

# Stereotypes

Pedro Bordalo, Nicola Gennaioli, Andrei Shleifer\*

First draft, November 2013.

## Abstract

We present a model of stereotypes in which a decision maker assessing a group recalls only that group's most representative or distinctive types relative to other groups. Because stereotypes highlight differences between groups, and neglect likely common types, they are especially inaccurate when groups are fairly similar. In this case, stereotypes consist of unlikely, extreme types. When stereotypes are inaccurate, they exhibit a form of base rate neglect. They also imply a form of confirmation bias in light of new information: beliefs overreact to information that confirms the stereotype and ignore information that contradicts it. However, stereotypes can change – or rather, be replaced – if new information changes the group's most distinctive trait.

---

\*Royal Holloway, University of London, Università Bocconi and IGIER, Harvard University. We are grateful to Sendhil Mullainathan and Josh Schwartzstein for extremely helpful comments and to Rohan Kekre for research assistance.

# 1 Introduction

The Oxford English Dictionary defines a stereotype as a “widely held but fixed and oversimplified image or idea of a particular type of person or thing”. Stereotypes are ubiquitous. Among other things, they cover racial groups (“Asians are good at math”), political groups (“republicans are creationists”), genders (“male drivers are aggressive”), and activities (“flying is dangerous”).

Stereotypes play an important cognitive role. As stressed by psychologists, stereotypes “... are mental representations of real differences between groups [...] allowing easier and more efficient processing of information. [S]tereotypes are selective, however, in that they are localized around group features that are the most distinctive, that provide the greatest differentiation between groups, and that show the least within-group variation” (Hilton and von Hippel 1996). While stereotypes allow for a quick and intuitive assessment of groups, they may also cause distorted judgment and biased behavior, such as discrimination and inter-group conflict. The nature of stereotypes is not completely understood and there are many open questions: How do stereotypes form? How do they affect beliefs and actions? Why do some stereotypes have a reasonable amount of validity (“men are aggressive drivers”), while others have much less (“flying is dangerous”)? How do stereotypes change?

We present a cognitive model of stereotype and belief formation building on Gennaioli and Shleifer’s (GS, 2010) paper “What comes to mind”. Stereotypes come quickly and automatically to mind – they are readily available in memory when evaluating a group – and are identified by the group’s “representative feature or type”. The representative type is defined as the type that most distinguishes the group in question from other groups, in line with Kahneman and Tversky’s (1972) notion that “an attribute is representative of a class if it is very diagnostic; that is, the relative frequency of this attribute is much higher in that class than in a relevant reference class.”

Formally, we follow GS (2010) in assuming that the most representative realization of a type  $x$  in a group  $S$  is defined as the value  $x_1$  that is most diagnostic of  $S$  relative to a

comparison group  $-S$ , in that it maximizes the likelihood ratio:

$$x_1 = \operatorname{argmax}_x \frac{\Pr(S|x)}{\Pr(-S|x)}.$$

Because of limited working memory, only the most representative types are recalled and used in judgments, be it for inference or for prediction. The stereotype is then formed by truncating the true probability distribution  $\Pr(x|S)$  to the  $d \geq 1$  most representative group types  $x_1, \dots, x_d$ . The trademark of stereotypes is the neglect of non-representative types, not the distortion of the relative probabilities of the types that do come to mind. Relative to a Bayesian, distortions in beliefs can be drastic, but only when the recalled states are not the most likely ones.

The concept of representativeness is critical to thinking about stereotypes. A type  $x$  is representative of a group  $S$  if, after observing  $x$ , a Bayesian DM assessing the likelihood of  $S$  relative to  $-S$  would increase the odds that he is facing  $S$ . Informally, we would say that a type  $x$  is representative of  $S$  if it is ex-post informative about  $S$ . If a decision maker observes a stereotypical type (a creationist, an aggressive driver), he would significantly update the probability that the observation comes from the stereotyped group (republicans, males). Crucially, this implies that stereotypes do not necessarily include types which are frequent in the group, namely types that maximize  $\Pr(x|S)$ . In fact, ex-post informative observations often come from the tails of distributions, which are unlikely to occur ex-ante.

To illustrate this logic, consider the formation of the stereotypes “Republicans are creationists” and “Democrats believe in Evolution”. In May 2012, Gallup conducted a public opinion poll assessing the beliefs about Evolution of members of the two main parties in the US. The results on the beliefs of Republicans and Democrats, largely unchanged in the three decades over which such polls have been conducted, are presented below:<sup>1</sup>

---

<sup>1</sup>The three options were described as “God created Humans in present form in the last 10,000 years”, “Humans evolved, God has no part in process” and “Humans evolved, God guided the process”. See <http://www.gallup.com/poll/155003/Hold-Creationist-View-Human-Origins.aspx> for details.

	<i>Creationism</i>	<i>Evolution</i>	<i>Evolution guided by God</i>
<i>Republicans</i>	58%	5%	31%
<i>Democrats</i>	41%	19%	32%

The table shows that being a creationist is the distinguishing feature of the Republicans, not only because most Republicans are creationist but also because more Republicans are creationists than Democrats. In this sense, stereotyping a Republican as a creationist yields a fairly accurate assessment. Formally,  $x = \textit{Creationism}$  maximizes not only  $\Pr(\textit{Republicans}|x)/\Pr(\textit{Democrats}|x)$  but also  $\Pr(x|\textit{Republicans})$ .

On the other hand, the distinguishing feature of the Democrats is to believe in the “standard” Darwinian Evolution of humans, a belief four times more prevalent than it is among Republicans. However, and perhaps surprisingly, only 19% of Democrats believe in Evolution. Most of them believe either in creationism (41%) or in Evolution guided by God (32%), just like Republicans do. Formally,  $x = \textit{Evolution}$  maximizes  $\Pr(\textit{Democrats}|x)/\Pr(\textit{Republicans}|x)$  but not  $\Pr(x|\textit{Democrats})$ . Evolution is not the most likely belief of Democrats, but rather the belief that occurs with the highest relative frequency. As a consequence, a stereotype-based prediction that a Democrat would believe in the standard evolutionary account of human origins, and would not believe in Creationism, is a bad prediction.

The predictive power of stereotypes turns critically on the difference between likelihood and representativeness (see GS 2010). Stereotypes of groups are good predictors when representativeness and likelihood rankings are consistent, as in the case of a creationist Republican. But stereotypes are bad predictors when representativeness differs from likelihood and stereotypes are biased towards ex-post informative but ex-ante unlikely types, as in the case of a democrat who believes in Darwinian evolution.

In this paper, we develop a theory of stereotypes along these lines. The theory has several implications, including the following:

- Whether stereotypes are accurate (recalling likely types) or inaccurate (recalling unlikely types) depends on the underlying distribution of group types. Because stereotypes highlight the differences between groups, they are inaccurate when groups are

fairly similar and differ only in the tails. Our theory thus explains why stereotypes are often extremely unlikely, as in “Arabs are terrorists.”<sup>2</sup>

- In many circumstances, stereotypes do not just emphasize differences across groups, they also minimize within group variability.
- To the extent that representativeness does not depend on the base rates of types in the overall population, group stereotypes may exhibit specific types of neglect of base rates. Because different types are neglected for different groups, this mechanism is distinct from – and yields different predictions than – a model where the impact of base rates on Bayesian updates is dampened.
- When group features are multidimensional, stereotypes are formed using the dimension along which the groups being compared differ the most. Football players in Ivy League schools may be stereotyped as “less smart” in the population of Ivy League students, but as “less athletic” in the population of all football players (including professionals). Stereotypes leave room for substantial variability along other dimensions.
- The model makes strong predictions about how decision makers thinking with stereotypes react to information. It predicts that, so long as stereotypes do not change, people display a form of confirmation bias in that they over-react to information consistent with stereotypes, and under-react or even ignore information inconsistent with stereotypes. Base rate neglect and confirmation bias are thus two sides of the same coin of representativeness based recall.
- Stereotypes can change – or rather, be replaced – if sufficient contrary information is received (e.g. observing sufficiently more creationist Democrats than creationist

