01] (see Figure 7).

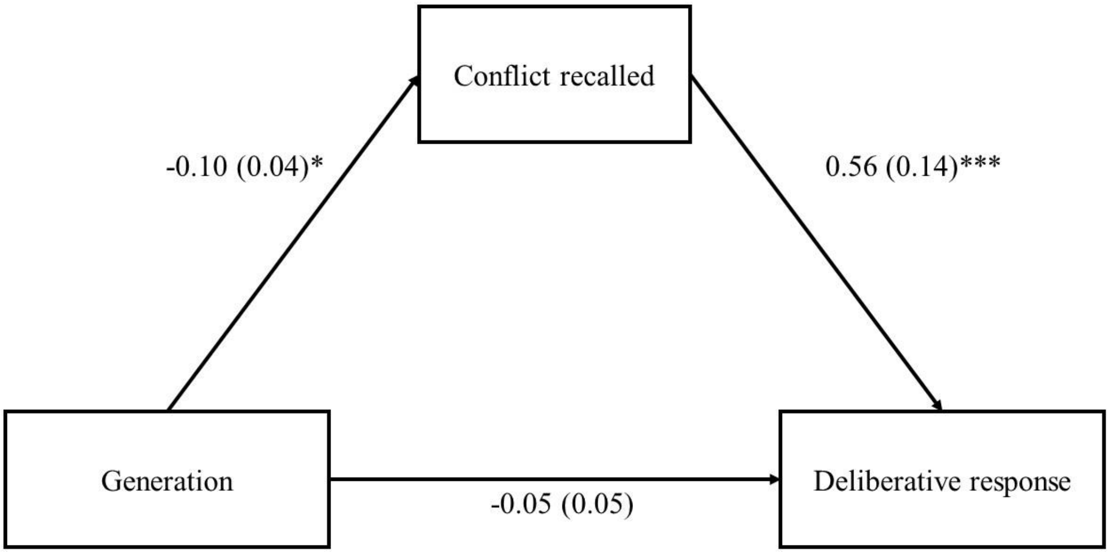


Figure 7. Unstandardized regression coefficients (with respective standard errors inside brackets) for Experiment 1's mediation model of the impact of generation on deliberative responses via the proportion of problems recalled with conflict; * = *p* < .05; *** = *p* < .001.

Corruption of Subsequent Reasoning. Another predicted effect of misrepresenting a problem on reasoning is that, if participants misrepresent a problem, they will subsequently fail to solve it, even if the problem is presented in its original conflict-intact version. Thus, we tested whether the effects of conflict filtering could also be observed in the responses that participants gave to the original problems that they saw at the end of each trial, after they had already solved and recalled a version of the problem. A first correlational analysis shows that recall performance correlated with reasoning performance (i.e., the proportion of correct deliberative responses) when the problems were presented for the second time, in their original version, *r*(88) = .37, *p* < .001. Moreover, a mediational analysis using the procedure described above reveals a significant indirect effect of generation on this reasoning performance via recall: *b* = -0.05, *SE* = 0.03, 95% CI [-0.11, -0.01] (see Figure 8).

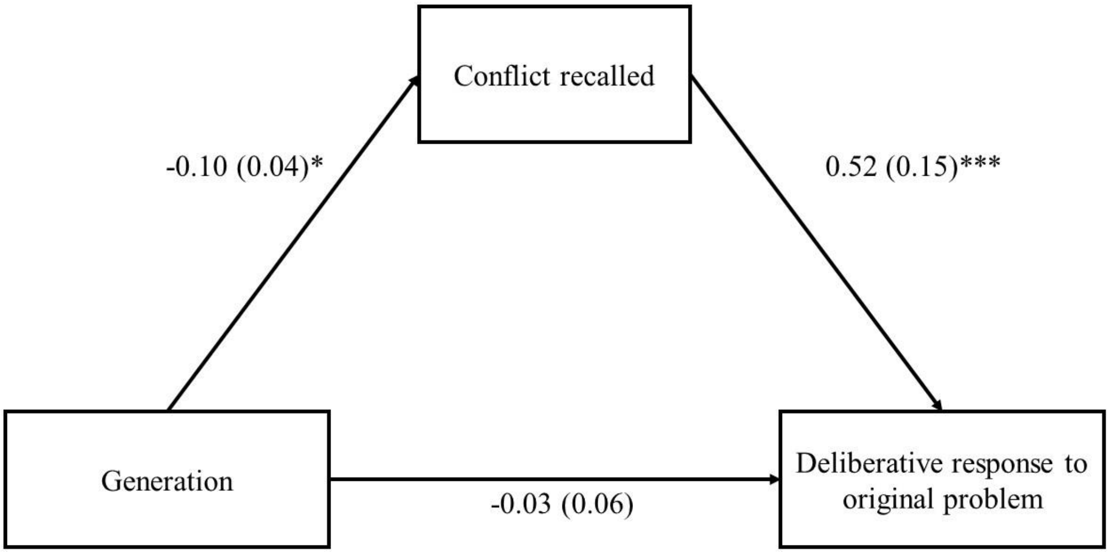


Figure 8. Unstandardized regression coefficients (with respective standard errors inside brackets) for Experiment 1's mediation model of the impact of generation on deliberative responses to the original version of the problems via the proportion of problems recalled with conflict; * = *p* < .05; *** = *p* < .001.

Types of Memory Error. Finally, we analyzed the memory errors according to premise location and type (distortion versus filtering). Our coding scheme broke down each sentence into simple elements. For example, the first sentence in problem 1, "A TV and a DVD are on sale", was broken down into three elements: TV, DVD, on sale. We coded whether each element was distorted (e.g., "DVD player" or naming a different object) or filtered out (i.e., not mentioning the DVD). We then calculated the proportion of distorted and filtered elements in each premise. And in a regression analysis, we used the filtering and distortion of each premise to predict the proportion of deliberative answers. As expected, the filtering of the critical premise led to a significant decrease in reasoning performance, *b* = -0.99, *SE* = 0.38, *p* = .011. Of all the other seven possible errors (4 premises * 2 types of error), only the distortion of the second premise had a significant impact as well, *b* = -0.97, *SE* = 0.41, *p* = .020.

Using the same coding to analyze the impact of filtering and distortion on deliberation when the problems were presented a second time (i.e., in their original versions), the same pattern was observed: filtering of the critical premise, *b* = -1.01, *SE* = 0.42, *p* = .018; distortion of the second premise, *b* = -0.94, *SE* = 0.45, *p* = .039; the other predictors were not significant.

5.2 – Experiment 2

Experiment 2 sought to replicate the results of Experiment 1 with the following differences: A more varied set of problems was used. Sample size increased from 30 to 50 three-person chains (total *N* = 150), providing 90% power to detect the smallest effect in Experiment 1 (filtering conflict: *r*(88) = -.26). And we altered the way that participants' recalls were transmitted from one generation to the next. In Experiment 1 we used a one-to-one transmission, such that all the problems that a participant received were recalled by the same person in the previous generation. In Experiment 2 we used a many-to-one transmission, such that a participant could receive each problem from a different person, to eliminate potential source effects. Moreover, this communication setup comes closer to the actual way in which people transmit to, and receive information from, different sources.

5.2.1 – Method

Participants. We aimed for 50 three-participant chains, for a total of 150 participants (86 women, 62 men, M_{age} = 33.21, SD_{age} = 10.42), recruited through Prolific (the criteria for participant selection were the same as in Experiment 1).

In Experiment 1, the recall protocols of some participants were suspiciously identical to the problems that they had read, raising the possibility that some participants copy-pasted the problems. In Experiment 2 we developed a strategy to detect this: An extra space was added between two words of the premises. Participants whose recall protocol was an exact copy of the problem that they read, including that extra space, were excluded from the analysis.

A total of 236 participants took part in the experiment. Of these, 47 were excluded because at least one of their recalls met our copy-paste detection criterion. Eighteen participants were excluded because all their recalls were invalid using the same criteria as in Experiment 1, four were excluded because the URL query which assigned participants to the correct chain did not function properly (the URL query became empty, causing the experiment to malfunction), and seven participants were assigned by experimenter error to an already-completed chain (only the first participant to start the experiment in a given chain was considered). Finally, a total of ten participants participated in the experiment only to provide valid recalls to replace invalid ones. After these 86 participants were excluded, 150 valid participants remained, forming 50 chains of three participants each.

Materials. There were nine problems: five conflict problems (see Table 8) and four no-conflict filler problems (Thomson & Oppenheimer, 2016). These filler problems are not considered in the analysis, as they were always presented in their original form, and were used only to lower participants' suspicion about the tricky nature of the problems, which might suppress spontaneous comprehension and reasoning.

Table 8. Problems used in Experiment 2.

Problem	Problem description	Intuitive answer	Deliberative answer
1	“A jacket and a tie together cost 44 dollars. The jacket costs 40 dollars **more than the tie**. How much does the tie cost?” (adapted from Frederick, 2005)	4	2
2	“A computer virus is spreading through the system of a computer. Every minute, **the number of infected files doubles**. If it takes 100 minutes for the virus to infect all of the system, how long would it take for the virus to infect half of the system?” (adapted from Frederick, 2005)	50	99
3	“Ellen and Kim are running around a track. **They run equally fast** but Ellen starter later. When Ellen has run 5 laps, Kim has run 15 laps. When Ellen has run 30 laps, how many has Kim run? ___ laps” (Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2016)	90	40
4	“A farmer had 15 sheep and **all but** 8 died. How many are left? (Thomson & Oppenheimer, 2016)”	7	8
5	“**You have a book of matches** and enter a cold, dark room. You know that in the room there is an oil lamp, a candle, and a heater. What do you light first?” (Boland, 2013)	Any of the set of three	Matches

Note. The critical conflict-relevant premises are highlighted in bold.

As in Experiment 1, second- and third-generation participants received the problems as recalled by participants of the previous generation. In this experiment, we used a many-to-one transmission instead of a one-to-one transmission. To achieve this, we created a pool with all valid recalls of a given problem reproduced by the previous generation and then randomized (without replacement) the assignment of problems, such that a given participant could receive each of the five problems from five different participants.

A total of 21 protocols (15 in the first generation, 6 in the second generation) out of 500 were considered invalid (by the same criteria as in Experiment 1) and replaced using donors.

Editing of recall protocols was kept to a minimum, using the same criteria as in Experiment 1.

Procedure. The procedure was the same as in Experiment 1.

5.2.2 – Results

The means for all dependent variables per generation are shown in Figure 9.

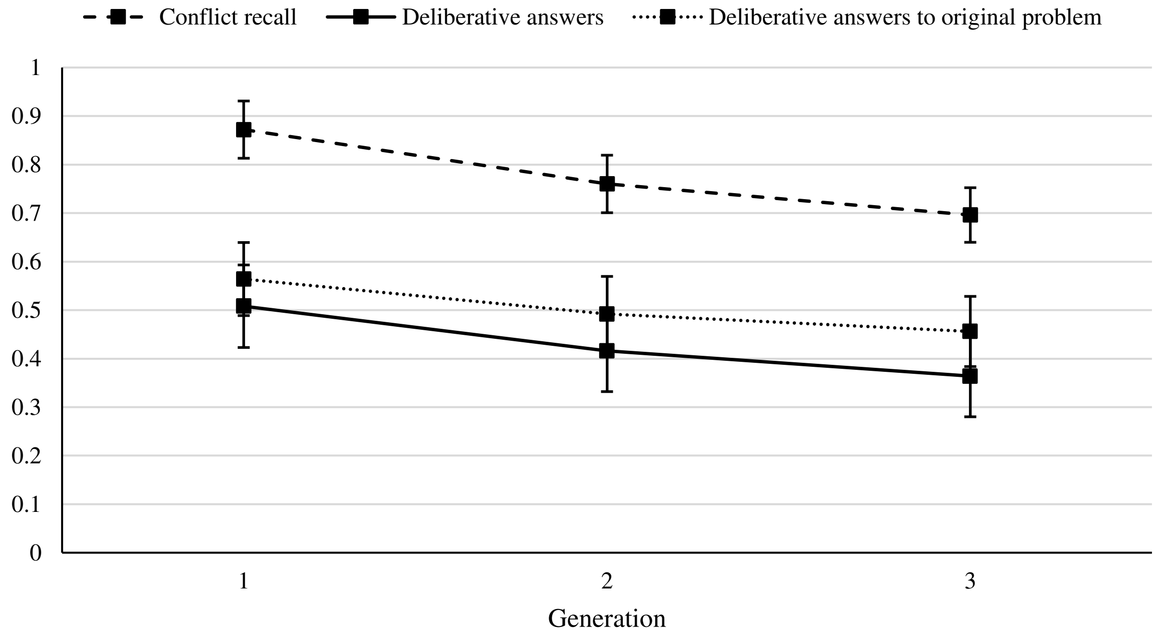


Figure 9. Proportion of conflict recall, deliberative answers (first trial), and deliberative answers given to the original version of problems (second trial) with 95% CI error bars (Experiment 2).

Memory Dependence. Responses to the problems and recall protocols were coded as in Experiment 1. The proportion of deliberative responses correlated with the proportion of recall protocols that preserved the critical premise, *r*(148) = .59, *p* < .001.

Conflict Filtering. The proportion of conflict-sensitive recall protocols decreased across generations, *r*(148) = -.33, *p* < .001.

Erosion of Deliberation. In turn, this progressive misrepresentation of the problems led to more simple solutions across generations. A mediational analysis using the same process as in Experiment 1 revealed a significant indirect effect of generation on the proportion of deliberative responses mediated by the proportion of recall protocols that contained the conflict premise: *b* = -0.06, *SE* = 0.02, 95% CI [-0.10, -0.03] (see Figure 10).

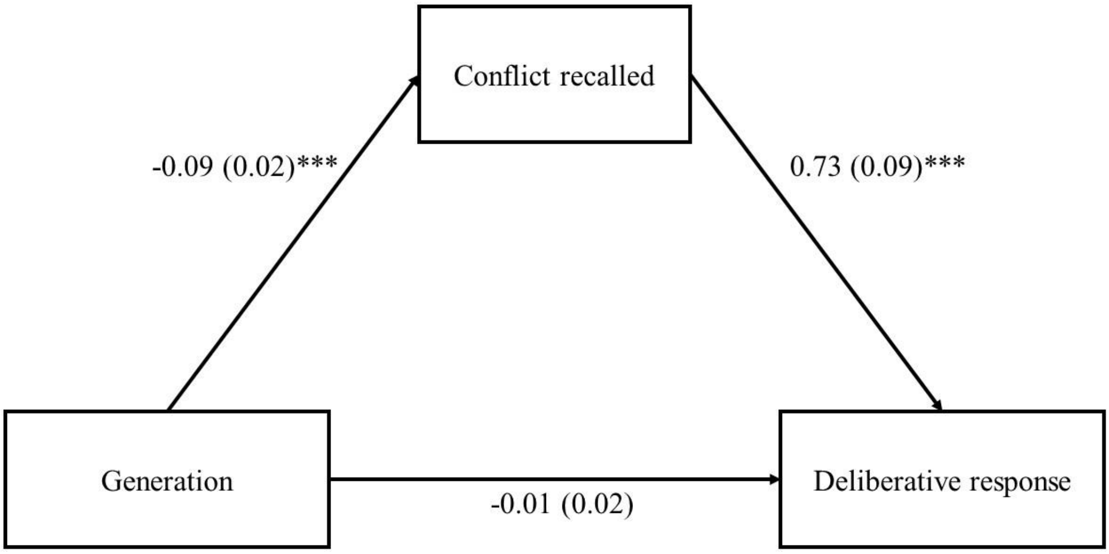


Figure 10. Unstandardized regression coefficients (with respective standard errors inside brackets) for Experiment 2's mediation model of the impact of generation on deliberative responses via the proportion of problems recalled with conflict; *** = *p* < .001.