---

<sup>2</sup>A Gallup poll conducted shortly after the 1993 terrorist bombing of the World Trade Center found that “majorities of Americans said the following terms applied to Arabs: religious (81%), terrorists (59%), violent (58%) and religious fanatics (56%). Related, a recent poll by Pew’s Global Attitudes Project found that Westerners view Muslims as fanatical (58% of respondents) and violent (50%), while Muslims view Westerners as selfish (68%), violent (66%) and greedy (64%). Curiously, selfishness and greed are among the traits that Westerners least associate with Muslims. Sources: <http://www.gallup.com/poll/4939/Americans-Felt-Uneasy-Toward-Arabs-Even-Before-September.aspx> and <http://www.pewglobal.org/2011/07/21/muslim-western-tensions-persist/>.

Republicans), or if an entirely different feature becomes more representative (e.g. observing multiple Democrats denouncing the rich). A change of stereotypes then leads to a drastic reevaluation of already available data. We note, however, that more information does not necessarily lead to a better (more likely) stereotype.

Since Kahneman and Tversky's (1972, 1974) work on heuristics and biases, there have been several attempts to formally model heuristics about probabilistic judgments and to incorporate them in economic models. This includes work on the confirmation bias (Rabin and Schrag 1999) and on biased probabilistic extrapolation (Grether 1980, Barberis, Shleifer, and Vishny 1998, Rabin 2002, Rabin and Vayanos 2010, Benjamin, Rabin and Raymond 2011), in which the DM has an incorrect model in mind or incorrectly processes the data available to him. Our approach is instead based on the single assumption that not all information comes to the DM's mind when making judgments. The fact that the DM neglects some information is a form of simplifying the judgment problem that is reminiscent of models of categorization, such as Mullainathan (2002) and Fryer and Jackson (2007). In these models, however, decision makers represent reality with the aid of categories that are organized according to likelihood, not representativeness (as defined here): these categories are best suited to describe likely or frequent events, and sacrifice precision in the representation of unlikely events. This coarsening generates imprecision in assessing the likelihood of events but does not create a systematic bias for overestimating unlikely events, nor does it allow for a role of context in shaping the stereotypes of a group. Our approach is more closely related to our previous work on salience (BGS 2012, 2013), which we discuss further in Section 3.3. Finally, the notion of stereotypes also plays a role in models of statistical discrimination (Arrow 1973, Phelps 1972). In these models, stereotypes fill up for the lack of information about agents, though in equilibrium these stereotypes end up being accurate in expectation. DMs in our model have in principle full information, yet distort their beliefs about groups at the ex-ante stage and this distortion can emphasize tail, ex ante unlikely, types.

In the next section, we introduce the notion of representativeness in the context of categorical (discrete) distributions and describe our model of stereotypes. In Section 3, we explore the forces that shape stereotypes and their accuracy (in the sense of likelihood). Several examples that connect to well-known biases of judgment are described in this set-

ting. In Section 4, we describe how stereotypes respond to new information. Section 5 extends the analysis to continuous distributions. Section 6 concludes.

## 2 The Model

Gennaioli and Shleifer (GS, 2010) present a model of how decision makers (henceforth DM's) intuitively solve *inference* problems, i.e. of how they assess the probability that a hypothesis is true given some data. The model builds on the idea that DMs think about and thus assess each hypothesis by recalling only the hypothesis' most representative scenarios, relative to alternative hypotheses.<sup>3</sup> In this paper, we use the same principle of representativeness based recall to explore the *prediction* problem, namely how the DM intuitively represents the distribution of types of a group  $S$ . In the examples of the introduction,  $S$  is Asian, or Republican, or male, or flights.

Formally, a DM must assess the distribution of a categorical random variable  $X \in \{1, \dots, N\}$  in a given group  $S$ . The group  $S$  is a proper subset of the entire population  $\Omega$ , on which the random variable  $X$  is defined. In most of our examples  $S$  is a social group, for instance when the DM assesses occupational choices or educational attainment within a certain group (e.g.,  $X = \text{occupation}$  and  $S = \text{the French}$ ). However, the theory is broader, and  $S$  may reflect past history, such as when the DM evaluates the future prospects of an industry (e.g.,  $X = \text{future earnings}$ ,  $S = \text{high tech stocks}$ ), or categories in the natural world ( $X = \text{ability to fly}$ ,  $S = \text{birds}$ ). Realizations of  $X$  may be multi-dimensional, capturing a bundle of attributes (e.g., occupation and nationality), or they may be quantifiable and naturally ordered (e.g. years of education or income). We return to these possibilities, but for now we take  $X$  to be unordered and unidimensional.

We denote by  $(\pi_{x,S})_{x=1,\dots,N}$  the true conditional distribution  $Pr(X = x|S)$  of the random variable  $X$  in group  $S$  where  $x \in \{1, \dots, T\}$  indexes the possible types. We denote by  $\pi_x$  the true unconditional probability  $Pr(X = x)$  in  $\Omega$ .

As in GS (2010), the DM recalls only a subset of  $d \in \{1, \dots, T\}$  types. Parameter  $d$

---

<sup>3</sup>In GS (2010), an event  $x$  is defined to be more representative of a hypothesis  $h$  against an alternative  $-h$  if, conditional on  $x$ ,  $h$  is more likely to be true than  $-h$ .

captures the severity of working memory limits. When  $d = 1$ , memory limits are so severe that the DM recalls only one type for  $S$ . When  $d = T$ , there are no memory limits, and the agent recalls all possible types for  $S$ . Critically, recall is not just limited but also selective. For a given  $d$ , the recalled types are the most representative of group  $S$ , in the sense that they are most distinctive of  $S$  relative to other groups in  $\Omega$ . Following GS (2010), we formalize this notion as follows.

**Definition 1** *The representativeness of type  $x$  for group  $S$  is defined as  $R(x, S) = \Pr(S|X = x)/\Pr(-S|X = x)$ , where  $-S = \Omega/S$ . Bayes rule implies that representativeness increases in the likelihood ratio:*

$$\frac{\Pr(X = x|S)}{\Pr(X = x|-S)} = \frac{\pi_{x,S}}{\pi_{x,-S}}. \quad (1)$$

The definition states that a type  $x$  is representative of a group  $S$  if, after observing  $x$ , a Bayesian DM assessing the likelihood of  $S$  relative to  $-S$  would be more confident that he is facing  $S$ . We say that a type  $x$  is representative of  $S$  if it is ex-post informative about  $S$  in this sense. This role of ex post diagnosticity arises even though diagnosticity relative to  $-S$  is normatively irrelevant for the DM’s assessment of the distribution of types in  $S$ . Equation (1) shows how this notion of representativeness captures Kahneman and Tversky’s (1972) intuition: a type  $x$  is representative of  $S$  if it is relatively more likely to occur in  $S$  than in  $-S$ . For instance, when thinking about occupations held by members of a group  $S$ , our mind finds it easy to retrieve occupations that are more common in  $S$  relative to all other groups.