Corruption of Subsequent Reasoning. Once again, when participants represented the problem without the conflict premises (in their recall protocols), they were less likely to solve the problem correctly when they subsequently saw it in its original conflict version:

Recall performance correlated with reasoning performance when the problems were presented again, in their original version, *r*(148) = .47, *p* < .001. And there is a significant indirect effect of generation on this final reasoning performance mediated by recall: *b* = -0.06, *SE* = 0.02, 95% CI [-0.09, -0.03] (see Figure 11).

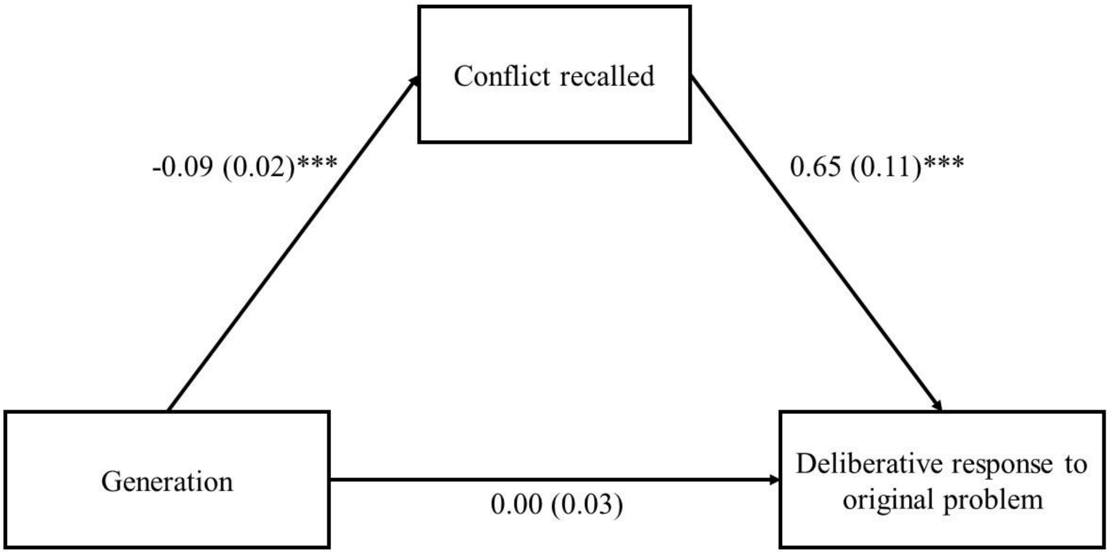


Figure 11. Unstandardized regression coefficients (with respective standard errors inside brackets) for Experiment 2's mediation model of the impact of generation on deliberative responses to the original version of the problems via the proportion of problems recalled with conflict; *** = *p* < .001.

Types of Memory Error. Using the same coding scheme as in Experiment 1, we analyzed the specific impact of filtering and distortion across premises. In this case, it was not possible to average across problems, as some had more premises than others, and the position of the critical premise also varied (see Table 8). Therefore, the results for each problem are presented separately. The filtering of the critical premise had a significant effect on deliberation for four out of the five problems: problem 1, *b* = -0.78, *SE* = 0.26, *p* = .003; problem 2, *b* = -2.98, *SE* = 0.30, *p* < .001; problem 3, *b* = -0.42, *SE* = 0.14, *p* = .003; problem 4, *b* = -0.53, *SE* = 0.10, *p* < .001; only in problem 5 was this effect not significant, *b* = -0.15, *SE* = 0.21, *p* = .477. The only other significant effects, out of the other 25 possible errors, were the filtering of the non-critical premise 2 in problem 2, *b* = 1.40, *SE* = 0.34, *p* < .001, and the distortion of the non-critical premise 3 in problem 5, *b* = -0.82, *SE* = 0.30, *p* = .007.

Using the same coding to analyze the impact of specific memory errors on deliberation when the problems for the second time (i.e., in their original versions), a similar pattern was observed. The effect of the filtering of the critical premise was significant for most problems: problem 2, *b* = -2.00, *SE* = 0.31, *p* < .001; problem 3, *b* = -0.32, *SE* = 0.15, *p* = .034; problem 4, *b* = -0.27, *SE* = 0.11, *p* = .014; though not significant for problem 1, *b* = -0.36, *SE* = 0.28, *p* = .201, and problem 5, *b* = -0.16, *SE* = 0.21, *p* = .451. The only other significant effects were the filtering of the non-critical premise 2 in problem 2, *b* = 1.18, *SE* = 0.34, *p* = .001, and the distortion of the non-critical premise 3 in problem 5, *b* = -0.88, *SE* = 0.31, *p* = .005.

5.3 – General Discussion

Research on judgment and reasoning has mostly focused on single-generation studies, whereby all participants are shown the same experimenter-controlled premises. This control is essential when studying how people reason about problems with certain key characteristics that are relevant to the experimenter. However, this approach misses two critical facts: First, participants might not represent the problem as intended by the experimenter, as shown in the present studies. And more importantly, reasoning in the wild (i.e., outside the lab) does not operate on carefully controlled information, but rather on problems that are passed on from person to person. This research shows that, in this process of social transmission, problems get distorted in systematic ways. Specifically, problems that started out as challenging, with critical premises that pose a conflict between easy intuitive solutions and harder deliberative ones, gradually lost these challenging features. Moreover, and as a result of this filtering process, the responses that people gave to those problems were less likely to reveal signs of sophisticated deliberation. Indeed, any individual differences in analytical ability or thinking styles that there may have existed in the first generation necessarily lost their predictive power as the collective understanding of the problems became distorted and over-simplified across generations.

Whereas some social amplification processes are random and unpredictable (Salganik et al., 2006), the amplification effects observed here are 1) systematic (i.e., there is a consistent distortion in the premises that turns conflict problems into no-conflict problems), 2) irreversible (i.e., once such distortions have occurred, they are unlikely to be undone by noise or variations that occur as problems are communicated), and 3) lead to predictable downstream effects (the corruption of subsequent reasoning).

This work makes several contributions: The first pertains to the social dynamics aspect of reasoning, and how interpersonal communication can affect it. Recent work has started to explore situations where people try to solve this kind of problems in groups (Trouche et al., 2014), and how people reproduce the solutions of these problems (Claidière et al., 2017), but a key element was missing from that research: how the problems are communicated to people by others, and how they in turn convey them to others. Indeed, though related, that work is different from the present research where we studied how people communicate (and eventually distort) the problems themselves. Before people get to discuss solutions to a problem, they need to communicate its premises. In those studies, the experimenters controlled the presentation of the premises, so that the problems were always presented in their original conflict-preserved version. Moreover, their conclusions were optimistic, such that argumentation was shown to promote the transmission of the correct, counterintuitive solutions. But outside the lab, when people transmit the problems to others, errors of attention and comprehension might compromise how the problems are conveyed, with downstream consequences. In this case, argumentation might be of less help.

Moreover, the methods used in this research bring together different areas of psychological research. In particular, serial reproduction has been used to study phenomena as diverse as stereotyping, rumor and risk perception (Allport & Postman, 1947; Kashima, 2000; Moussaïd et al., 2015). And the use of memory as a way to assess how accurately people represent problems has been used in developmental psychology and education (Fazio et al., 2016; Laski & Siegler, 2014; McNeil & Alibali, 2004; Siegler, 1976).

In addition, these results suggest a novel effect of misrepresenting the conflict in a problem: When participants first formed a simplistic representation of a problem, disregarding the critical conflict-relevant premises, they became less apt at solving the problem subsequently, when the problem was presented again in its original conflict-preserved version. Thus, forming an incorrect interpretation of the problem compromises reasoners' ability to correctly understand it and detect the conflict in it in subsequent attempts to solve it. This is a novel effect in the literature that has consequences for the possibility of debiasing: Research has already shown that the way people interpret a problem influences whether they are able to solve it correctly. These studies suggest that it can also bias the way in which people interpret the problem thenceforth.

Moreover, our results suggest representational dependence, which is relevant for the debate on whether having precise memory for premises is essential for good reasoning (Brainerd & Reyna, 1992a, 1992b; M. Chapman & Lindenberger, 1992a, 1992b). Our results suggest that forming an accurate representation of certain critical premises in conflict problems is essential for sound reasoning to occur. Furthermore, they suggest that forming an accurate representation of these premises is important in ways that go beyond the immediate context in which one is trying to solve the problem; it is also critical for subsequent attempts at solving the problem. In this way, the downstream consequences of earlier misrepresentation are not just interpersonal (as demonstrated in the serial reproduction findings), but also intrapersonal (as demonstrated by the corruption effect).

Still on the topic of memory dependence, the results were remarkably consistent with regard to the type of memory error that was most detrimental for deliberation. In line with previous work (Mata et al., 2014, 2017), the filtering of the critical premise was the error that most consistently predicted reasoning performance. Other types of error (distortion), or errors in other premises, did not have a consistent effect on reasoning. And the filtering of the critical premise compromised not only immediate reasoning but also subsequent reasoning for the original problems when they were presented a second time. These results also add to the literature on conflict detection (De Neys, 2012; Pennycook, Fugelsang, & Koehler, 2012), and they lend further support to the notion that sound comprehension is the basis for good reasoning (Mata et al., 2014, 2017; Mata & Ferreira, 2018).

These studies served as a modest demonstration of the effects of filtering in reasoning, but we suspect that these effects also hold for a variety of other domains. As an example, moral dilemmas are known to trigger powerful intuitions, for instance about universal deontological mandates, which may be in conflict with consequentialist reasoning (Greene & Haidt, 2002; Haidt, 2001). The way in which these dilemmas are understood and represented should be equally subject to the same kind of filtering that we observed in the present studies, with certain key premises getting distorted or completely dropped in the process of social transmission, and with the potential for deliberative reasoning about conflicting aspects in the dilemma (Landy & Royzman, 2018) getting eroded in the process. And the same might apply to many other domains where there is important information on societal issues that is nuanced and complex, but where people tend to attend to only part of it, either because of cognitive limitations or motivated biases, with the potential for erosion and corruption effects to emerge. Examples include health and risk information (Frost et al., 1997; Moussaïd et al., 2015; Reyna & Brainerd, 2008), information about groups and stereotypes (Hamilton & Gifford, 1976; Hamilton & Rose, 1980; Kashima, 2000; Lyons & Kashima, 2003), negativity or partisan media biases (Knobloch-Westerwick, Mothes, & Polavin, 2017; Rozin & Royzman, 2001), and the myside bias (Baron, 1995; Stanovich, West, & Toplak, 2013).

One might argue that the pernicious effects observed in these studies are more likely to occur in one-on-one linear transmission rather than in group discussion. Indeed, group discussion has been shown to remedy incorrect reasoning, such that, through sound argumentation, people with the correct solution are able to persuade others to adopt that solution even if they considered other solutions (Trouche et al., 2014). There are, however, two factors that should temper this optimism: arguably, one-on-one communication takes place just as often, if not more often, than group discussion. That is, in many instances when we first hear about a problem, we do not have the luxury to be surrounded by a group working to get to the correct solution. Moreover, although group discussion has a great potential to remedy certain reasoning errors, it will only succeed in getting people to the correct solution to the extent that the group members are discussing on the basis of correct premises, and not distorted premises that were fed by previous sources.

Finally, the present studies did not measure reasoners' confidence in their responses, but it would be interesting to track the course of confidence as conflict gets filtered. As the collective understanding of a problem converges towards increasingly simpler representations, the confidence that people have in their solutions should increase. Thus, reasoners would grow more confident about a problem whose complexity they understand more poorly. This dynamic might shed light on how certain positions become extreme (e.g., van Prooijen & Krouwel, 2017), even though the strength and resolve with which people defend those positions is not necessarily related to how deeply they understand it. In fact, it can even be inversely related, such that those who have a less deep understanding of a topic are those with stronger feelings about it (Fernbach, Rogers, Fox, & Sloman, 2013).

6 – Chapter VI – Discussion and conclusions

In this work, we investigated the social amplification of judgment biases: As information travels from one generation to another, the messages accumulate systematic distortions, such that the level of bias in the population progressively increases the further information travels away from its origin, resulting in a population that is more biased than if all members had first-hand contact with the data.

This social amplification of bias was consistently observed in several studies with several methodological variations: In particular, it was observed in various judgment and decision-making domains (risk perception, illusory correlations, denominator neglect, and cognitive reflection), with both frequency estimates (Chapters II and III), forced recognition (Chapter III), and free recall (Chapters II, IV, and V), when people transmitted the problem (Chapters IV and V) or their answers to the problem (Chapters II and III), with lab (Chapters II, III, and IV) or online participants (Chapter V), and by way of distorting (Chapters II to IV), adding (Chapter II), or filtering (Chapter V) information. These studies also demonstrated how the social amplification of bias can lead to downstream effects in people's attitudes (Chapter III), choices (Chapter IV), and subsequent understanding of problems (Chapter V).

These results are in line with previous studies using serial reproduction that reveal a convergence with shared cognition (Bangerter, 2000; Bartlett, 1932; Kalish et al., 2007; Kashima, 2000; Mesoudi et al., 2006). Sometimes the bias was not revealed by the majority (e.g., only around 13% of participants revealed a bias to filter the critical premise in Chapter V), thus also supporting the results of Navarro et al. (2018) regarding the power of minorities to exert influence beyond what would be expected given their size. Still, while the study of

Navarro et al. (2018) found this power to be due to extremists' higher resistance to changing their attitudes, the results of Chapter V reveal that this can also happen in the absence of extreme attitudes, simply by filtering out important aspects of the information.

These findings also inform the ecological vs. intrapsychic debate about the origin of judgment and decision-making biases (Fiedler, 2014). In the ecological approach, the primary reasons for thinking biases can be found in the information environment, without a need to postulate intrapsychic biased mechanisms (Fiedler, 2000a; Fiedler & Wänke, 2009). The social amplification of bias presents a hybrid approach that merges the intrapsychic and environmental approaches. It does so by showing how individual biases of intrapsychic origin (e.g., frequency estimates in Chapter II that were inflated by varied levels of vividness of different causes of death or the filtering of certain elements of problems in Chapter V) can lead to a biased information environment. As the biased information generated by an intrapsychic bias is communicated to others and thus becomes part of the information environment, even perfectly unbiased persons should become biased by making use of this information (and note that advice-taking is a perfectly rational behavior in many cases; e.g., Yaniv, 2004a, 2004b).