According to Definition 1, representativeness is a relative concept, that depends on which group  $-S$  the target group  $S$  is compared to, or equivalently on the set of possible groups  $\Omega$  (given that  $S = \Omega/S$ ). As a consequence, the specification of  $\Omega$  can influence the stereotype for a certain group  $S$ . For example, the stereotype of college athletes in the population of all college students might be “poor academic” or “very athletic”, and we examine in Section 3.4 which one of them obtains. If, however, the same college athletes are compared to professional athletes, their stereotype might be “not very athletic.” This observation raises the question of how  $\Omega$  is determined. In many cases,  $\Omega$  is specified by the problem itself. For



instance, in assessing the academic performance of a college athlete it is natural to compare him to other college students. We do not as of yet have a theory of what determines  $\Omega$  when it is not pinned down by the problem itself. One approach is to assume that in these cases  $\Omega$  is the comprehensive probability space upon which a rational DM would form his beliefs.

Definition 1 leads to the following property.

**Remark 1** *Suppose that  $\pi_{x,S} \geq \pi_{x,-S}$ . Then, the representativeness  $R(x,S)$  of type  $x$  for group  $S$ :*

- i) increases, for given baseline probability  $\pi_{x,-S}$ , in the difference  $(\pi_{x,S} - \pi_{x,-S})$ .*
- ii) decreases, for given difference  $\pi_{x,S} - \pi_{x,-S}$ , in the baseline probability  $\pi_{x,-S}$ .*

Our model captures the idea that the mind is attuned to perceive and recall differences, in the spirit of Weber’s law of sensory perception: property i) says that a type is more representative the more likely it is to occur under  $S$  than under  $-S$ .<sup>4</sup> Critically, though, a given increase in probability is more representative if it occurs in an infrequent type, making infrequent types *ceteris paribus* more representative. Property ii) thus captures a form of diminishing sensitivity. It implies that types or features that are not particularly likely *ex ante* can be very representative of a group  $S$ , precisely because such types are *ex post* informative.

The DM’s assessment of the distribution of group  $S$  works as follows:

**Definition 2** *Denote by  $r \in \{1, \dots, N\}$  the representativeness ranking of types, and denote by  $x_r$  the  $r$ -th most representative type for  $S$ . The DM forms his beliefs according to the modified probability distribution:*

$$\pi_{x_r,S}^{st} = \begin{cases} \frac{\pi_{x_r,S}}{\sum_{r'=1}^d \pi_{x_{r'},S}}, & \text{for } r \in \{1, \dots, d\}. \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

The DM’s beliefs about  $S$  consist of a truncated probability distribution on the  $d$  most representative types. We call the distribution  $(\pi_{x_r,S}^{st})_{r=1,\dots,d}$  the stereotype for  $S$ ; informally,

---

<sup>4</sup>This feature connects to our previous work on salience, which also builds on Weber’s law. In BGS (2012) we postulated that, in a choice among two lotteries, a lottery outcome is more salient when it entails: i) a larger payoff difference (ordering), and ii) a lower payoff level (diminishing sensitivity). In Definition 1, these same properties characterize recall in the domain of probabilities.

we also refer to the represented types,  $\{x_1, \dots, x_d\}$ , as the stereotype of  $S$ . In the extreme case where  $d = 1$ , the DM recalls only the most representative type  $x_1$  about group  $S$ , and assigns it probability  $\pi_{x_1, S}^{st} = 1$ . In this case, the stereotype consists only of what psychologists call an “exemplar” type. In less extreme, and perhaps more realistic cases,  $d > 1$  and the stereotype of  $S$  includes the exemplar and some less representative types, that psychologists call “subtypes.” When thinking about Italians, people will not just recall the fact that they eat pasta, but also that they dress elegantly.

According to Definition 2, the DM does not take into account the  $(T - d)$  least representative types. The DM does not view these types as impossible; rather, he neglects them in his assessments.<sup>5</sup> Put differently, these types are in the back of his mind, but the DM attaches a zero probability to them in his current thinking. This formulation opens up a way of modeling surprises or reactions to “zero probability events,” which we exploit in Section 4.

Stereotypes depend on “true” probabilities. Past experiences or information are stored in the DM’s long-term memory, but recall is incomplete and selective: only the most distinctive types of group  $S$  come to mind. Equation (1) shows that, conditional on coming to mind, the assessed odds ratios of any two types are consistent with the DM’s experience and information. This is in contrast with models where conditional probabilities are distorted through the action of some bias, for instance a direct neglect of base rates. We expand on the link to base rate neglect in Section 3.4. Finally, since past experiences or information may vary across individuals, our model allows for individual heterogeneity in stereotypes. This allows for stereotypes to be influenced by culture, for example (see Section 3.3).

---

<sup>5</sup>Aside from zero probability events, qualitatively similar properties can be obtained under a smooth discounting of less representative outcomes. Formally, given a weighting function  $\delta(\pi_{x, S}/\pi_{x, -S})$  which increases in the likelihood ratio (i.e.,  $\delta'(\cdot) > 0$ ) one can define:

$$\pi_{x, S}^{st} = \frac{\delta(\pi_{x, S}/\pi_{x, -S}) \cdot \pi_{k, S}}{\sum_k \delta(\pi_{k, S}/\pi_{k, -S}) \cdot \pi_{k, S}}$$

In this formulation, the probability of types that have a higher likelihood ratio is inflated.

### 3 Stereotype Formation

We now illustrate how the model sheds light on the formation of stereotypes. Section 3.1 begins with the benchmark case in which memory limitations are so severe that the stereotype coincides with the exemplar, namely  $d = 1$ . Section 3.2 considers the case in which recall is more (but not fully) complete. Section 3.3 discusses the broad properties of the model. Section 3.4 discusses the implications of our model to the literature on heuristics and biases. Section 3.5 presents the implications of the model when types are multidimensional.

#### 3.1 Exemplars: Representativeness versus Likelihood

Suppose that  $d = 1$ , so the stereotype for  $S$  is its exemplar, namely the most representative type (ties are resolved randomly). This extreme case entails a potentially large loss of information. Still, we can assess whether the recalled information is optimal given the constraints. Using any standard measure of the accuracy of an assessed distribution  $(\pi_{x,S}^{st})_{x=1,\dots,N}$  against a true distribution  $(\pi_{x,S})_{x=1,\dots,N}$ , for instance that given by the quadratic loss function  $L[(\pi_{x,S}^{st})_{x=1,\dots,N} | (\pi_{x,S})_{x=1,\dots,N}] = \sum_x (\pi_{x,S}^{st} - \pi_{x,S})^2$ , an exemplar is optimal if it collapses the stereotype  $(\pi_{x,S}^{st})_{x=1,\dots,N}$  on the most likely type of the true distribution  $(\pi_{x,S})_{x=1,\dots,N}$ . The accuracy of the exemplar decreases as it becomes less and less likely.

To explore the link between representativeness and likelihood, consider the case where the likelihood of types in group  $-S$  is a simple transformation of that in group  $S$ . The result below highlights the forces that determine the optimality of exemplars in this case:

**Proposition 1** *Let  $(\pi_{x,S})_{x=1,\dots,N}$  be the conditional distribution in group  $S$ , and suppose that the conditional distribution in the comparison group  $-S$  is defined by  $\pi_{x,-S} = \pi^* \cdot \pi_{x,S}^\alpha$  for all  $x$ , where  $\alpha$  is a real number and  $\pi^* = 1 / \sum_x \pi_{x,S}^\alpha$  is a normalizing constant. We then have:*

- i) If  $\alpha < 1$ , the exemplar for  $S$  is the most likely type (i.e.,  $\operatorname{argmax}_x \pi_{x,S}$ );*
- ii) If  $\alpha > 1$ , the exemplar for  $S$  is the least likely type (i.e.,  $\operatorname{argmin}_x \pi_{x,S}$ ).*

The parameter  $\alpha$  controls the relationship between the distributions for groups  $S$  and  $-S$ . If  $\alpha > 0$ , group  $-S$  has the same likelihood ranking of types as group  $S$ ; the distributions

are “similar” and in particular have the same modal type. If in addition  $\alpha > 1$ , then group  $-S$  is more concentrated around the mode than group  $S$ , which has fatter tails.