These results also reveal another way in which bias can be not only an outcome (e.g., of the use of heuristics, Tversky & Kahneman, 1974), but also a cause in itself (Keren & Teigen, 2004). Thus, it is not only possible that one type of bias (e.g., the matching bias) causes other biases (e.g., errors in the Wason selection test, the construction and evaluation of truth tables; Evans, 1989), but also that some person's biases causes biases in others. As such, by ignoring the social amplification of bias (as is always the case when ecological factors are ignored), we could wrongly assume that a person that behaves in a biased way is a biased *person*, while indeed what we witness is the result of a biased *environment*, leading us to implement the wrong type of interventions to correct the bias.

As for the method of serial reproduction, it has once again revealed itself as a magnifying glass (Paul, 1959), allowing small and even sometimes non-significant effects to grow larger and significant. For example, in Experiment 1 of Chapter III, the illusory correlation was only small and marginally significant in the first generation ($d = 0.29$, $p = .097$), yet the illusory correlation revealed by the third generation grew to a close-to-large effect size (i.e., $d = 0.80$; Cohen, 1988), and became highly significant ($d = 0.73$, $p < .001$).

This result provides further support to the defense of the relevance of small effects, as argued by Greenwald, Banaji, and Nosek's (2015). These authors claimed that small effects should be considered consequential if they satisfy one of two conditions: 1) if they affect a significant portion of the population, or 2) if they repeatedly affect the same people. Of course, the list of aspects that are important in interpreting effect sizes is more extensive than that, and includes experimental control, measurement validity, and knowledge of the particular research area and practical implications involved (Cohen, 1988). The aspects presented by Greenwald et al. (2015) may be considered as the type of aspects that can be included in the contextual knowledge necessary to interpret effect sizes, alongside others, such as the size of the manipulation (e.g., nudging vs. shoving) and how difficult it is to influence the dependent variable (Prentice & Miller, 1992). The studies presented in the current dissertation add another criterion to this group of aspects: If small effects found with studies using isolated individuals can become stronger as information is passed from one person to another, those small effects should also be taken into account. Conversely, large biases may also become smaller and inconsequential, if the conditions are present for the social attenuation of bias (e.g., resulting from argumentation; Claidière et al., 2017), revealing the importance of serial reproduction to study the impact of biases in society.

The present set of studies, in particular the combination of serial reproduction and repeated reproduction in Chapter III, also has implications for the study of collective memory. Roediger et al. (2009) note that collective memories are created through the retrieval of individual memories, but repeated retrieval (which is studied with the repeated reproduction paradigm) is the process that strengthens and maintains these collective memories over time. As repeated retrieval is frequently performed in the presence of others, to whom people retell the information they recall, serial reproduction allows the study of a second important aspect of collective memory: how it spreads beyond those who had first-hand contact with the events (e.g., as with the generations that lived through the second world war and those who were born after it). Future studies combining the repeated and serial reproduction methods can increase our understanding of the process of formation and maintenance of collective memories in a population over time, particularly if feedback both within (Roediger et al., 2009) and between generations is allowed. Of course, this combination of paradigms may also shed light on other collective concepts, such as stereotypes and risk perception.

Furthermore, the experiments in Chapter V are a demonstration of how the serial reproduction method may be used to discover new effects: in that chapter, participants first received others' reproductions of the problem, but were subsequentially (after recall) presented with the problem in its original, intact version. This led to a corruption of subsequent reasoning, as participants who were first exposed to a retelling of the problem without its critical element were more likely to give an incorrect response than those who were first exposed to a retelling that preserved the critical element. This methodological modification allows the serial reproduction method to capture new dynamics of real-life situations, such as when meeting someone after being first exposed to someone else's opinion about them, going on a vacation in a city that a friend told us about, or when receiving information about a task by someone who had first attempted to solve it. While some studies (e.g., in rumor and reputation; Dalal, Diab, & Tindale, 2015; Jones & Skarlicki, 2005) have captured this sort of dynamic, the results of Chapter V show how such a dynamic may also be important in judgment and decision-making. The reverse situation, - in which people first have direct contact with the material and are then exposed to it as recalled by someone else, is also potentially interesting, as it might attenuate the social amplification of bias. This would be in line with results in persuasion research, where it has been found that people are less susceptible to persuasion if they first have direct contact with the object of attitudes (e.g., food products; Wu & Shaffer, 1987) and the egocentric discounting found in the advice-taking literature (e.g., Yaniv & Kleinberger, 2000).

6.1 – Limitations and potential future studies

At this point, it is important to discuss some limitations of these studies and consider potential future experiments that might circumvent those limitations.

First, unidirectional one-to-one communication is only one of many types of communication in society. Research using groups (i.e., bidirectional many-to-many communication) reveals that an important predictor of group judgment and decision-making is whether the problem has a demonstrable solution (Laughlin & Ellis, 1986; Trouche et al., 2014): When a demonstrable solution exists, it takes only one member of the group to convince others and for the correct response to be adopted by the majority. When a demonstrable solution does not exist, then the majority's opinion (which may or not align with the correct answer) typically wins, unless the minority can make use of a shared representation to persuade the majority (Kerr & Tindale, 2004).

The studies in the current dissertation do not reveal the same pattern of results as research using groups. For example, in Chapter V, the Cognitive Reflection Test (Frederick, 2005) can be considered a problem with a demonstrably correct solution: a study by Trouche et al. (2014) found that 45% of participants switched from an incorrect to the correct answer after being given an argument, created by another participant, supporting the correct answer.

The reverse, switching from correct to incorrect after receiving an argument supporting the incorrect answer, never happened.

Yet, in the studies reported in Chapter V, the minority who recalled the problem in a biased manner (13.33% and 12.87% in the first and second experiments) did not decrease as generations progressed, - the number of biased messages in the population actually doubled after three generations (to 34.44% and 31.60%). It is simple to understand why the results differ: In the serial reproduction studies, participants have only the previous participants' retelling of the problem, and so cannot demonstrate the correct answer even if they could produce it given the intact version of the problem, - if groups had to work with problems with corrupted premises, truth-wins results would also be precluded. If, in serial reproduction, participants have direct experience with the problem before receiving the retelling[14] (making these studies more like advice-taking research; e.g., Yaniv & Kleinberger, 2000) or if they are able to receive and compare retellings from different persons, then results may be more in line with those observed in groups and with Claidière and collaborators' (2017) serial reproduction study.

Second, while the current studies bring a neglected Social Psychology concept into the judgment and decision-making research area, a set of important social factors have still been left out of the current dissertation. Indeed, in all studies, participants were never informed of any characteristic of the sender or of the receiver of the information. On the side of how receivers perceive the sender, they may make adjustments for factors such as perceived incompetence (Wilson & Sherrell, 1993), bias (Pronin et al., 2002), or lack of candidness (Walster & Festinger, 1962). On the side of how senders perceive the receivers, they may, for example, adapt the information to be more congruent with the perceived

[14] Note that the results of Chapter V suggest that if people first receive biased, second-hand information, and only afterwards come into direct contact with the original information, they are still influenced by being first exposed to biased information. Future studies may test whether this is an effect specific of the Cognitive Reflection Test or whether it generalizes to other types of materials.

preferences of the audience (Higgins, 1992) or display accountability effects (Lerner & Tetlock, 1999). A study by Lyons and Kashima (2003) reveals how this approach may be fruitful: when participants assumed or were informed that others did not endorse a given stereotype, they communicated less stereotype-consistent information (i.e., a social attenuation of bias) than participants who assumed or were informed that others did endorse the stereotype. Thus, these social perception factors may either amplify (e.g., chains composed of people with extreme political views, as frequently occurs in social media and their eco-chamber effect; Barberá, Jost, Nagler, Tucker, & Bonneau, 2015) or attenuate the social amplification of bias (e.g., chains composed of people with diverse attitudes, possibly leading the bias of one to be corrected by the bias of the other). Future studies should explore these different dynamics and test boundary conditions to the general effect.

Third, and related to the previous point, another limitation has to do with the computer-mediated transmission of information. If the transmissions were done by the two participants transmitting information verbally in a live interaction, 1) factors related to the source and sender of information would probably take effect, including subtle non-verbal factors that may affect both bias and confidence, and 2) two-way communication would become possible, such that people could ask for clarifications, counterexamples, etc., possibly bringing the results of such a paradigm closer to those of research with groups.

Fourth, we did not explore individual differences that may influence the social amplification of bias. Following Paul (1959), homogenous chains may be used to assess the extent to which a given individual difference has an impact on the progression of information distortion over time. Many individual differences may influence the social amplification of bias, - for the judgment and decision-making biases, the work of Stanovich and West (1998) suggests two broad types of individual biases that are relevant for performance in heuristics-and-biases problems: 1) variables that reflect algorithm-level limitations, such as those measured by Scholastic Aptitude Test scores, and 2) variables that reflect thinking dispositions that may moderate competence, such as those measured by the Actively Open-Minded Thinking scale (Stanovich & West, 1998). The correlation between these general competence and thinking disposition measures and each specific bias can vary significantly (Stanovich & West, 1998), so future studies should explore particular individual differences presumed to be important for a particular bias. The intensification and attenuation of other (e.g., social) types of biases may be moderated by other individual differences. For example, certain individuals may be less willing to communicate stereotype-consistent information in general. A type of individual difference that may have an impact in all types of social amplification is motivation. For example, in the judgment and decision-making literature, it has been found that the motivation to achieve certain desirable conclusions leads to different levels of reasoning sophistication (e.g., Mata et al., 2015), and so the social amplification of bias may be moderated by the person's motivation to achieve a certain conclusion.

Fifth, the current results can be easily complemented by future studies including confidence measures. For example, the social amplification of illusory correlations guided by a priori attitudes may lead to increased confidence as the data becomes closer to people's expectations (Schultze et al., 2015) and the social filtering of the Cognitive Reflection Test may be associated with a progressive increase in confidence, as those who fail to answer the problem correctly think the problem is easier than those who solve them (Frederick, 2005). Thus, the inclusion of confidence may contribute to the understanding of how certain positions become extreme (e.g., van Prooijen & Krouwel, 2017), and in general it would add a metacognitive dimension to this research.

A final limitation of this work pertains to the combination of repeated reproduction and serial reproduction in Chapter III, following the footsteps of Roediger and collaborators (2014). Theoretically, the two methods have significant differences. For example, repeated reproduction can lead to reminiscence (i.e., the recovery in a given test of original material forgotten in a previous test; Ballard, 1913; Wheeler & Roediger, 1992) and testing-like effects (i.e., the act of retrieving information can, by itself, be protective for long-term learning by comparison with re-study; Roediger & Karpicke, 2006; Rowland, 2014). Yet, reminiscence and testing effects are not possible in serial reproduction because there is not a memory that can allow old material to re-emerge or a memory to be protected against forgetting by each recall.[15] Still, the combination of these two paradigms does allow us to: 1) gauge which of the two effects, individual or social, leads to stronger bias (Roediger et al., 2014), and 2) construct a complete image of the memory process, as memory typically involves both many retellings of the event by the same person (as in repeated reproduction) and the receiving of information from others (as in serial reproduction; Roediger et al., 2009).

Nevertheless, the combination of repeated and serial reproduction raises important methodological questions, chief among them how to deal with the fact that there is full horizontal dependence (in the repeated reproduction, as all participants repeat recall the same number of times), but only partial vertical dependence (in the serial reproduction, only the first recall of a given participant can influence the first recall of the next participant in the communication chain; the second recall of a participant has no effect on the second recall of the next participant). While we opted for a simple and computationally light approach (by comparing the correlations of serial and repeated reproduction through a *Z* test, thus ignoring this dependence), future studies should address this statistical issue by using multilevel mixed models (e.g., Goldstein, 2010) to account for this complex structure.

6.2 – Modelling the social amplification of bias

[15] Serial reproduction may serve as an interesting control for repeated reproduction. For example, when using lists of associated words, the resurgence of an item from the original list can both be due to guessing based on the list theme (e.g., animals) or the creation of a false memory instead of actual reminiscing. To the extent that a participant in the serial reproduction can never reminisce, an increase in performance over time in repeated reproduction by comparison with serial reproduction is a robust test of reminiscence.

In the introductory chapter, the social amplification of bias hypothesis was described using the following equation:

$$\text{Message at destination}_t = \text{Original information} + \text{Bias}_t + \text{Noise}_t.$$

While this is a useful first approach, future research on the social amplification of bias may be improved by using agent-based modeling. Agent-based models are models in which the modelling happens at the level of individual agents (that can represent people, groups, neurons, etc., depending on the desired level of analysis) through a description of how the agents behave and interact with an environment (that can be composed of other agents and/or inactive elements such as obstacles or resources; E. R. Smith & Conrey, 2007). So, while in the equation-approach we would predict that a significant level of bias will lead to a significant amount of distortion in messages, in the agent-based model approach theses variables also have an impact, but not by the attribution of weights to factors (e.g., that a 1 unit of bias in the population will lead to 15% increase in distortion of messages), but rather by the determination of how agents behave and the consequences of their behavior for the variable of interest (e.g., there is a given probability that any agent is biased, - in case the agent is biased, then the agent will give a biased answer).

As an example of an agent-based model, consider Schelling's (1971) classical spatial proximity model. The purpose of this model was to explain segregation, such as the creation of racial neighborhoods, as resulting from individual (not group-level) preferences and behavior. The model consists of two groups of agents, whose membership is permanent and recognizable. Agents are initially distributed at random on some spatial dimension, have preferences about the membership composition of their neighborhoods, and will move from one location that does not meet their individual neighborhood preferences to the closest (available) location that does. The model is easily implemented in any agent-based modelling software (e.g., NetLogo; Wilensky, 1999), but can also be easily demonstrated using two types of coins on a checkerboard and proceeding from one coin to another, checking whether the coin's preferences are met in its neighborhood (which can be defined as all coins adjacent to the target coin), changing the coin to the nearest location that meets its demands in case they are unmet, and repeating the procedure until all coins are in a satisfying location. Segregation can then be measured as agents' average proportion of in-group neighbors (see Figure 12 for a sample run, using NetLogo).

The surprising result of Schelling's (1971) spatial proximity model is that, by defining agents that are mostly tolerant, strong segregation is observed. For example, with individuals that only move if they are surrounded by less than 37.5% in-group neighbors, an initial random distribution (of around 50% segregation) does not lead to a decrease in segregation, but rather to an increase to about 75%. Thus, individual-level preferences or behaviors (e.g., individual tolerance) can lead to opposing group-level outcomes (e.g., group segregation).

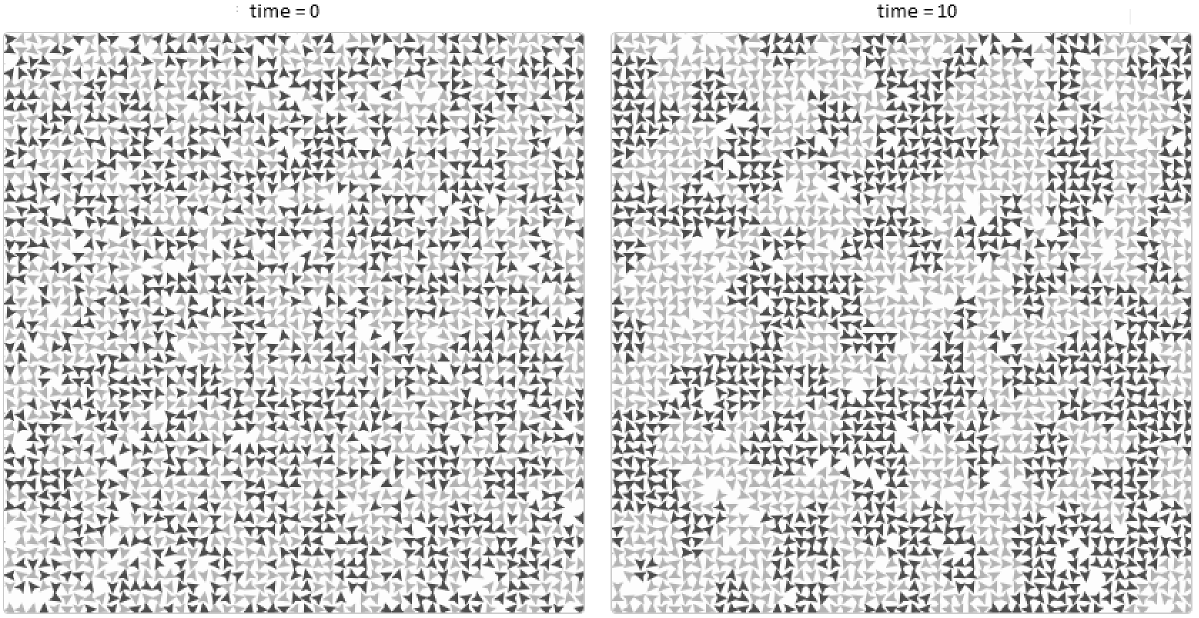


Figure 12. Example of a Shelling's spatial proximity model run, as implemented on NetLogo (Wilensky, 1997), with a preference for ≥30% in-group neighbors. Light and dark triangles represent the two groups of agents. On the left, the model as initiated (randomly distributed agents), on the right, the model after 10 time units (revealing ~70% segregation).

Smith and Conrey (2007) have called for the use of agent-based modelling in Social Psychology, arguing that this type of modelling has the following advantages: 1) it offers insight into the generative processes, as it focuses on modelling individual-level behaviors (e.g., people move depending on the quantity of in-group neighbors); 2) it does not require simplifying assumptions of rationality, such as self-interest or rational utility maximization, to make predictions, allowing different assumptions of this kind to be compared; 3) it allows bidirectional casual relationships (e.g., feedback effects); 4) it allows nonlinear, conditional, or qualitative effects (e.g., threshold effects, such as only moving after a certain level of out-group neighbors is reached); and 5) it allows varying levels of abstraction (e.g., neurons, individuals, groups). Smith and Conrey (2007) contrast agent-based modelling with the more traditional, in Social Psychology, causal modelling approach, concluding that both are complementary: Agent-based modelling serves as a tool for developing theory and understanding the theory's implications, while causal modelling serves the purpose of testing hypotheses in a given dataset.

A short demonstration of the usefulness of the agent-based approach for research into the social amplification of bias follows. A model was created using NetLogo 6.1.0 (Wilensky, 1999), inspired on the paradigm used in Chapter IV (denominator neglect).

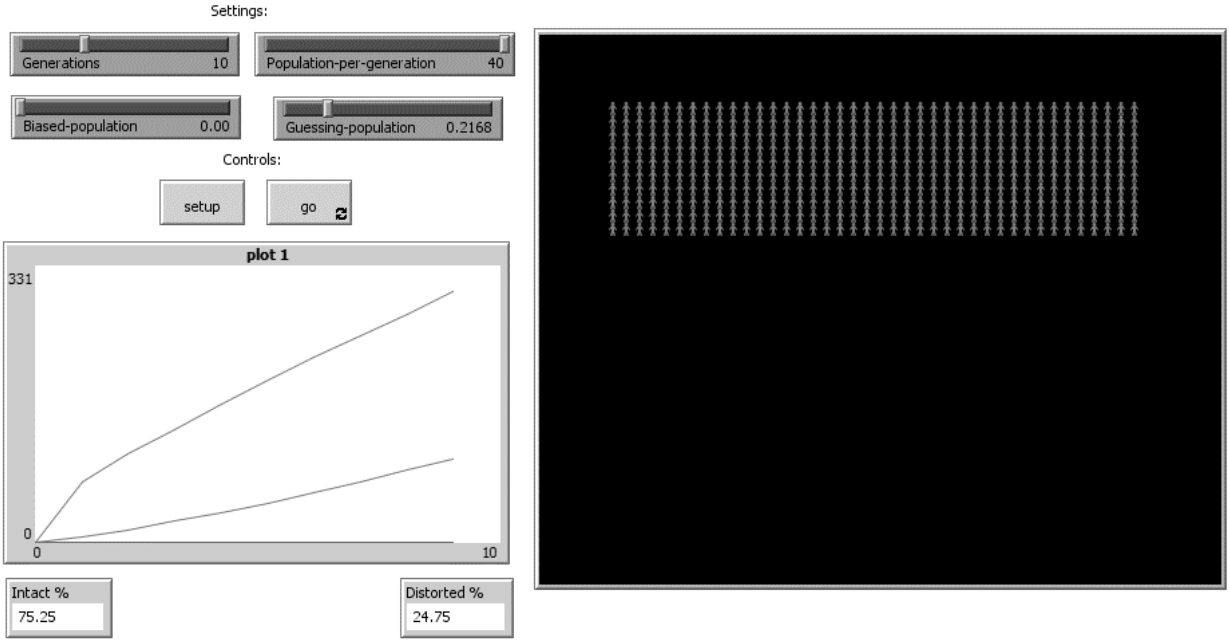
In Chapter IV, participants performed a free (and forced) recall task, so that three possibilities existed: 1 = the recall revealed a difference between the two options in line with the original, 0 = a distortion leads to no difference between the two options, and -1 = the difference is distorted to the opposite of the original. Yet, no participant, out of 120, ever freely recalled a lack of difference and only one did so during forced numeric recall, where guessing is more likely. Thus, it is difficult to model the results of Chapter IV, because it is not easy to assess the amount of guessing that took place. Based on the forced recall, one would expect about 0.01% of guessing, but one could think of alternative guessing strategies:

For example, maybe all participants did perceive a difference between the two options and the errors observed simply resulted from a confusion between the labels of the options (i.e., Car A vs. Car B, Treatment X vs. Treatment Z). This alternative is not very likely, though, as 12.92% of judgments participants made were judgments of no difference, so the lack of participants guessing an absence of difference cannot be because all of them represented the task as being between two different options.

Still, a binary version of the denominator neglect paradigm, in which participants can only either guess a correct or an incorrect version of the material, presents itself as a simplified scenario that is easy to implement in agent-based modelling and is sufficient to illustrate some of the ways in which this type of modelling may be used to improve research on the social amplification of bias. This model is described next.

During the setup phase of the model, a 2-dimensional space (that has no attributes other than allowing agents to share the same space or not) is populated by virtual agents that are positioned to form lines (each horizontal line represents a generation, and each vertical line a transmission chain). Two model parameters determine 1) the number of agents, and 2) the number of generations. See Figure 13 for the graphic user interface of the model.

Agents have only two properties, 1) their location, and 2) their memory state, which can include an intact or a distorted version of the original message. Furthermore, agents have only two actions, 1) movement, and 2) communication. Movement is performed one generation at a time. When a given generation is active, all agents of that generation move one space down, entering another agent's space (unless it is the last generation that is active). The communication behavior is nested inside movement and is only triggered if the active agent has entered the location of another agent. Communication represents one of three types of communication behaviors: the sender agent communicates 1) the exact message that is stored in its memory (i.e., unbiased communication), 2) the distorted version, regardless of the contents of its memory (i.e., biased communication), and 3) randomly chooses between the intact or distorted, regardless of the contents of its memory (i.e., guessing communication). Two parameters control the probabilities of each agent adopting each of the communication types: one that controls the probability of guessing, and another of biased communication (unbiased communication is whatever probability remains). Upon setup, the guessing and bias parameters also determine the probability that each first-generation agent 1) memorizes the intact version, 2) guesses between the intact and distorted version, or 3) memorizes the distorted version. This setup simulates the first-generation's contact with the original material.



Settings:

Generations 10

Population-per-generation 40

Biased-population 0.00

Guessing-population 0.2168

Controls:

Intact % 75.25

Distorted % 24.75

Figure 13. The graphic user interface of a binary agent-based social amplification of bias model, showing, on the left side, parameter settings, controls, and basic graphic and statistics, and, on the right side, the environment.

In Figure 14 one can see a basic calibration of the model, using the sample characteristics of the experiment in Chapter IV (40 participants and three generations). A total of 1000 simulations of each scenario were run using NetLogo's Behavior Space tool. As can be expected from logic, when all participants guess (all-guessing scenario), the average immediately starts at the middle of the scale and remains close to it. When all participants are unbiased (all-unbiased scenario), no distortion is ever introduced, and when all participants are biased (all-biased), all messages are immediately distorted and remain so.

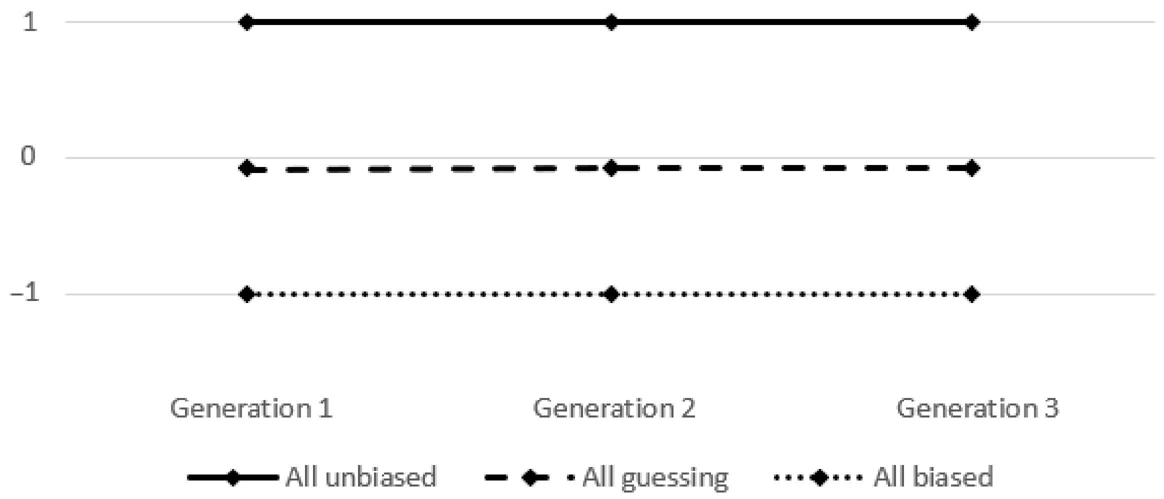


Figure 14. Simulated averages of information distortion (1 = intact, -1 = distorted) with 1000 simulations of three generations of 40 agents each when all agents are unbiased (solid line), guessing (dashed line), or biased (dotted line).

Next, to show how the model can be used to guide future studies, we compare two scenarios: 1) all the distortion observed in Chapter IV was due to bias (all-distortion-is-bias scenario), and 2) all the distortion observed in Chapter IV was due to guessing (all-distortion-is-guessing scenario). In Chapter IV's experiment, 27 out of 213 (i.e., 12.68%) messages that were received intact were then distorted, and one can use that value as an estimate of distortion individuals added on average (representing only new distortion, not distortion that is received from a previous participant and was simply kept). Thus, in the all-distortion-is-bias scenario, the bias parameter was set to 0.1268 and the guessing parameter to 0. For the all-distortion-is-guessing scenario, as the guessers randomly memorize one of the two possible version of the message, the probability of the guessing parameter was the double of the observed distortion (0.2536) to take into account that half of the agents will guess the intact version, and the bias parameter was set to 0. Figure 15 compares 1000 simulations of each scenario (all-distortion-is-guessing and all-distortion-is-bias) with Chapter IV's observed means.

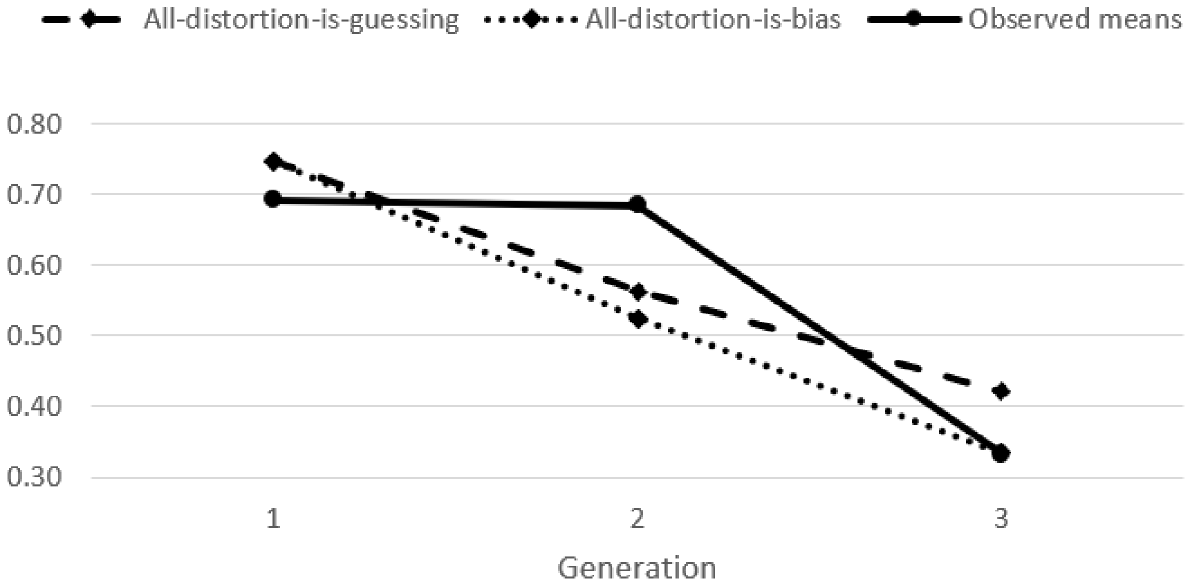


Figure 15. Simulated averages of information distortion (1 = intact, 0 = distorted) with 1000 simulations of 30 generations of 40 agents each when the observed first-generation distortion (solid line) of Chapter IV is assumed to be all-guessing (dotted line) or all-bias (solid line). The y-axis does not cover the full range of the scale (1 to -1) to allow the small differences between the three lines to be visible.