In broad terms, although the best stereotype is the most likely type, that type is not always selected by representativeness. Consider first the case of  $\alpha > 0$ , illustrated in Figure 1. For simplicity, we plot the case where types are ordered (along the x-axis) and sufficiently fine to represent  $(\pi_{x,S})_{x=1,\dots,N}$  by a continuous frequency distribution (on the y-axis). When  $\alpha > 0$ , the likelihood ranking of types is the same for  $S$  (solid line) and  $-S$  (dashed line). Formally, the distributions are co-monotonic and both groups share the same modal type.

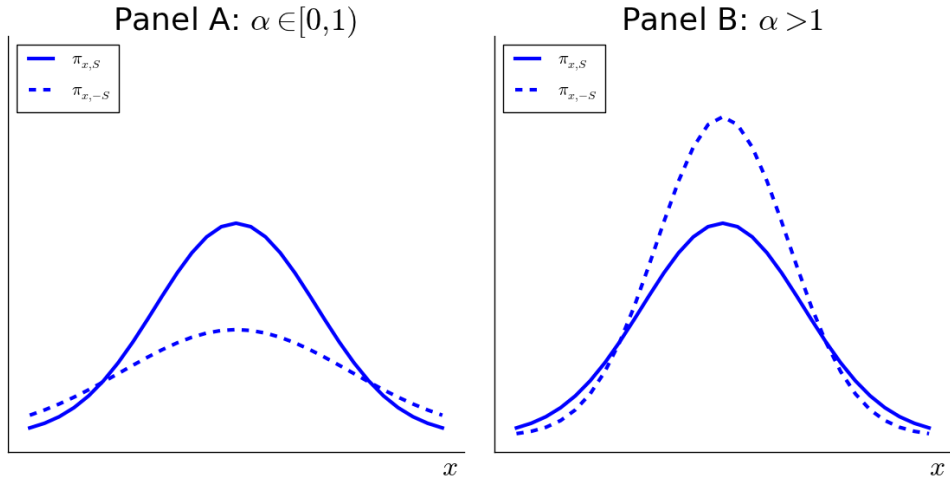


Figure 1: Likely and unlikely stereotypes.

Panel A describes case i) when  $0 < \alpha < 1$ . The fact that  $\alpha < 1$  implies that the distribution in group  $-S$  has heavier tails than  $S$ , which is in turn more concentrated around its mode. In this case, the most ex post informative and thus representative type  $x_1$  for  $S$  is its modal type. As a consequence, the exemplar for  $S$  is its most likely type. (In contrast, the exemplar for  $-S$  is its least likely type).

Panel B represents case ii), in which  $\alpha > 1$ . Now, the fact that  $\alpha > 1$  implies that the distribution in group  $-S$  has thinner tails than that in  $S$ . In this case, the ex post informative and thus representative type  $x_1$  for  $S$  lies in the tails. Thus, the exemplar for  $S$  is its least likely type. (In contrast, the exemplar for  $-S$  is its most likely type).

Proposition 1 captures one sense in which comparing similar groups, namely those with

the same modal type, leads to bad stereotyping for one of them. This is because, from Definition 1, the representativeness ranking of types for group  $S$  is the opposite of that for group  $-S$ . As a consequence, if the distributions have the same likelihood ranking, as in the case  $\alpha > 0$  in Proposition 1, then either  $S$  or  $-S$  will inevitably be represented by an unlikely stereotype. In other words, when distributions are co-monotonic, the process of stereotyping entails a “competition” in which one group obtains a good stereotype while the other obtains a bad stereotype.

To see this, consider the following example. Suppose that  $X$  measures work habits. For the sake of simplicity, we consider two types, namely  $X = \{\text{time spent on work, time spent on vacation}\}$ . Group  $S$  (the Americans) work 49 weeks per year, so the conditional distribution of work versus vacation time is  $\{0.94, 0.06\}$ . In contrast, group  $-S$  (the Europeans) work 47 weeks per year, with work habits  $\{0.9, 0.1\}$ . In both cases, work is by far the most likely activity. However, because the Americans’ work habits are more concentrated around their modal activity, the stereotypical American activity is work. Because Europeans have fatter vacation tails, their stereotypical activity is enjoying the dolce vita. This stereotype is inaccurate, precisely because the vast majority of time spent by Europeans is at work. Still, due to its higher representativeness, vacationing is the distinctive mark of Europeans, which renders the image of holidays highly available when thinking of that group.

Consider now the case  $\alpha < 0$ , where the distributions for  $S$  and  $-S$  are “dissimilar”, in that they have opposite likelihood (and representativeness) rankings over types. Then we have the following result.

**Corollary 1** *Consider again the case of Proposition 1 where  $\pi_{x,-S} = \pi^* \cdot \pi_{x,S}^\alpha$ , for  $\alpha \in R$ .*

- i) If  $\alpha \leq 0$ , the stereotypes for both  $S$  and  $-S$  are each group’s most likely types;*
- ii) If  $\alpha > 0$ , for either  $S$  or  $-S$  the exemplar is the group’s least likely type.*

In broad terms, both groups are represented by good exemplars only if the two distributions are sufficiently different, and in particular if they have different modal types. Figure 2 illustrates this result.

Panel A shows the case in which  $\alpha < 0$ . Here the two groups are very different, in the sense that the most likely type for  $S$  is the least likely for  $-S$  and vice versa. In this case,

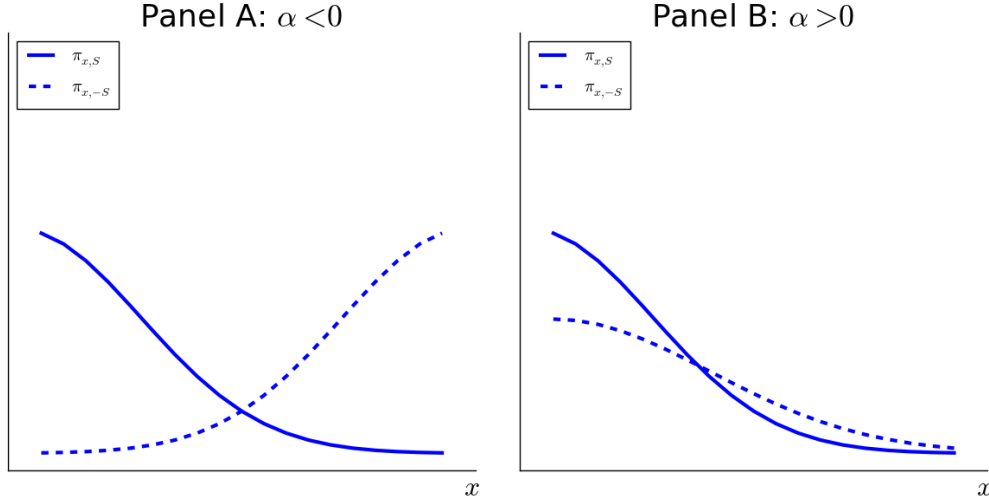


Figure 2: Overall quality of stereotypes.

the exemplar for both groups is good and coincides with the modal type. Panel B shows the case where  $\alpha > 0$ . Now the most likely types in  $S$  are the most likely types also in  $-S$ . In this case, only one group ( $S$  in Figure 2) is represented with a likely type, while the other obtains an unlikely, and thus bad, exemplar.