The results of Figure 15 show that the two scenarios make similar predictions (differing in 0, 0.04, and 0.09, in the first, second, and third generation, respectively), with the sum of squared errors between the observed and predicted means being identical up to two decimal places (SSE_{guessing} = 0.025, SSE_{bias} = 0.029, *F*(2, 2) = 1.13, *p* = .53). Yet, a useful feature of agent-based modelling is that one can extrapolate beyond the data obtained by simulating more generations and seeing how they progress over time. Figure 16 shows the result of 1000 simulations with 30 generations of 40 agents each for the all-distortion-is-guessing and the all-distortion-is-bias scenarios.

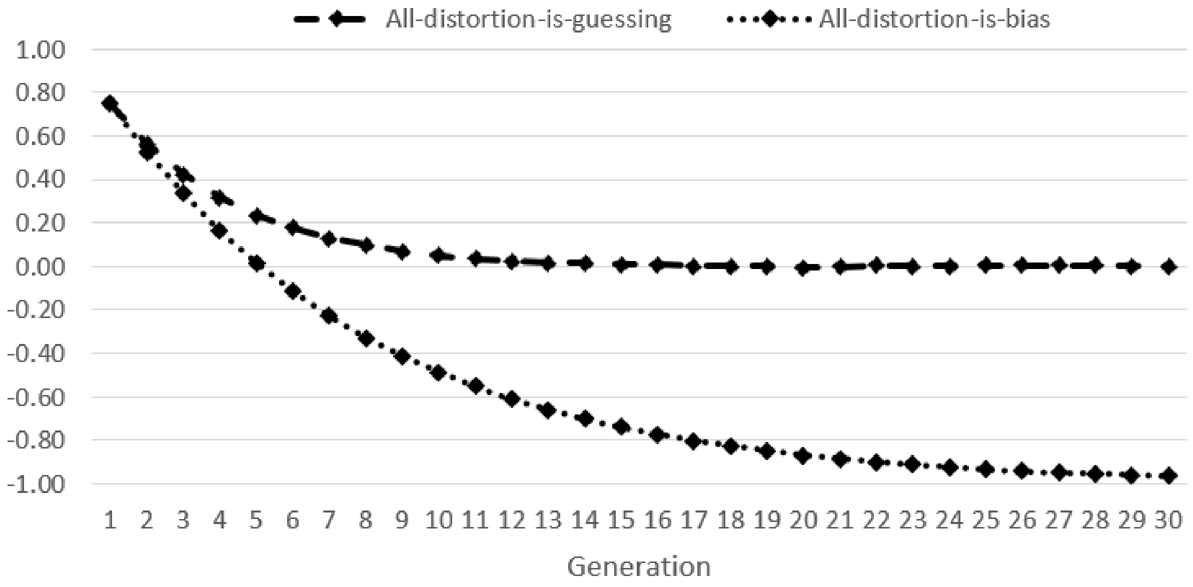


Figure 16. Simulated averages of information distortion (1 = intact, 0 = distorted) with 1000 simulations of 30 generations of 40 agents each when the observed first-generation distortion of Chapter IV is assumed to be all-distortion-is-guessing (dotted line) or all-distortion-is-bias (solid line).

Three interesting aspects can be noted from Figure 16. First, despite having the same starting point, the two lines start to diverge, with the all-distortion-is-bias scenario revealing higher levels of distortion. This is because, in the all-distortion-is-bias scenario, all new distortion is preserved, while in the all-distortion-is-guessing scenario, some of the distortion that is introduced by guessing is also reversed by guessing (i.e., an agent who received a distorted version can guess the intact version, thus reverting the distortion). Second, the progression is not linear, but exponential, although with few generations, a linear trend seems to suffice (see Figure 15). Third, the two lines stabilize around different values: The all-distortion-is-guessing scenario stabilizes in the middle of the scale (0), while the all-distortion-is-bias scenario sinks to the lower end (-1).

If, when all distortion is merely a result of guessing, the average converges with zero, strong evidence of a binary social amplification of bias could be obtained by running a study with enough generations to observe a mean significantly below the middle of the scale. In fact, the agent-based model can be used to estimate, by simulations, how many generations would be required to obtain a desired probability of finding a difference between the middle of the scale and the observed mean, if the expected difference does exist (i.e., to obtain a desired level of power). To do so, it is possible to reuse the simulations of the all-distortion-is-bias scenario, calculating a *t*-test between each simulation's observed mean and the middle of the scale (one-tailed, as only a significant difference *below* the mean would support a social amplification of bias) and checking the percentage of significant *t*-tests obtained. The results of such a simulation can be found in Figure 17.

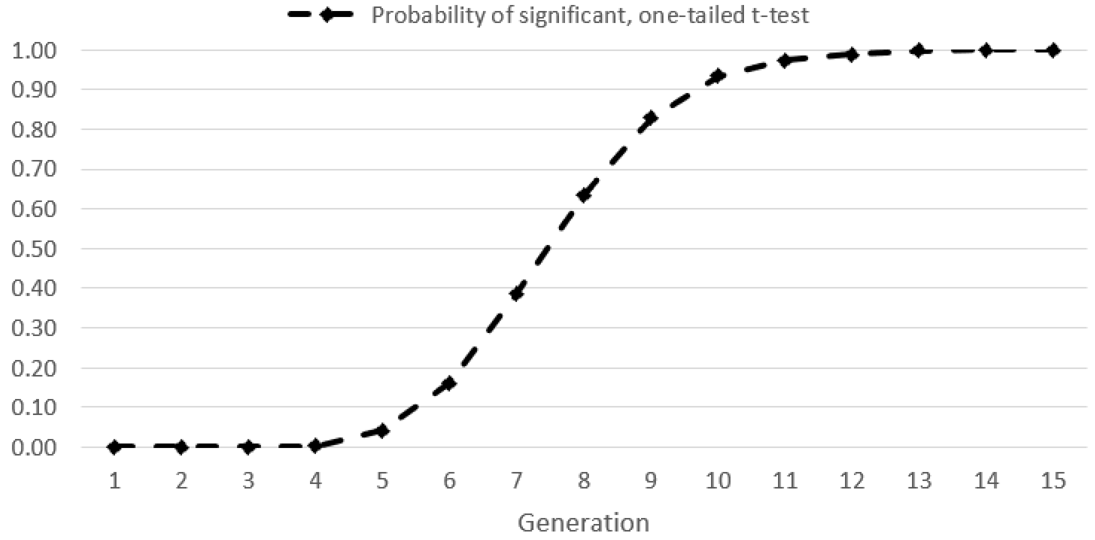


Figure 17. Simulated probability, per generation, of a significant one-tailed test revealing an average distortion significantly below the middle of the scale (0), given an average 12.68% distortion caused by bias and using 1000 simulations.

Thus, to obtain an 80% "power" to detect a bias that, on average, causes a distortion of 12.68% of the material in a dichotomous variable, nine generations, with 40 participants each, are required. Any of the many possible future studies suggested in this discussion, such as the inclusion of the possibility of feedback or of source effects, can easily be modelled by making adaptations to the current model, but two further particularly interesting uses of agent-based modelling for the social amplification of bias research are highlighted next.

First, the agent-based modelling approach may be used to compare different patterns of amplification: 1) biases that grow continuously (or until they reach a ceiling), 2) biases that grow until they reach a non-truncated level and then stabilize, and 3) biases that self-correct. For example, do the illusory correlations guided by a priori beliefs (such as stereotypes) grow until they become as strong as possible, such that all doctors are rich, and no biologist is rich? Do they stabilize once they reach a shared prevalence judgment in the population (e.g., almost all doctors are rich and almost no biologists are rich, but one or two incongruent cases exist)? Or, finally, do they self-correct, as the infrequent cases become distinctive and then guide the illusory correlation in the reverse direction?

Second, the agent-based modelling approach can be used to contrast the results of groups (by allowing all agents to communicate simultaneously with one another) and serial reproduction, as well as of any of the more complex and mixed types of communication structures, such as networks. For example, the results of Navarro and collaborators (2018), in which population heterogeneity led to minority-wins outcomes, differ from the results of groups, in which group composition does not by itself lead to minority-wins outcomes (it is required that minorities make use of shared cognitions in order to influence the majority). One reason for this discrepancy may be due to the possibility of receiving (in groups vs. serial reproduction) 1) feedback, or 2) information from more than one source. Using agent-based modelling, one could vary theses aspects and find out which is the likely factor behind this difference.

These interesting questions, as all others suggested, can be first modelled through agent-based modelling, to test their logical soundness and to determine the number of participants and generations necessary, and then tested in the laboratory.

6.3 – Recovering the original information

The current studies suggest that it is inadequate to neglect the impact of bias in communication, particularly when trying to retrieve the original information. Two techniques that can be used to retrieve the original information, or at least to obtain good approximations to the original information, are the use of abundance distributions (Dunn & Kirsner, 2011) and the wisdom of the crowd (Galton, 1907). If we use the data of Chapter III's experiments (about illusory correlations), which enable us to test both techniques, we can show how they perform.

In the abundance distribution technique, we choose as our best guess of the original information the most frequent informational variant. In Experiment 1 of Chapter III, participants were exposed to the following frequencies of four categories of stimuli: nice behaviors by group A: 18, rude behaviors by group A: 8, nice behaviors by group B: 9, and rude behaviors by group B: 4. Considering only the first generation of Experiment 1 (in Dunn & Kirsner's, 2011, study, all participants were first-hand witnesses of the event), the modal frequencies were: 10 nice behaviors by group A, 5 rude behaviors by group A, 6 nice behaviors by group B, and 10 rude behaviors by group B. Thus, the abundance distribution has not revealed the original information (that both group A and B have the same proportion of the two types of behaviors), but the bias (that group A has more nice than rude behaviors, while group B has more rude than nice behaviors). For Experiment 2, again considering only the first generation (and the first recall), participants saw typical, non-typical, and neutral traits associated with members of three professions: doctor, biologist, and designer. Participants saw a total of 4 typical and neutral traits per profession, and 8 non-typical traits per profession. The modes were in line with Experiment 1, such that there was an inflation, in the very first generation, of the traits considered typical of each profession (doctor = 5, biologist = 5, designer = 6), at the expense of some non-typical traits (doctor = 7, biologist = 7, designer = 8), without influencing the neutral traits (doctor = 4, biologist = 4, designer =

4). This technique, thus, leads to the recovery of the bias, not signal, in the cells influenced by bias (i.e., regarding stereotypic traits; Hamilton & Rose, 1980).

In the wisdom of the crowd technique, we average participants' estimates to obtain our best guess of the original information. In Experiment 1, Chapter III, we used participants' frequency estimates to calculate an index that assessed the difference in the proportion of behaviors between the two groups. If only error is present, the errors should be able to cancel each other out. We again consider only the first generation, as wisdom of the crowd techniques were developed for first-hand estimations (Galton, 1907). We obtained an average difference between proportions with *M* = 0.05, *SE* = 0.17, marginally different from zero, *t*(34) = 1.71, *p* = .097, and 10 participants (28.57%) making better approximations than the wisdom of the crowd. In Experiment 2, for participants in the first generation (and first recall), the proportion (relative to the original frequency) of typical traits was above the original proportion (*M* = 1.34, *SD* = 0.42), *t*(49) = 5.64, *p* < .001, non-typical traits were below the original proportion (*M* = 0.83, *SD* = 0.23), *t*(49) = -5.27, *p* < .001, and proportion of neutral traits were successfully recovered to two decimal places (*M* = 1.00, *SD* = 0.29), *t*(49) = 0.08, *p* = .935. In terms of performance, 30 (60%) participants outperformed the wisdom of the crowd in their estimates of proportion of typical traits, 24 (48%) in the proportion of non-typical traits, and only 6 (12%) in the proportion of neutral traits.

Given the above, one can see how these otherwise effective methods start to malfunction if the population contains significant bias. It is important to notice that this is not a result of the serial reproduction method in itself, but rather a result of the existence of significant populational bias (see, e.g., Einhorn, Hogarth, & Klempner, 1977; Yaniv, 2004b), which in the case of the wisdom of the crowd pulls the average away from the original information, and in the case of abundance distributions generates an alternative variant that competes with the original. The addition of a social amplification of bias, of course, further intensifies this problem, allowing small and possibly negligible biases to accumulate and become stronger, further undermining these techniques.

If one again recalls the social amplification of bias equation…

$$\text{Message at destination}_t = \text{Original information} + \text{Bias}_t + \text{Noise}_t.$$

…one can see that, to be able to retrieve the original information, it is necessary to identify the bias parameter. This may be done if the bias in the population is known and if one uses only people from the first-generation or, alternatively, if one knows how far each member of the population is from the origin and makes certain assumptions about the spread of information (that all members of the population are equally likely to communicate with one another, that people tend to communicate within homogenous groups, etc.). While this may be a feasible technique in some situations (e.g., on social media, where one can map the social network and trace the information from its source to its various destinations), in many others it is not.

6.4 – Combating the social amplification of bias

If, in many situations, not much can be done to recover the original information transmitted within a biased population, maybe something can be done to avoid the social amplification of bias from emerging in the first place. Next, we will see interventions that focus on 1) the individuals who communicate, 2) the message that is communicated, and 3) the environment surrounding the communication.

First, the need to inoculate people against biases, much like governments inoculate whole populations against infectious diseases, has been raised for a while now (e.g., Turk & Salovey, 1986). Indeed, all of the biases that we covered can be consequential: 1) Heightened risk perception of dramatic risks may lead to local impact on business sales, residential property values, and economic activity, while lowered risk perception of ordinary risks may lead to lack of political and social pressure to implement risk reduction interventions

(Kasperson et al., 1988); 2) illusory correlations may foster xenophobic beliefs, as people perceive immigrants as more likely to commit crimes than residents, even though immigrants commit the same amount of crimes as residents when controlling for demographic variables such as gender and education (Rumbaut & Ewing, 2007); 3) the denominator neglect may lead to irrational health-related choices (Garcia-Retamero, Galesic, & Gigerenzer, 2010); and 4) and the filtering of certain, critical aspects of texts may lead to simplistic, non-stimulating environments that have no "desirable difficulties" (Bjork, 1994), or to simplistic representations that may lead to overconfidence and even extremism (e.g., Fernbach, Light, Scott, Inbar, & Rozin, 2019).

Interventions that focus on making the individuals who communicate resistant to biases can consist of three components: 1) interventions that promote rationality, such as teaching about the existence of bias (Fischhoff, 1982), teaching the correct algorithm (e.g., Fischhoff, 1982; Nisbett, Fong, Lehman, & Cheng, 1987; Lehman & Nisbett, 1990), or promoting favorable thinking dispositions (e.g., Stanovich & West, 1998; Tishman, Jay, & Perkins, 1993); 2) interventions that promote a general skepticism towards second-hand information or that increase epistemic vigilance (Mata, Fiedler, et al., 2013; Sperber et al., 2010), and 3) interventions that promote the search for new information, particularly from scientific sources, but also from diverse, opposing, or independent sources, a rationale that also underlies the consider-the-opposite debiasing technique (Lord, Lepper, & Preston, 1984). Combining interventions that increase rationality, skepticism, and thinking dispositions should 1) reduce the amount of bias that is introduced in the messages by reducing the amount of people who add bias, 2) reduce the impact of second-hand information by leading people to discount it, and 3) promote the exposure to different, independent sources of information, regarding the same fact, which should allow a better estimation of the truth (e.g., Yaniv, 2004a; Yaniv & Kleinberger, 2000).