Corollary 1 emphasises the point that the overall quality of stereotypes is worse when the groups being compared are fairly similar. Intuitively, the DM focuses on differences among groups and neglects features that are likely to arise in both groups. This causes large inaccuracies when groups are similar. Accordingly, the representation of a group  $S$  can switch from inaccurate to accurate when the comparison group  $-S$  changes appropriately, as depicted in Figure 2.

To see this, consider the following example. Suppose  $X$  measures occupations by sector,  $X = \{\text{agriculture, industry, market services, non-market services}\}$ . The distribution of group  $S$ , say the French, across them is  $\{.03, .22, .38, .37\}$ , while that of  $-S$ , the Germans, is  $\{.015, .28, .40, .305\}$ .<sup>6</sup> The modal French occupation, like the modal German occupation, is market services, such as trade, transportation, finance and so on. However, the stereotypical French sector is its very small agriculture, while the stereotypical German sector is its larger,

<sup>6</sup>Source: [http://epp.eurostat.ec.europa.eu/statistics\\_explained/index.php/Labour\\_market\\_and\\_labour\\_force\\_statistics](http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Labour_market_and_labour_force_statistics).

but still relatively infrequent, industry. The reason is that these two economies are very similar in their overall distribution of labor force, and therefore the most representative differences arise in tail sectors.

Compare now the French economy to the Chinese economy, which is characterized by the distribution  $\{.36, .28, .15, .21\}$ .<sup>7</sup> In this context, the stereotypical French sector is market services while the stereotypical Chinese sector is agriculture – the exemplars coincide with the modal types! This is not only because the exemplars are likely in the respective groups, but also because they are relatively unlikely in the comparison group. Because probability mass is concentrated in different types in the different groups – service sector for the French, and farming for the Chinese – the most distinguishing features of each group coincide with the most likely ones. Because social groups tend to be similar in many dimensions, this logic implies that social or ethnic stereotypes tend to be particularly inaccurate. Such stereotypes focus either on relatively minor cultural differences or on tail extremist members. This mechanism could be behind countless negative stereotypes held about ethnic groups.<sup>8</sup>

When the random variable  $X$  is ordered, for instance when it measures years of education, height, or earnings, types can be indexed according to the natural ordering of  $X$ , in which case type  $x = 1$  corresponds to the lowest value of  $X$  while  $x = T$  corresponds to the highest value of  $X$ . We then have:

**Corollary 2** *If the distribution of  $x$  across groups  $S$  and  $-S$  satisfies the monotone likelihood ratio property (MLRP), then:*

*If  $\pi_{x,S}/\pi_{x,-S}$  increases in  $x$ , the exemplar for  $S$  is the highest type  $x = T$ .*

*If  $\pi_{x,S}/\pi_{x,-S}$  decreases in  $x$ , the exemplar for  $S$  is the lowest type  $x = 1$ .*

Corollary 2 is a useful result because many economic models satisfy the monotone likelihood ratio property (e.g., many agency models). When MLRP holds, the exemplar for group  $S$  is located either on the left or the right tail of the support of  $X$ . That is, the exemplar is an extreme observation. Intuitively, under MLRP extreme observations are most informative about – and thus representative of – the group they come from. For example, let  $x$

---

<sup>7</sup>Adapted from: <https://www.cia.gov/library/publications/the-world-factbook/geos/ch.html>.

<sup>8</sup>The anthropology and psychology literatures documents that extremely negative ethnic stereotypes play an important role in inter-group conflict (Tajfel, 1982).

measure firms' earnings growth bundled in three broad types,  $X = \{\text{low, moderate, high}\}$ . Consider the set  $S$  of firms in a growth sector such as the biotech industry, with conditional distribution  $\{.1, .5, .4\}$ , and the set  $-S$  of firms in a stable sector such as utilities, with conditional distribution  $\{.2, .5, .3\}$ . The distribution of earnings growth across the two sectors satisfies MLRP, with  $\pi_{x,S}/\pi_{x,-S}$  increasing in  $x$ . As a consequence, the exemplars of  $S$  and  $-S$  are extreme: the exemplar biotech firm has high earnings growth, while the exemplar utility firm has low growth. Note that the exemplars for both sectors are bad, in the sense that the modal earnings growth in both sectors is moderate. In assessing the prospects of a newly listed firm belonging to a "hot" sector that has recently outperformed the aggregate, an investor spontaneously recalls representative extremely successful firms belonging to that sector, even though extreme success is very unlikely.<sup>9</sup>

As this example illustrates, when the distribution of  $x$  has thin tails the exemplar is not just an extreme but also a relatively unlikely type. When we analyze continuous distributions in Section 5, we will see that this is indeed the case for normal distributions.

### 3.2 Stereotypical Beliefs

Suppose now that DMs recall more than one type, namely  $d > 1$ . In this case, under the assumptions of Proposition 1, the formation of stereotypes follows a very simple rule.

**Corollary 3** *If the distribution in  $-S$  is the power transformation of the distribution in  $S$  as defined in Proposition 1, then:*

*i) If  $\alpha < 1$ , the stereotype for  $S$  consists of its  $d$  most likely types.*

*ii) If  $\alpha > 1$ , the stereotype for  $S$  consists of its  $d$  least likely types.*

*In both cases, the beliefs of the DM are represented by distribution  $a$  that truncates away the  $N - d$  types that do not constitute the stereotype.*

As in Proposition 1, when the distribution of  $S$  differs from that of  $-S$  in likely types, case i), then the stereotype for  $S$  is good. The DM makes an imperfect assessment of  $S$  and

---

<sup>9</sup>By emphasizing stereotypical outcomes in the valuation of firms' stock, this logic provides a new mechanism for the growth-value puzzle in asset pricing (Lakonishok, Shleifer and Vishny 1994). Because stereotypical outcomes are also extreme, this mechanism is very similar to that described using the model of salience (Bordalo, Gennaioli and Shleifer 2013). In general, however, stereotypes and salience can produce different results, see Section 3.3.



in particular fails to fully account for its variability across types. However, because the most likely types are recalled, beliefs are minimally distorted given the constraints on memory. When instead the distribution in  $S$  is more dispersed in unlikely types than that of  $-S$ , case ii), the stereotype for  $S$  is bad. Here the DM fails in two ways. First, as before, the DM suppresses variability across types. Second, the DM now disproportionately recalls unlikely types. For instance, the DM typically perceives social groups to be more homogeneous than they really are, but his distorted beliefs may be compounded by the fact that he generalizes to the entire group a trait that may be very infrequent. In this case, stereotypes indeed “provide the greatest differentiation between groups, and [...] show the least within-group variation” (Hilton and von Hippel 1996).

The characterization of stereotypes is particularly interesting when  $X$  is ordered. Suppose that  $X = \{x_1, \dots, x_N\}$  is ordered, that type  $x_i$  has base rate  $\pi_{x_i}$  in the population  $\Omega = S \cup -S$ , and write  $\mathbb{E}^{st}(x|S) = \sum_1^d \pi_{x_i, S}^{st} \cdot x_i$ .

**Corollary 4** *If MLRP holds, then:*

*i) if the likelihood ratio is increasing, the stereotype for  $S$  is the right tail  $\{x_{N-d+1}, \dots, x_N\}$ .*

*Moreover,*

$$\mathbb{E}^{st}(x|S) > \mathbb{E}(x|S) > \mathbb{E}(x)$$

*ii) if the likelihood ratio is decreasing, the stereotype for  $S$  is the left tail  $\{x_1, \dots, x_d\}$ .*

*Moreover,*

$$\mathbb{E}^{st}(x|S) < \mathbb{E}(x|S) < \mathbb{E}(x)$$

*If MLRP does not hold, and the likelihood ratio is U-shaped and symmetric, the stereotype for  $S$  consists of the  $d$  most extreme types on the left and right tails.*

When MLRP holds, the DM’s belief about  $S$  are formed by truncating from the original distribution the least representative tail. This leads to three important effects. First, the DM’s mean assessment of group  $S$  is shifted in the same direction as the true conditional mean  $\mathbb{E}(x|S)$  relative to the unconditional mean  $\mathbb{E}(x)$ . Second, because his assessments are biased in the direction of the (extreme) exemplar, the DM’s estimate of the mean type is too extreme. Third, as types in the non-stereotypical tail are neglected, the DM’s assessment

of the variance in types is dampened relative to its true value. In this case, stereotyping effectively leads to a form of overconfidence in which the DM both holds extreme views and overestimates the precision of his assessment.