An interesting way to implement these different interventions is by adopting an inoculation procedure similar to the one used in attitudes research (e.g., Banas & Rains, 2010; McGuire, 1961). For example, skepticism in second-hand information may be increased by explaining the need for skepticism and offering people weak second-hand information that goes against their interests (e.g., "I heard Psychology students are actually less altruistic than most people"). As another example, illusory correlations may be attenuated if people are taught how to apply the correct algorithm to calculate a correlation with a small set of data in which all types of events have the same frequency (and thus, no illusory correlation can emerge) or using a correlation that goes against their own interests (thus making use of people's capacity to engage in higher-quality reasoning when motivated to defend a desired motivation; e.g., Mata et al., 2015).

Second, interventions can focus on the messages that are communicated. Three examples are presented: 1) the promotion of noiseless-medium communication could, to some extent, reduce the social amplification of bias. In many social media sites, for example, users can copy or share the original source of information. This sharing, while allowing some biases (for example, in the selection of which parts of the text to share, or which sources to share), should lead to some improvement over verbal transmission; 2) sources of information could make attempts at crafting messages that are more likely to be transmitted without bias, for example by communicating ratio information in a range that promotes the desired risk-level: a wide range to increase risk perception (e.g., 1,000 deaths by suicide in the national population of ten millions) and a narrow range to decrease risk perception (e.g., one death by suicide per ten thousand national residents; Pinto-Prades, Martinez-Perez, & Abellán-Perpiñán, 2006), and 3) it may be important for certain sources of information to be curated and promoted in detriment of others, in an attempt to combat an impoverished or biased information environment, an idea in line with Kang's (2005) suggestion that, to reduce implicit bias against minorities, the media should limit the amount of crime reports it communicates.

Third, interventions can consist of changes to the environment in which people communicate, which can be harder or softer (as in nudging, Thaler & Sunstein, 2008). As an example of the latter, social media sites could add a special field to posts (after the main text of the post) where people link to the source of their messages. As an example of the former, a social norm (N. J. Goldstein, Cialdini, & Griskevicius, 2008; Reno, Cialdini, & Kallgren, 1993) could be communicated or otherwise explicitly created that promotes the checking of the original sources of information (and in so doing adds a reputational cost to infractions), thus inhibiting the sharing of false or partial information.

Out of the three approaches, bias reduction strategies that focus on debiasing individuals should be the preferable approach to reducing the social amplification of bias, as doing so not only preserves but should actually increase people's freedom. One way to define rationality is that it is the ability to select the best path to satisfy one's objectives (Evans, Over, & Manktelow, 1993). Bias, to the extent that it impedes one from achieving one's desired ends, reduces freedom (e.g., when a person is kept poor because of spending a significant amount of money on lotteries). Furthermore, strategies that focus on debiasing individuals are probably the most efficient of the three, as teaching the population to think for themselves is probably more time and cost-effective than engineering information and environments to deal with each biased behavior revealed by the population. Of course, this assumes that a method to immunize the population to a large spectrum of biases can be developed, implemented at a low cost, and re-applied whenever necessary (or having long-lasting effects). So far, no such method seems to exist (Soll, Milkman, & Payne, 2016), but is without a doubt a worthwhile endeavor.

6.5 – Conclusion

The present dissertation sought to contribute by revealing how four biases, - the tendency to communicate more dramatic than ordinary causes of death, illusory correlations, denominator neglect, and the filtering of critical elements of reasoning problems, can accumulate in the messages that people communicate to one another and, thus, lead to a social amplification of bias. This progressive increase in the biased messages had an impact on attitudes, choices, and subsequent ability to think properly about reasoning problems. Ideally, these findings can contribute to important discussions in the psychological literature and emphasize the importance of research on how to reduce bias and promote rationality. The limitations of these studies revealed interesting and potentially important avenues for future research, and an agent-based modelling approach was suggested for future studies interested in continuing research into the social amplification of bias.

In conclusion, while a lot of research has focused on bringing the heuristics-and-biases approach into social psychology, much can be gained by bringing social psychology into the heuristics-and-biases approach, as this dissertation hoped to show. Indeed, it is only by considering how people interact and how the information they communicate evolves and transforms over time that we can gain a better understating of how biases spread in our societies and how to combat them.

7 – References

Ahmad, N., Peterson, N., & Torella, F. (2015). The micromort: A unit for comparing and communicating risk to patients. *International Journal of Clinical Practice*, *69*(5), 515–517. https://doi.org/10.1111/ijcp.12643

Allport, G. W., & Postman, L. (1947). *The Psychology of Rumor*. Oxford, England: Holt, Henry.

Alonso, D., & Fernández-Berrocal, P. (2003). Irrational decisions: Attending to numbers rather than ratios. *Personality and Individual Differences*, *35*, 1537–1547. https://doi.org/10.1016/s0191-8869(02)00368-9

Alper, T. G., & Korchin, S. J. (1952). Memory for socially relevant material. *Journal of Abnormal and Social Psychology*, *47*(1), 25–37. https://doi.org/10.1037/h0055826

Ballard, P. B. (1913). Oblivescence and reminiscence. *British Journal of Psychology Monograph Supplements*, *1*(2), 1–82. Retrieved from https://archive.org/details/obliviscenceremi02ball/page/n14

Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, *77*(3), 281–311. https://doi.org/10.1080/03637751003758193

Bangerter, A. (2000). Transformation between scientific and social representations of conception: The method of serial reproduction. *British Journal of Social Psychology*, *39*(4), 521–535. https://doi.org/10.1348/014466600164615

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking & Reasoning*, *1*(3), 221–235. https://doi.org/10.1080/13546789508256909

Baron, J. (2008). *Thinking and Deciding*. New York, NY: Cambridge University Press.

Barrett, J. L., & Nyhof, M. A. (2001). Spreading non-natural concepts: The role of intuitive conceptual structures in memory and transmission of cultural materials. *Journal of Cognition and Culture*, *1*(1), 69–100. https://doi.org/10.1163/156853701300063589

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press.

Bebbington, K., MacLeod, C., Ellison, T. M., & Fay, N. (2017). The sky is falling: Evidence of a negativity bias in the social transmission of information. *Evolution and Human Behavior*, *38*(1), 92–101. https://doi.org/10.1016/j.evolhumbehav.2016.07.004

Bennett, K., & Corcoran, R. (2010). Biases in everyday reasoning: Associations with subclinical anxiety, depression and paranoia. *Psychosis*, *2*(3), 227–237. https://doi.org/10.1080/17522431003592649

Berger, J. (2011). Arousal increases social transmission of information. *Psychological Science*, *22*(7), 891–893. https://doi.org/10.1177/0956797611413294

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–206). Cambridge, MA: MIT Press.

Boland, S. C. (2013). Reflection expanded: Expanding the cognitive reflection test. Retrieved from https://www.scribd.com/document/138024235/Reflection-Expanded-Expanding-the-Cognitive-Reflection-Task

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Bonner, C., & Newell, B. R. (2010). In conflict with ourselves? An investigation of heuristic and analytic processes in decision making. *Memory and Cognition*, *38*(2), 186–196. https://doi.org/10.3758/MC.38.2.186

Bornstein, B. H., & Emler, A. C. (2001). Rationality in medical decision making: A review of the literature on doctors' decision-making biases. *Journal of Evaluation in Clinical Practice*, *7*(2), 97–107. https://doi.org/10.1046/j.1365-2753.2001.00284.x

Brainerd, C. J., & Reyna, V. F. (1992a). Developmental independence of memory and reasoning. *Psychological Science*, *3*(6), 332–339. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1016.8132&rep=rep1&type=pdf

Brainerd, C. J., & Reyna, V. F. (1992b). The memory independence effect: What do the data show? What do the theories claim? *Developmental Review*, *12*(2), 164–186. https://doi.org/10.1016/0273-2297(92)90007-O

Busenitz, L. W., & Barney, J. B. (1997). Differences between entrepreneurs and managers in large organizations: Biases and heuristics in strategic decision-making. *Journal of Business Venturing*, *12*(1), 9–30. https://doi.org/10.1016/S0883-9026(96)00003-1

Carter, R. A. (2018). Illusory correlation and perceived criminality. *College of Arts & Sciences Senior Honors Thesis*. Retrieved from https://ir.library.louisville.edu/honors/179

Castelli, L., & Carraro, L. (2011). Ideology is related to basic cognitive processes involved in attitude formation. *Journal of Experimental Social Psychology*, *47*(5), 1013–1016. https://doi.org/10.1016/j.jesp.2011.03.016

Chapman, L. J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior*, *6*(1), 151–155. https://doi.org/10.1016/S0022-5371(67)80066-5

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, *72*(3), 193–204. https://doi.org/10.1037/h0024670

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, *74*(3), 271–280. https://doi.org/10.1037/h0027592

Chapman, M., & Lindenberger, U. (1992a). How to detect reasoning-remembering dependence (and how not to). *Developmental Review*, *12*(2), 187–198. https://doi.org/10.1016/0273-2297(92)90008-P

Chapman, M., & Lindenberger, U. (1992b). Transitivity judgments, memory for premises, and models of children's reasoning. *Developmental Review*, *12*(2), 124–163. https://doi.org/10.1016/0273-2297(92)90006-N

Claidière, N., Trouche, E., & Mercier, H. (2017). Argumentation and the diffusion of counter-intuitive beliefs. *Journal of Experimental Psychology: General*, *146*(7), 1052–1066. https://doi.org/10.1037/xge0000323

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. L. (1979). On the psychology of prediction: Whose is the fallacy? *Cognition*, *7*, 385–407. https://doi.org/10.1016/0010-0277(79)90023-4

Collins, R. L., Taylor, S. E., Wood, J. V., & Thompson, S. C. (1988). The vividness effect: Elusive or illusory? *Journal of Experimental Social Psychology*, *24*(1), 1–18. https://doi.org/10.1016/0022-1031(88)90041-8

Combs, B., & Slovic, P. (1979). Newspaper coverage of causes of death. *Journalism Quarterly*, *56*(4), 837–849. https://doi.org/10.1177/107769907905600420

Cosmides, L. (1985). *Deduction or Darwinian algorithms? An explanation of the "elusive" content effect on the Wason selection task*. Harvard University. Retrieved from https://www.cep.ucsb.edu/papers/cosmides_1985.pdf

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*(3), 187–276. https://doi.org/10.1016/0010-0277(89)90023-1

Dalal, D. K., Diab, D. L., & Tindale, R. S. (2015). I heard that...: Do rumors affect hiring decisions? *International Journal of Selection and Assessment*, *23*(3), 224–236. https://doi.org/10.1111/ijsa.12110

Dale, D., Rudski, J., Schwarz, A., & Smith, E. (2007). Innumeracy and incentives: A ratio bias experiment. *Judgment and Decision Making*, *2*(4), 243–250. Retrieved from http://journal.sjdm.org/jdm07007.pdf

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299. https://doi.org/10.1016/j.cognition.2007.06.002

Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology*, *66*(5), 819–829. https://doi.org/10.1037//0022-3514.66.5.819

Denes-Raj, V., Epstein, S., & Cole, J. (1995). The generality of the ratio-bias phenomenon. *Personality and Social Psychology Bulletin*, *21*(10), 1083–1092. https://doi.org/10.1177/01461672952110009

Detmer, D. E., Fryback, D. G., & Gassner, K. (1978). Heuristics and biases in the medical decision-making. *Academic Medicine*, *53*(8), 682–683. https://doi.org/10.1097/00001888-197808000-00012

Dunn, J. C., & Kirsner, K. (2011). The search for HMAS Sydney II: Analysis and integration of survivor reports. *Applied Cognitive Psychology*, *25*(4), 513–527.

https://doi.org/10.1002/acp.1735

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, *84*(1), 158–172. https://doi.org/10.1037/0033-2909.84.1.158

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, *49*(8), 709–724. https://doi.org/10.1037/0003-066x.49.8.709

Epstein, S., & Pacini, R. (2001). The influence of visualization on intuitive and analytical information processing. *Imagination, Cognition and Personality*, *20*(3), 195–216. https://doi.org/10.2190/G4VG-AKQP-2Q91-JQHP

Evans, J. S. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Evans, J. S. B. T., Over, D. E., & Manktelow, K. I. (1993). Reasoning, decision making and rationality. *Cognition*, *49*(1–2), 165–187. https://doi.org/10.1016/0010-0277(93)90039-X

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Eysenck, M. W., & Eysenck, M. C. (1980). Effects of processing depth, distinctiveness, and word frequency on retention. *British Journal of Psychology*, *71*(2), 263–274. https://doi.org/10.1111/j.2044-8295.1980.tb01743.x

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fazio, L. K., Kennedy, C. A., & Siegler, R. S. (2016). Improving children's knowledge of fraction magnitudes. *PLoS ONE*, *11*(10), 1–14. https://doi.org/10.1371/journal.pone.0165243

Fernbach, P. M., Light, N., Scott, S. E., Inbar, Y., & Rozin, P. (2019). Extreme opponents of genetically modified foods know the least but think they know the most. *Nature Human Behaviour*, *3*(3), 251–256. https://doi.org/10.1038/s41562-018-0520-3

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, *24*(6), 939–946. https://doi.org/10.1177/0956797612464058

Ferreira, M. B., Mata, A., Donkin, C., Sherman, S. J., & Ihmels, M. (2016). Analytic and heuristic processes in the detection and resolution of conflict. *Memory & Cognition*, *44*, 1050–1063. https://doi.org/10.3758/s13421-016-0618-7

Fiedler, K. (1991). The tricky nature of skewed frequency tables: An information loss account of distinctiveness-based illusory correlations. *Journal of Personality and Social Psychology*, *60*(1), 24–36. https://doi.org/10.1037//0022-3514.60.1.24

Fiedler, K. (2000a). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*(4), 659–676. https://doi.org/10.1037/0033-295X.107.4.659

Fiedler, K. (2000b). Illusory correlations: A simple associative algorithm provides a convergent account of seemingly divergent paradigms. *Review of General Psychology*, *4*(1), 25–58. https://doi.org/10.1037/1089-2680.4.1.25

Fiedler, K. (2014). From intrapsychic to ecological theories in social psychology: Outlines of a functional theory approach. *European Journal of Social Psychology*, *44*(7), 657–670. https://doi.org/10.1002/ejsp.2069

Fiedler, K., Russer, S., & Gramm, K. (1993). Illusory correlations and memory performance. *Journal of Experimental Social Psychology*. https://doi.org/10.1006/jesp.1993.1006

Fiedler, K., & Wänke, M. (2009). The cognitive-ecological approach to rationality in social psychology. *Social Cognition*, *27*(5), 699–732.