Broadly, Corollary 4 implies that the DM overemphasises the correlation implied by MLRP between types and groups. When assessing a hot sector, the investor recalls not just highly successful but also moderately successful exemplars in the same sector. However, he neglects (or underweights) the possibility of failures, because failure is statistically non-diagnostic, and psychologically non-representative, of a growing sector. This causes both excessive optimism (in that the expectation of growth is unreasonably high) and overconfidence (in that the variability in earnings growth considered possible is truncated). True, the hot sector may have better growth opportunities on average, but representativeness exaggerates this feature and induces the investor to neglect significant uncertainty and thus the risk of failure.

Similarly, when assessing an agent's skill level, a principal attributes high performance to high skill, because high performance is the distinctive mark of a talented agent. Because he neglects the possibility that talented agents perform poorly and that non-talented agents perform well (perhaps due to stochasticity in the environment), the principal has too much faith in skill, and neglects the role of luck in accounting for the output.

Consider now the case where the likelihood ratio is U-shaped, and MLRP does not hold (case ii). This case arises when the distribution of  $S$  has heavy tails (both right and left) relative to  $-S$ . In this case, the stereotype for  $S$  includes both the right and the left tails while it truncates away the less extreme, intermediate types. The extreme types capture the distinctive trait of the represented group, namely fatter tails. In this case, stereotyping leads the DM to exaggerate the variance of potential outcomes, effectively displaying a form of under-confidence. When assessing the performance of an erratic student, a teacher may perceive extreme variability by recalling examples of the spectacular performance as well as of failures. These extreme outcomes may be very infrequent, but they are recalled, and over-weighted, because they constitute the distinguishing and thus representative feature of the student in question. The teacher may be under confident in predicting the performance of the student precisely because he recalls such high variability.

### 3.3 Discussion

Our model captures the notion that stereotypes are mental representations of situations or groups which focus on “features that are the most distinctive” among those situations or groups (Hilton and Hippel 1996). In our model, what makes a type, or a feature, distinctive and stereotypical for a group is its representativeness. As discussed in Section 2, our definition of representativeness is related to Kahneman and Tversky’s formulation (1972) and is equivalent to the notion of representativeness presented in GS (2010). As in that paper, representativeness leads the DM to represent each group with stereotypes that emphasize ex post informative types for that group. While GS (2010) focused on the inference problem, namely how stereotypes affect probabilistic assessment of competing hypothesis (answering the question: which group  $S$  does this observation  $x$  belong to?), here we explore the more basic questions of how stereotypes are formed, what determines their accuracy, and how they respond to new information.

Three aspects of our model highlight the importance of understanding stereotype formation and evolution. First, stereotypes exaggerate the distinctive trait of the group they represent. The main implication here is that stereotypes might correspond to types or features that are ex-post informative about the group but ex-ante unlikely to occur. If most rug dealers are Persian, it would be correct to conclude that in all likelihood a given rug dealer is Persian. The problem, of course, is to conclude ex-ante that a Persian is likely to be a rug dealer, since the vast majority of them are not. But stereotyping does exactly this. When a group exhibits a higher than average frequency of tail types, these tails become over-weighted in the stereotype. When instead a group is concentrated around its most likely type, that type becomes stereotypical as long as other groups are relatively more concentrated in other types.

The second key aspect of stereotyping is its context dependence. A type is diagnostic for group  $S$  if it ex post informative about  $S$  relative to a reference group  $-S$ , which provides the measure of context. Proposition 1 and Corollary 1 show how the stereotypes for  $S$  change depending on normatively irrelevant changes in the distribution of  $-S$ . Context dependence allows us to connect two important properties of stereotypes: i) stereotypes are often deeply

held and perhaps culturally transmitted beliefs, and ii) stereotypes can also be generated on the fly, in a context specific manner. For instance, when are asked to provide a stereotype of a Ivy League football player, one might think of the tall, popular quarterback. While this caricature is rooted in background knowledge of college life, in our view it follows from an implicit – though stable – comparison to other groups. But stereotypes are also generated by explicit and momentary comparisons, in which case they pop up automatically and on the fly. When we assess a prospective student who is a football player, we are likely to imagine him as bigger – though less dedicated to academics – than other students, but as small when compared to an average football player. While the cultural stereotype typically evokes a specific exemplar, the explicit comparison done for instrumental purposes is likely to evoke a more nuanced stereotypical distribution. We suggest that in both cases, the same mechanism of representativeness shapes beliefs about a group, in a context dependent way. As we show later, context dependence has key implications for the dynamic updating of stereotypes, explaining why – as the distribution of types in a group changes – the stereotype may either persist (if the comparison group changes in the same direction) or be radically updated (if the comparison group changes in an asymmetric way).

Lastly, a central property of stereotypes is that, while the most representative traits of the group are recalled, other less diagnostic types simply do not come to mind. Representativeness can therefore shed light on whether extreme types are recalled, and if so which ones. In our theory, extreme types do not come to mind when assessing a group unless they are more likely to occur in that group relative to the comparison group. When going out for a walk, we do not typically think of the risk of being hit by a car unless traffic is very intense. At a restaurant, we typically do not think of the food being dangerous unless utensils are particularly dirty. In contrast, when we judge the prospects of a company whose earnings are growing faster than average, extreme right tail realizations are stereotypical, while left tail realizations are not. We focus on the former, while the latter do not come to mind and are ignored.

Understanding the recall of extreme events connects with the phenomenon of payoff salience, which we first formalized in BGS (2012). In that model, a state is salient when it has extreme payoffs. Here a state is representative and thus stereotypical when it is ex-post

most informative of the group the decision maker is assessing. While salience describes how the DM allocates attention between pieces of information he has in mind, representativeness accounts for *which* information he brings to mind in the first place. Our analysis suggests an important interaction between salient and representative events. When  $X$  has a natural order (e.g. years of education, or income), extreme states that are representative are also salient. In this case, salience reinforces representativeness in inducing biased decisions: for example, a decision maker takes on excess risk when he focuses (both in the stereotype and in the salience sense) on high tail payoffs. On the other hand, some extreme states may not be representative and thus do not come to mind, even though they would be salient if recalled. This is the case of Corollary 2 when MLRP holds, or the example of the restaurant’s dirty kitchen. This representativeness based neglect may sometimes lead to better decisions since, given the payoff salience of these unlikely risks, life could be problematic if we always attended to them. For instance, dining at a restaurant would make us overly anxious.

Although we believe that representativeness captures a significant part of the stereotype phenomenon, several significant aspects lie outside our approach. Foremost among them, assessing a group (or a member of a group) can elicit automatic associations that, while very available, are not necessarily the most diagnostic.<sup>10</sup> Related, not every diagnostic type gets selected for active recall. When thinking of Muslims, DMs may be more likely to recall a religious fanatic than a Bedouin. Both are representative types in our model, but the former is currently widely available in public discourse. Representativeness is aided by accessibility in memory of diagnostic types, but accessibility itself can bring non diagnostic types to the top of the mind.

A robust finding in social psychology is that stereotypes minimise within-group variability, and that stereotypes of in-groups tend to be more detailed and more positive than those of out-groups (Hilton and von Hippel 1996). A strong prediction of our model is that, by truncating the true distribution of types, the DM perceives both in- and out-groups to be

---

<sup>10</sup>For example, in the example France vs China, the occupations of Chinese workers that comes to the minds of American consumers might not only be the farmer (the most likely and most representative type) but also the industrial worker (because American consumers are continuously faced with Chinese industrial exports). This issue is related to Kahneman and Tversky’s availability heuristic, and the question of when it dominates representativeness in tasks involving probability assessments. This is an important topic for further study.

more homogeneous than they really are. Although our model does not explicitly predict systematic differences between in-group and out-group stereotypes, these differences might emerge if the DM is assumed to have less (and more biased) information about the out-group. This asymmetry might also be attributed to the greater availability of positive types in the in-group.

A final point to highlight is that stereotypes may themselves shape which new information is obtained and how it is processed. Psychologists have documented a tendency to search for information that confirms one’s beliefs (Nickerson 1988). Schwartzstein (2012) proposes a model of biased learning in which information is used to update beliefs only about dimensions that are attended to. This can be viewed as a model of what features and dimensions characterize the groups under assessment (in other words, what the random variable  $X$  is). This approach is complementary to our model. The representativeness mechanism that links priors to stereotypes can naturally be coupled with a non-Bayesian updating process.

### 3.4 Stereotypes and Biases in Judgment

Stereotypes relate to several strands of the experimental literature on biases in judgment. As shown by GS (2010), the neglect of non-representative types can account for several biases identified by Tversky and Kahneman (1974) and associated with the representativeness heuristic. Because representativeness of a type for a group does not depend on its base rate in the overall population, representativeness-based recall can lead to a neglect of base rates in judgment: a DM will underestimate the likelihood of observing a type in a group if that type is not representative of the group, regardless of its base rate. Consider for instance a DM asked to assess the modal race of the population in the US living in poverty. Because a much higher proportion of blacks are poor than whites (27.4% vs 9.9% as of 2010), the DM stereotypes blacks as poor and whites as not poor. By neglecting the poor white type, he thus estimates poor blacks to outnumber poor whites. In fact, because the white population is over five times larger than the black population, white poor outnumber black poor by 2 to 1.

Our model provides a psychological foundation for base rate neglect and yields predictions different from other approaches seeking to capture the same phenomenon. To see this,

consider the classic example in which a medical test for a particular disease has a 99% rate of true positives and a 1% rate of false positives. The test is informative in that the relative likelihood of having the disease is higher if the test is positive than if it is negative. Such a test generates extreme stereotypes: the stereotypical person who tests positive is a person with the disease, the stereotypical person who tests negative does not have the disease. Accordingly, the DM greatly boosts his assessment that a positively tested person is sick (as in the standard example), but also that a negatively tested person is healthy. Similarly, a faculty member who is judging a job market candidate from the quality of his talk rushes to the conclusion that the candidate is highly qualified if the talk is good, and that the candidate is incompetent if the talk is bad. In both examples, the over-reaction is stronger the less likely the stereotypes are: if most people are healthy, or if most candidates are qualified, the decision maker’s assessment is fairly accurate when the individual being judged falls in the group with a likely stereotype (negative test, good talk) but is severely biased otherwise (positive test, bad talk).

This example is a direct application of Corollary 4: when the DM finds out that a person belongs to group  $S$ , he updates his assessment in the right direction. However, because the group has an extreme stereotype, he updates too much. This effect is starkly different from the mechanical neglect of base rates in Bayes rule. In that alternative model, the DM can update his assessment in the wrong direction: because he neglects the base rates when assessing the conditional – but not the unconditional – likelihood of sickness, the DM can be *less* confident that a person is healthy after a negative test than without any information, provided the probability of being sick is sufficiently low.<sup>11</sup>

The neglect of non-representative types is also related to the conjunction fallacy, as shown in GS (2010). Consider a DM assessing the likelihood of several different occupations for

---

<sup>11</sup>One way to model base rate neglect is to postulate a modified Bayes rule (Grether 1980, Bodoh-Creed, Benjamin and Rabin 2013):

$$\pi_{x,S} = \frac{\pi_{S,x} \cdot \pi_x^\eta}{\pi_{S,x} \cdot \pi_x^\eta + \pi_{S,-x} \cdot \pi_{-x}^\eta}$$

where  $\pi_{S,x} = Pr(S|X = x)$ ,  $\pi_x$  is  $x$ ’s base rate,  $-x$  is the complement of type  $x$  in  $X$  and parameter  $\eta \in [0, 1]$  modulates the strength of base rate neglect. When  $\eta = 1$ , the DM follows Bayes’ rule. When  $\eta < 1$ , the DM dampens the base rates of  $x$  and  $-x$ . The key implication is that the conditional probability of type  $x$  in any group  $S$  is boosted if and only if its base  $\pi(x)$  in the entire population is small, namely if  $\pi_x < \pi_{-x}$ . The frequency of a low probability type is overestimated in any target group  $S$ , independent of its frequency in  $S$  relative to that in the entire population.

a woman who was a social activist in college (the Linda problem, Tversky and Kahneman 1972). When computing the probability that the woman is a bank teller, the DM compares the distribution of features in the group of bank tellers to that of non-bank tellers. The exemplar of a bank teller is highly unlikely given that the woman is a former activist. However, should the non-stereotypical “feminist bank teller” type be brought to the DM’s attention, he would assign it a positive probability, above that of the “bank teller exemplar”, in violation of the conjunction rule of probability.

### 3.5 Multidimensional Types

In the real world, types are often multidimensional. Members of social groups vary in their occupation, education, religion, income and other dimensions. Firms differ in their sector, location and management style. The state of the economy includes GDP growth, interest rates, and inflation. While multiple dimensions are subsumed in our previous analysis, in which each of the  $N$  types may consist of a unique specification of a possibly large set of attributes, for many groups stereotypes are formed along specific dimensions. Thus, some social groups are stereotyped by their occupations (“immigrants work in menial jobs”), others by their political views (“the young are liberal”), still others by their religious customs (“Buddhists meditate”).<sup>12</sup> How are these dimensions selected?

Our model of representativeness provides a parsimonious perspective on this issue: the stereotype for group  $S$  will be organized around the dimension along which  $S$  is most different from  $-S$ . To see what this means, consider an example in which social groups in the US are described in terms of educational attainment (share of group members with higher education degree) and demand for social services (share of group members on welfare). Suppose that 35% of the white population has a college degree and 2% are on welfare, while 21% of the black population has a college degree and 10% are on welfare.<sup>13</sup> In terms of representativeness, the black population differs the most from the white population along

---

<sup>12</sup>As alluded to in Section 3.3, stereotypes may vary depending on circumstances according to changes in the comparison group. Walking in a deserted neighborhood may evoke a crime-based stereotype, while watching a sport event may evoke an athleticism-based stereotype for the same ethnic group.

<sup>13</sup>Data from the National Center for Education Statistics ([http://nces.ed.gov/programs/digest/d12/tables/dt12\\_008.asp](http://nces.ed.gov/programs/digest/d12/tables/dt12_008.asp)) and from Statistic Brain (<http://www.statisticbrain.com/welfare-statistics/>).



the welfare dimension, not along the educational attainment dimension. This follows from the diminishing sensitivity of representativeness (Remark 1): even though the difference in educational attainment is larger (79% of blacks versus 65% of whites without college degrees), the most distinguishing feature of the black population is its higher relative demand for welfare (10% versus 2%). In this sense, to be formalized below, the stereotype for the black population is to be on welfare, despite the fact that only a small minority is on welfare (and even if the higher share of blacks on welfare were partially driven by their lower rates of college graduation). Conversely, the stereotype for whites is their higher share of graduates, not the fact that fewer are on welfare, even though a minority of whites go to college and a majority of whites are not on welfare. This is both because relatively more whites go to college and because most blacks are also not on welfare. The example shows that when groups are characterized by multidimensional types, they can be stereotyped along different dimensions. In particular, due to diminishing sensitivity, both groups can be stereotyped with unlikely types.