https://doi.org/10.1521/soco.2009.27.5.699

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge, UK: Cambridge University Press.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Frewer, L. J., Miles, S., & Marsh, R. (2002). The media and genetically modified foods: Evidence in support of social amplification of risk. *Risk Analysis*, *22*(4), 701–711. https://doi.org/10.1111/0272-4332.00062

Frost, K., Frank, E., & Maibach, E. (1997). Relative risk in the news media: A quantification of misrepresentation. *American Journal of Public Health*, *87*(5), 842–845. https://doi.org/10.2105/AJPH.87.5.842

Galton, F. (1907). Vox populi. *Nature*, *75*(1949), 450–451. https://doi.org/10.1038/075450a0

Garcia-Marques, L., & Hamilton, D. L. (1996). Resolving the apparent discrepancy between the incongruency effect and the expectancy-based illusory correlation effect: The TRAP model. *Journal of Personality and Social Psychology*, *71*(5), 845–860. https://doi.org/10.1037/0022-3514.71.5.845

Garcia-Retamero, R., Galesic, M., & Gigerenzer, G. (2010). Do icon arrays help reduce denominator neglect? *Medical Decision Making*, *30*, 672–684. https://doi.org/10.1177/0272989X10369000

Garrido, M. (2003). Afinal o que fazem os simpáticos e os inteligentes? Um pré teste de descrições comportamentais. *Laboratório de Psicologia*, *1*(1), 45–55. https://doi.org/10.14417/lp.771

Gauld, A., & Stephenson, G. M. (1967). Some experiments relating to Bartlett’s theory of remembering. *British Journal of Psychology*, *58*(1), 39–49.

https://doi.org/10.1111/j.2044-8295.1967.tb01054.x

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond “heuristics and biases.” *European Review of Social Psychology*, *2*(1), 83–115. https://doi.org/10.1080/14792779143000033

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review*, *103*(3), 592–596. https://doi.org/10.1037/0033-295x.103.3.592

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition*, *43*(2), 127–171. https://doi.org/10.1016/0010-0277(92)90060-u

Ginossar, Z., & Trope, Y. (1987). Problem solving in judgment under uncertainty. *Journal of Personality and Social Psychology*, *52*(3), 464–474. https://doi.org/10.1037/0022-3514.52.3.464

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*(1), 75–90. https://doi.org/10.1037/0033-295x.109.1.75

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Chichester, UK: Wiley. https://doi.org/10.1016/s1364-6613(02)02011-9

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, *35*(3), 472–482. https://doi.org/10.1086/586910

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in*

Cognitive Sciences, *6*(12), 517–523. https://doi.org/10.1016/s1364-6613(02)02011-9

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*(4), 553–561. https://doi.org/10.1037/pspa0000016

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037//0033-295x.108.4.814

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, *12*, 392–407. https://doi.org/10.1016/s0022-1031(76)80006-6

Hamilton, D. L., & Rose, T. L. (1980). Illusory correlation and the maintenance of stereotypic beliefs. *Journal of Personality and Social Psychology*, *39*(5), 832–845. https://doi.org/10.1037/0022-3514.39.5.832

Hamilton, D. L., & Sherman, S. J. (1989). Illusory correlations: Implications for stereotype theory and research. In D. Bar-Tal, C. F. Graumann, A. W. Kruglanski, & W. Stroebe (Eds.), *Stereotyping and Prejudice* (pp. 59–82). New York, NY: Springer.

Hayes, A. F. (2013). *Introduction to meditation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: The Guilford Press.

Hertwig, R., Pachur, T., & Kurzenhäuser, S. (2005). Judgments of risk frequencies: Tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning Memory and Cognition*, *31*(4), 621–642. https://doi.org/10.1037/0278-7393.31.4.621

Higgins, E. T. (1992). Achieving “shared reality” in the communication game: A social action that creates meaning. *Journal of Language and Social Psychology*, *11*(3), 107–131. https://doi.org/10.1177/0261927x92113001

Hogarth, R. M. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin*, *90*(2), 197–216. https://doi.org/10.1037//0033-2909.90.2.197

Hogarth, R. M. (2005). The challenge of representative design in psychology and economics. *Journal of Economic Methodology*, *12*(2), 254–263. https://doi.org/10.1080/13501780500086172

Hoyert, D. L., & Xu, J. (2012). Deaths: preliminary data for 2011. *National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, *61*(6), 1–51. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24984457

Hunt, R. R., & Mitchell, D. B. (1982). Independent effects of semantic and nonsemantic distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(1), 81–87. https://doi.org/10.1037/0278-7393.8.1.81

Hunzaker, M. B. F. (2014). Making sense of misfortune: Cultural schemas, victim redefinition, and the perpetuation of stereotypes. *Social Psychology Quarterly*, *77*(2), 166–184. https://doi.org/10.1177/0190272514521219

Hunzaker, M. B. F. (2016). Cultural sentiments and schema-consistency bias in information transmission. *American Sociological Review*, *81*(6), 1223–1250. https://doi.org/10.1177/0003122416671742

Huq, S. F., Garety, P. A., & Hemsley, D. R. (1988). Probabilistic judgements in deluded and non-deluded subjects. *The Quarterly Journal of Experimental Psychology*, *40*(4), 801–812. https://doi.org/10.1080/14640748808402300

Isen, A. M., & Daubman, K. A. (1984). The influence of affect on categorization. *Journal of Personality and Social Psychology*, *47*(6), 1206–1217. https://doi.org/10.1037/0022-3514.47.6.1206

Jacobs, J. E., & Klaczynski, P. A. (2002). The development of judgment and decision making during childhood and adolescence. *Current Directions in Psychological Science*, *11*(4), 145–149. https://doi.org/10.1111/1467-8721.00188

Jacobs, R. C., & Campbell, D. T. (1961). The perpetuation of an arbitrary tradition through several generations of a laboratory microculture. *Journal of Abnormal and Social Psychology*, *62*(3), 649–658. https://doi.org/10.1037/h0044182

Jagiello, R. D., & Hills, T. T. (2018). Bad news has wings: Dread risk mediates social amplification in risk communication. *Risk Analysis*, *38*(10), 2193–2207. https://doi.org/10.1111/risa.13117

Janis, I. (1991). Groupthink. In E. Griffin (Ed.), *A first look at communication theory* (pp. 235–246). New York, NY: McGraw-Hill.

Johnston, L., & Hewstone, M. (1992). Cognitive models of stereotype change: 3. Subtyping and the perceived typicality of disconfirming group members. *Journal of Experimental Social Psychology*, *28*(4), 360–386. https://doi.org/10.1016/0022-1031(92)90051-K

Jones, D. A., & Skarlicki, D. P. (2005). The effects of overhearing peers discuss an authority’s fairness reputation on reactions to subsequent treatment. *Journal of Applied Psychology*, *90*(2), 363–372. https://doi.org/10.1037/0021-9010.90.2.363

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697–720. https://doi.org/10.1037/0003-066X.58.9.697

Kahneman, D. (2011). *Thinking fast and slow*. New York, NY: Farrar, Straus, and Giroux.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.21236/ada045771

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge. *Psychonomic Bulletin & Review*, *14*(M), 288–294. https://doi.org/10.3758/BF03194066

Kang, J. (2005). Trojan horses of race. *Harvard Law Review*, *118*(5), 1489–1593. Retrieved from http://www.jstor.org/stable/4093447

Kashima, Y. (2000). Maintaining cultural stereotypes in the serial reproduction of narratives. *Personality and Social Psychology Bulletin*, *26*(5), 594–604. https://doi.org/10.1177/0146167200267007

Kasperson, R. E., Renn, O., Slovic, P., Brown, H. S., Emel, J., Goble, R., … Ratick, S. (1988). The social amplification of risk: A conceptual framework. *Risk Analysis*, *8*(2), 177–187. https://doi.org/10.1111/j.1539-6924.1988.tb01168.x

Keren, G., & Teigen, K. H. (2004). Yet another look at the heuristics and biases approach. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* (pp. 89–109). Oxford, England: Blackwell Publishing Ltd.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*(1), 623–655. https://doi.org/10.1146/annurev.psych.55.090902.142009

King, R. N., & Koehler, D. J. (2000). Illusory correlations in graphological inference. *Journal of Experimental Psychology: Applied*, *6*(4), 336–348. https://doi.org/10.1037/1076-898X.6.4.336

Kirkpatrick, L. A., & Epstein, S. (1992). Cognitive-experiential self-theory and subjective probability: Further evidence for two conceptual systems. *Journal of Personality and Social Psychology*, *63*(4), 534–544. https://doi.org/10.1037//0022-3514.63.4.534

Knobloch-Westerwick, S., Mothes, C., & Polavin, N. (2017). Confirmation bias, ingroup bias, and negativity bias in selective exposure to political information. *Communication Research*. https://doi.org/10.1177/0093650217719596

Korobkin, R. (2011). What comes after victory for behavioral law and economics? *University of Illinois Law Review*, *2011*(5), 1653–1674. Retrieved from https://www.illinoislawreview.org/wp-content/ilr-content/articles/2011/5/Korobkin.pdf

Kuran, T., & Sunstein, C. R. (1999). Availability cascades and risk regulation. *Stanford Law Review*, *51*(4), 683–768. https://doi.org/10.2307/1229439

Laibson, D., & Zeckhauser, R. (1998). Amos Tversky and the ascent of behavioral economics. *Journal of Risk and Uncertainty*, *16*, 7–47. https://doi.org/10.1023/A:1007717224343

Landy, J. F., & Royzman, E. B. (2018). The moral myopia model: Why and how reasoning matters in moral judgment. In G. Pennycook (Ed.), *The new reflectionism in cognitive psychology: Why reason matters* (pp. 70–92). New York, NY: Psychology Press.

Laski, E. V., & Siegler, R. S. (2014). Learning from number board games: You learn what you encode. *Developmental Psychology*, *50*(3), 853–864. https://doi.org/10.1037/a0034321

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*(3), 177–189. https://doi.org/10.1016/0022-1031(86)90022-3

Lee, T. L., Gelfand, M. J., & Kashima, Y. (2014). The serial reproduction of conflict: Third parties escalate conflict through communication biases. *Journal of Experimental Social Psychology*, *54*, 68–72. https://doi.org/10.1016/j.jesp.2014.04.006

Lefebvre, M., Vieider, F. M., & Villeval, M. C. (2011). The ratio bias phenomenon: Fact or artifact? *Theory and Decision*, *71*(4), 615–641. https://doi.org/10.1007/s11238-010-9212-9

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. https://doi.org/10.1037/0033-2909.125.2.255

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 551–578. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0096151507603168

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*(6), 1231–1243. https://doi.org/10.1037/0022-3514.47.6.1231

Lyons, A., & Kashima, Y. (2001). The reproduction of culture: Communication processes tend to maintain cultural stereotypes. *Social Cognition*, *19*(3), 372–394. https://doi.org/10.1521/soco.19.3.372.21470

Lyons, A., & Kashima, Y. (2003). How are stereotypes maintained through communication? The influence of stereotype sharedness. *Journal of Experimental Psychology: Attitudes and Social Cognition*, *85*(6), 989–1005. https://doi.org/10.1037/0022-3514.85.6.989

Mandel, D. R. (2014). Do framing effects reveal irrational choice? *Journal of Experimental Psychology: General*, *143*(3), 1185–1198. https://doi.org/10.1037/a0034207

Mason, W. A., Conrey, F. R., & Smith, E. R. (2007). Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and Social Psychology Review*, *11*(3), 279–300. https://doi.org/10.1177/1088868307301032

Mata, A., & Ferreira, M. B. (2018). Response: Commentary: Seeing the conflict: An attentional account of reasoning errors. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00024

Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). Flexibility in motivated reasoning: Strategic shifts of reasoning modes in covariation judgment. *Social Cognition*, *31*(4), 465–481. https://doi.org/10.1521/soco_2012_1004

Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict: An attentional account of reasoning errors. *Psychonomic Bulletin & Review*, *24*(6), 1980–1986. https://doi.org/10.3758/s13423-017-1234-7

Mata, A., Fiedler, K., Ferreira, M. B., & Almeida, T. (2013). Reasoning about others' reasoning. *Journal of Experimental Social Psychology*, *49*(3), 486–491. https://doi.org/10.1016/j.jesp.2013.01.010

Mata, A., Schubert, A., & Ferreira, M. B. (2014). The role of language comprehension in reasoning: How “good-enough” representations induce biases. *Cognition*, *133*(2), 457–463. https://doi.org/10.1016/j.cognition.2014.07.011

Mata, A., Sherman, S. J., Ferreira, M. B., & Mendonça, C. (2015). Strategic numeracy: Self-serving reasoning about health statistics. *Basic and Applied Social Psychology*, *37*(3), 165–173. https://doi.org/10.1080/01973533.2015.1018991

Maxwell, R. S. (1936). Remembering in different social groups. *British Journal of Psychology: General Section*, *27*(1), 30–40. https://doi.org/10.1111/j.2044-8295.1936.tb00814.x

McGarty, C., Haslam, S. A., Turner, J. C., & Oakes, P. J. (1993). Illusory correlation as accentuation of actual intercategory difference: Evidence for the effect with minimal stimulus information. *European Journal of Social Psychology*, *23*(4), 391–410. https://doi.org/10.1002/ejsp.2420230406

McGuire, W. J. (1961). Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. *Journal of Abnormal and Social Psychology*, *63*(2), 326–332. https://doi.org/10.1037/h0048344

McKenzie, C. R. M. (2003). Rational models as theories - not standards - of behavior. *Trends in Cognitive Sciences*, *7*(9), 403–406. https://doi.org/10.1016/S1364-6613(03)00196-7

McNeil, N. M., & Alibali, M. W. (2004). You’ll see what you mean: Students encode equations based on their knowledge of arithmetic. *Cognitive Science*, *28*(3), 451–466. https://doi.org/10.1207/s15516709cog2803_7

Mercier, H., Majima, Y., & Miton, H. (2018). Willingness to transmit and the spread of pseudoscientific beliefs. *Applied Cognitive Psychology*, *32*(4), 499–505.

https://doi.org/10.1002/acp.3413

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74. https://doi.org/10.1017/s0140525x10000968

Mesoudi, A. (2007). Using the methods of experimental social psychology to study cultural evolution. *Journal of Social, Evolutionary, and Cultural Psychology*, *1*(2), 35–58. https://doi.org/10.1037/h0099359

Mesoudi, A., & Whiten, A. (2004). The hierarchical transformation of event knowledge in human cultural transmission. *Journal of Cognition and Culture*, *4*(1), 1–24. https://doi.org/10.1163/156853704323074732

Mesoudi, A., Whiten, A., & Dunbar, R. (2006). A bias for social information in human cultural transmission. *British Journal of Psychology*, *97*(3), 405–431. https://doi.org/10.1348/000712605X85871

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., & De Neys, W. (2015). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology*, *27*(2), 227–237. https://doi.org/10.1080/20445911.2014.986487

Miller, D. T., Turnbull, W., & McFarland, C. (1989). When a coincidence is suspicious: The role of mental simulation. *Journal of Personality and Social Psychology*, *57*(4), 581–589. https://doi.org/10.1037/0022-3514.57.4.581

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, *12*(2), 125–135. https://doi.org/10.1037/h0027568

Moussaïd, M., Brighton, H., & Gaissmaier, W. (2015). The amplification of risk in experimental diffusion chains. *Proceedings of the National Academy of Sciences*, *112*(18), 5631–5636. https://doi.org/10.1073/pnas.1421883112

Mullen, B., & Johnson, C. (1990). Distinctiveness-based illusory correlations and stereotyping: A meta-analytic integration. *British Journal of Social Psychology*, *29*, 11–28. https://doi.org/10.1111/j.2044-8309.1990.tb00883.x

Murphy, R. A., Schmeer, S., Vallée-Tourangeau, F., Mondragón, E., & Hilton, D. (2011). Making the illusory correlation effect appear and then disappear: The effects of increased learning. *Quarterly Journal of Experimental Psychology*, *64*(1), 24–40. https://doi.org/10.1080/17470218.2010.493615

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, *83*(4), 602–627. https://doi.org/10.1037/0033-2909.83.4.602

Navarro, D. J., Perfors, A., Kary, A., Brown, S. D., & Donkin, C. (2018). When extremists win: Cultural transmission via iterated learning when populations are heterogeneous. *Cognitive Science*, *42*(7), 2108–2149. https://doi.org/10.1111/cogs.12667

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Northway, M. L. (1936). The influence of age and social group on children's remembering. *British Journal of Psychology: General Section*, *27*(1), 11–29. https://doi.org/10.1111/j.2044-8295.1936.tb00813.x

Ortmann, A., & Gigerenzer, G. (1997). Reasoning in economics and psychology: Why social context matters. *Journal of Institutional and Theoretical Economics*, *153*, 701–710. https://doi.org/10.1007/978-3-642-59783-1_10

Paul, I. H. (1959). Studies in remembering. *Psychological Issues*, *1*(2), 1–152. Retrieved from https://archive.org/details/psychologicaliss00erik/page/n695

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition*, *124*(1), 101–106. https://doi.org/10.1016/j.cognition.2012.04.004

Pinto-Prades, J.-L., Martinez-Perez, J.-E., & Abellán-Perpiñán, J.-M. (2006). The influence of the ratio bias phenomenon on the elicitation of health states utilities. *Judgment and Decision Making*, *1*(2), 118–133. Retrieved from http://journal.sjdm.org/06126/jdm06126.htm

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811–832. https://doi.org/10.1037/t03881-000

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology*, *34*(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*(1), 160–164. https://doi.org/10.1037/0033-2909.112.1.160

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the Cognitive Reflection Test applying Item Response Theory (IRT). *Journal of Behavioral Decision Making*, *29*(5), 453–469. https://doi.org/10.1002/bdm.1883

Pronin, E., Gilovich, T., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369–381. https://doi.org/10.1177/0146167202286008

Pryor, J. B., & Stoller, L. M. (1994). Sexual cognition processes in men high in the likelihood to sexually harass. *Personality and Social Psychology Bulletin*, *20*(2), 163–169. https://doi.org/10.1177/0146167294202003

Pyszczynski, T., & Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. *Advances in Experimental Social Psychology*, *20*, 297–340. https://doi.org/10.1016/s0065-

2601(08)60417-7

Rachlinski, J. J. (2000). Heuristics and biases in the court: Ignorance or adaptation? *Cornell Law Faculty Publications*. Retrieved from scholarship.law.cornell.edu/facpub/810

Rawson, K. A., & Van Overschelde, J. P. (2008). How does knowledge promote memory? The distinctiveness theory of skilled memory. *Journal of Memory and Language*, *58*(3), 646–668. https://doi.org/10.1016/j.jml.2007.08.004

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328. https://doi.org/10.1016/j.cognition.2009.02.012

Reno, R. R., Cialdini, R. B., & Kallgren, C. A. (1993). The transsituational influence of social norms. *Journal of Personality and Social Psychology*, *64*(1), 104–112. https://doi.org/10.1037/0022-3514.64.1.104

Reyes, R. M., Thompson, W. C., & Bower, G. H. (1980). Judgmental biases resulting from differing availabilities of arguments. *Journal of Personality and Social Psychology*, *39*(1), 2–12. https://doi.org/10.1037/0022-3514.39.1.2

Reyna, V. F. (1991). Class inclusion, the conjunction fallacy, and other cognitive illusions. *Developmental Review*, *11*, 317–336. https://doi.org/10.1016/0273-2297(91)90017-I

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*, 89–107. https://doi.org/10.1016/j.lindif.2007.03.011

Rodríguez-Ferreiro, J., & Barberia, I. (2017). The moral foundations of illusory correlation. *PLoS ONE*, *12*(10), 1–10. https://doi.org/10.1371/journal.pone.0185758

Roediger, H. L. (1997). Remembering. *Contemporary Psychology*, *42*(6), 488–492. https://doi.org/10.1037/000282

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., Meade, M. L., Gallo, D. A., & Olson, K. R. (2014). Bartlett revisited: Direct comparison of repeated reproduction and serial reproduction techniques. *Journal of Applied Research in Memory and Cognition*, *3*(4), 266–271. https://doi.org/10.1016/j.jarmac.2014.05.004

Roediger, H. L., Zaromb, F. M., & Butler, A. C. (2009). The role of repeated retrieval in shaping collective memory. In P. Boyer & J. Wertsch (Eds.), *Memory in Mind and Culture* (pp. 138–170). Cambridge, UK: Cambridge University Press.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Rumbaut, R. G., & Ewing, W. A. (2007). The myth of immigrant criminality and the paradox of assimilation. *American Immigration Law Foundation: Immigration Policy Center*. Retrieved from https://www.americanimmigrationcouncil.org/sites/default/files/research/Imm Criminality %28IPC%29.pdf

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, *311*(5762), 854–856. https://doi.org/10.1126/science.1121066

Schaller, M., & Maass, A. (1989). Illusory correlation and social categorization: Toward an integration of motivational and cognitive factors in stereotype formation. *Journal of*

Personality and Social Psychology, *56*(5), 709–721. https://doi.org/10.1037/0022-3514.56.5.709

Schelling, T. C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology*, *1*(2), 143–186. https://doi.org/10.1080/0022250X.1971.9989794

Schmidt, S. R., & Saari, B. (2007). The emotional memory effect: Differential processing or item distinctiveness? *Memory and Cognition*, *35*(8), 1905–1916. https://doi.org/10.3758/BF03192924

Schultze, T., Rakotoarisoa, A., & Schulz-Hardt, S. (2015). Effects of distance between initial estimates and advice on advice utilization. *Judgment and Decision Making*, *10*(2), 144–171. Retrieved from http://journal.sjdm.org/vol10.2.html

Schwarz, N., Strack, F., Hilton, D., & Naderer, G. (1991). Base rates, representativeness, and the logic of conversation: The contextual relevance of "irrelevant" information. *Social Cognition*, *9*(1), 67–84. https://doi.org/10.1521/soco.1991.9.1.67

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: The University of Illinois Press.

Shanteau, J. (1978). When does a response error become a judgmental bias? Commentary on "Judged frequency of lethal events." *Journal of Experimental Psychology: Human Learning & Memory*, *4*(6), 579–581. https://doi.org/10.1037//0278-7393.4.6.579

Shedler, J., & Manis, M. (1986). Can the availability heuristic explain vividness effects? *Journal of Personality and Social Psychology*, *51*(1), 26–36. https://doi.org/10.1037/0022-3514.51.1.26

Shoemaker, P. J. (1987). Mass communication by the book: A review of 31 texts. *Journal of Communication*, *37*(3), 109–131. https://doi.org/10.1111/j.1460-2466.1987.tb00998.x

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*(4), 481–520. https://doi.org/10.1016/0010-0285(76)90016-5

Simonson, I., & Staw, B. M. (1992). Deescalation strategies: A comparison of techniques for reducing commitment to losing courses of action. *Journal of Applied Psychology*, *77*(4), 419–426. https://doi.org/10.1037/0021-9010.77.4.419

Smith, C. M., Tindale, R. S., & Steiner, L. (1998). Investment decisions by individuals and groups in "sunk cost" situations: The potential impact of shared representations. *Group Processes & Intergroup Relations*, *1*(2), 175–189. https://doi.org/10.1177/1368430298012005

Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, *11*(1), 87–104. https://doi.org/10.1177/1088868306294789

Smith, M. R., & Alpert, G. P. (2007). Explaining police bias: A theory of social conditioning and illusory correlation. *Criminal Justice and Behavior*, *34*(10), 1262–1283. https://doi.org/10.1177/0093854807304484

Smith, S. M., & Shaffer, D. R. (2000). Vividness can undermine or enhance message processing: The moderating role of vividness congruency. *Personality and Social Psychology Bulletin*, *26*(7), 769–779. https://doi.org/10.1177/0146167200269003

Smither, J. W., Collins, H., & Buda, R. (1989). When ratee satisfaction influences performance evaluations: A case of illusory correlation. *Journal of Applied Psychology*, *74*(4), 599–605. https://doi.org/10.1037/0021-9010.74.4.599

Soll, J. B., Milkman, K. L., & Payne, J. W. (2016). A user's guide to debiasing. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making* (pp. 924–951). Chichester, UK: Wiley.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161–188. https://doi.org/10.1037/0096-3445.127.2.161

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*(4), 259–264. https://doi.org/10.1177/0963721413480174

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

Strough, J., Karns, T. E., & Schlosnagle, L. (2011). Decision-making heuristics and biases across the life span. *Annals of the New York Academy of Sciences*, *1235*, 57–74. https://doi.org/10.1111/j.1749-6632.2011.06208.x

Sunstein, C. R., & Zeckhauser, R. (2011). Overreaction to fearsome risks. *Environmental and Resource Economics*, *48*(3), 435–449. https://doi.org/10.1007/s10640-010-9449-3

Surowiecki, J. (2004). *The wisdom of crowds*. New York, NY: Doubleday.

Talland, G. A. (1956). Cultural differences in serial reproduction. *Journal of Social Psychology*, *43*(1), 75–81. https://doi.org/10.1080/00224545.1956.9919201

Thaler, R. H. (2017). From cashews to the evolution of behavioral economics stories and thought experiments circa 1970s. *American Economic Review*, *108*(6), 1265–1287. https://doi.org/10.1257/aer.108.6.1265

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge*. New York, NY: Penguin Group.

Thompson, M. S., Judd, C. M., & Park, B. (2000). The consequences of communicating social stereotypes. *Journal of Experimental Social Psychology*, *36*(6), 567–599.

https://doi.org/10.1006/jesp.1999.1419

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113. Retrieved from http://journal.sjdm.org/vol11.1.html

Tindale, R. S. (1993). Decision errors made by individuals and groups. In N. J. J. Castellan (Ed.), *Individual and group decision making: Current issues* (pp. 109–124). Hillsdale, NJ: Erlbaum.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*(5), 1–38. https://doi.org/10.18637/jss.v059.i05

Tomarken, A. J., Mineka, S., & Cook, M. (1989). Fear-relevant selective associations and covariation bias. *Journal of Abnormal Psychology*, *98*(4), 381–394. https://doi.org/10.1037/0021-843X.98.4.381

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971. https://doi.org/10.1037/a0037099

Turk, D. C., & Salovey, P. (1986). Clinical information processing: Bias inoculation. In R. E. Ingram (Ed.), *Information processing approaches to clinical psychology* (pp. 305–323). San Diego, CA: Academic Press.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P.

Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). New York, NY: Cambridge University Press.

Ucbasaran, D., Westhead, P., Wright, M., & Flores, M. (2010). The nature of entrepreneurial experience, business failure and comparative optimism. *Journal of Business Venturing*, *25*(6), 541–555. https://doi.org/10.1016/j.jbusvent.2009.04.001

van Prooijen, J. W., & Krouwel, A. P. M. (2017). Extreme political beliefs predict dogmatic intolerance. *Social Psychological and Personality Science*, *8*(3), 292–300. https://doi.org/10.1177/1948550616671403

Von Restorff, H. (1933). Über die wirkung von bereichsbildungen im spurenfeld. *Psychologische Forschung*, *18*(1), 299–342. https://doi.org/10.1007/bf02409636

Wagoner, B. (2015). Qualitative experiments in psychology: The case of Frederic Bartlett’s methodology. *Forum Qualitative Sozialforschung*, *16*(3), 1885–1910. Retrieved from http://nbn-resolving.de/urn:nbn:de:0114-fqs1503239

Wallsten, T. S. (1983). The theoretical status of judgmental heuristics. *Advances in Psychology*, *16*(C), 21–37. https://doi.org/10.1016/S0166-4115(08)62192-5

Walster, E., & Festinger, L. (1962). The effectiveness of “overheard” persuasive communications. *Journal of Abnormal and Social Psychology*. https://doi.org/10.1037/h0041172

Ward, T. H. G. (1949). An experiment on serial reproduction with special reference to the changes in the design of early coin types. *British Journal of Psychology: General Section*, *39*(3), 142–147. https://doi.org/10.1111/j.2044-8295.1949.tb00213.x

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard’s (1913) and Bartlett’s (1932) Results. *Psychological Science*, *3*(4), 240–245. https://doi.org/10.1111/j.1467-9280.1992.tb00036.x

Whiten, A., & Mesoudi, A. (2008). Establishing an experimental science of culture: Animal social diffusion experiments. *Philosophical Transactions of the Royal Society B*, *363*, 3477–3488. https://doi.org/10.1098/rstb.2008.0134

Wiemer, J., & Pauli, P. (2016). Fear-relevant illusory correlations in different fears and anxiety disorders: A review of the literature. *Journal of Anxiety Disorders*, *42*, 113–128. https://doi.org/10.1016/j.janxdis.2016.07.003

Wilensky, U. (1999). NetLogo. Evanston, IL: Center for Connected Learning and Computer-Based Modeling, Northwestern University. Retrieved from http://ccl.northwestern.edu/netlogo/

Wilson, E. J., & Sherrell, D. L. (1993). Source effects in communication and persuasion research: A meta-analysis of effect size. *Journal of the Academy of Marketing Science*, *21*(2), 101–112. https://doi.org/10.1007/BF02894421

Wu, C., & Shaffer, D. R. (1987). Susceptibility to persuasive appeals as a function of source credibility and prior experience with the attitude object. *Journal of Personality and Social Psychology*, *52*(4), 677–688. https://doi.org/10.1037/0022-3514.52.4.677

Yaniv, I. (2004a). Receiving other people’s advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*, 1–13. https://doi.org/10.1016/j.obhdp.2003.08.002

Yaniv, I. (2004b). The benefit of additional opinions. *Current Directions in Psychological Science*, *13*(2), 75–78. https://doi.org/10.1111/j.0963-7214.2004.00278.x

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281. https://doi.org/10.1006/obhd.2000.2909