We now formalize the intuition described in this example. Suppose that the original random variable  $X$  is the product two categorical variables  $Y$  and  $Z$ , where  $Y \in \{1, \dots, N_Y\}$  and  $Z \in \{1, \dots, N_Z\}$ , where  $N_Y, N_Z > 1$ . In the previous notation,  $N = N_Y \times N_Z$  is the number of types. Types are indexed by realizations  $(y, z)$  of the two variables. According to Definition 1, the representativeness of type  $(y, z)$  for  $S$  is then defined by  $\Pr(y, z|S)/\Pr(y, z|-S)$ . In this setup, a stereotype consists of the  $d$  most representative realizations  $(y, z)$  of the two variables. To make progress, consider the special case in which the representativeness of a realization of  $Z$  does not depend on that of  $Y$ , formally  $\Pr(z|y, S)/\Pr(z|y, -S) = \Pr(z|S)/\Pr(z|-S)$  for all  $z$  and all  $y$  (an assumption implicit in the previous example). The representativeness of type  $(y, z)$  is then an increasing function of:

$$\frac{\Pr(z|S)}{\Pr(z|-S)} \cdot \frac{\Pr(y|S)}{\Pr(y|-S)}. \quad (3)$$

The representativeness of  $(y, z)$  is simply the product of the representativeness of  $y$  and  $z$  considered independently. This condition holds, for instance, when being uneducated (low  $y$ ) is predictive of lower income (low  $z$ ), but this correlation may be independent of group

identity and so acts uniformly across groups. Under this assumption, if one group has a higher share of poor members, that must be because it has also a higher share of uneducated members.

When equation (3) holds, the organization of a stereotype is pinned down by comparing the variation in representativeness along the two dimensions  $Y$  and  $Z$ . Denote by  $y_r$  the  $r$ -th most representative type of  $Y$ , when representativeness for type  $y$  is defined by  $\Pr(y|S)/\Pr(y|-S)$ . If  $y_1$  is much more representative than  $y_2$ , then type  $(y_1, z)$  is more representative than  $(y_2, z')$  for any  $z$  and  $z'$ . In this case, the stereotype intuitively becomes “lexicographic,” in the sense that it allows for little variation in types of the highly representative dimension  $Y$  and for much more variation in types of the less representative dimension  $Z$ . Specifically, the first  $N_Z$  types that come to mind are combinations of  $y_1$  with all possible realizations of  $Z$ . The result below characterizes the cases in which this lexicographic ranking arises.

**Proposition 2** *When (3) holds, the stereotype is lexicographic in dimension  $Y$  if:*

$$\min_r \left[ \frac{\Pr(y_r|S)}{\Pr(y_r|-S)} / \frac{\Pr(y_{r+1}|S)}{\Pr(y_{r+1}|-S)} \right] > \left[ \frac{\Pr(z_1|S)}{\Pr(z_1|-S)} / \frac{\Pr(z_{N_Z}|S)}{\Pr(z_{N_Z}|-S)} \right], \quad (4)$$

where  $v_r$  denotes the  $r$ -th most representative realization  $v = y, z$ , when representativeness for  $v$  is defined in isolation, formally  $\Pr(v_r|S)/\Pr(v_r|-S)$ . In particular, the stereotype is lexicographic in  $Y$  if  $Z$  is uninformative,  $\Pr(z|S) = \Pr(z|-S)$  for all  $z$ .

Equation (4) identifies a stark condition for the stereotype to be lexicographic, namely that the maximum percentage variation in the likelihood ratio along  $Z$  is lower than the minimum variation along  $Y$ . Not only the ranking of  $Y$  types by representativeness matters, but also how large an increase in representativeness is obtained by recalling  $y_1$  rather than  $y_2$ , and so on. In particular, the stereotype is lexicographic in  $Y$  when the non-diagnostic dimension  $Z$  is undistinguishable across groups. When comparing Americans and Europeans, stereotypes do not focus on particular age groups, in the sense that the stereotypical European or American can be of a wide range of ages.

More importantly, however, Proposition 2 says that stereotypes can be organized along a given dimension  $Y$  if each type along  $Y$  is sufficiently more representative than the next.

Remark 1 implies that representativeness of types becomes more extreme when the most representative types are unlikely. This suggests, as in the previous example on the demand for welfare, that Equation (4) tends to select bad stereotypes.

## 4 Stereotypes and Reaction to New Information

So far, we considered the formation and properties of stereotypes based on static distributions. The model, however, lends itself naturally to exploring how stereotypes and beliefs change by the arrival of new information over time. To this end, we suppose that, unlike in Section 3, the decision maker does not have perfect information about the categorical distribution  $(\pi_{x,S})_{x=1,\dots,N}$  of the group  $S$  of interest, or about the distribution  $(\pi_{x,-S})_{x=1,\dots,N}$  of the comparison group  $-S$ . Instead, at the outset the DM has priors over these distributions. To highlight the connection to our previous analysis of static categorical distributions, we assume these are described by conjugate Dirichlet priors:

$$g[\pi_{x,W}, \alpha_{x,W}]_{x=1,\dots,N} = \frac{\Gamma(\sum_x \alpha_{x,W})}{\prod_x \Gamma(\alpha_{x,W})} \cdot \prod_x \pi_{x,W}^{\alpha_{x,W}-1}, \quad \text{for } W = S, -S$$

Given the parameters  $\alpha_S = (\alpha_{x,S})_{x=1,\dots,N}$  and  $\alpha_{-S} = (\alpha_{x,-S})_{x=1,\dots,N}$  of the Dirichlet distribution, the prior expectation about the probability of type  $x$  in a given group  $W$  is given by the multinomial distribution:

$$Pr(X = x | \alpha_W) = \mathbb{E}(\pi_{x,W} | \alpha_W) = \frac{\alpha_{x,W}}{\sum_u \alpha_{u,W}}, \quad \text{for } W = S, -S. \quad (5)$$

In particular, the ex-ante probability assigned by a fully Bayesian decision maker to the event that the next observation from group  $W$  is in type  $x$  is the share of observations from  $W$  in type  $x$ , namely  $\Pr(X = x | \alpha_W)$ . Thus, it is natural to assume that the stereotype initially held by the DM is determined as a function of the probabilities in Equation (4) according to Definition 1. For simplicity, we set  $\sum_x \alpha_{x,S} = \sum_x \alpha_{x,-S}$ .

Suppose that a sample  $n_W = (n_{1,W}, \dots, n_{N,W})$  is observed, where  $n_{x,W}$  denotes the observation count in type  $x$  and let  $\sum_x n_{x,W}$  be the total number of observations for group

$W$ . Then, the posterior probability of observing  $x$  assessed by a Bayesian DM is

$$\Pr(X = x|\alpha_W, n_W) = \mathbb{E}(\pi_{x,W}|\alpha_W, n_W) = \frac{\alpha_{x,W} + n_{x,W}}{\sum_u (\alpha_{u,W} + n_{u,W})}, \quad (6)$$

which is a weighted average of the prior probability of Equation (4) and the sample proportion  $n_{x,W}/n_W$  of type  $x$ . As new observations arrive, the probability distribution in group  $W$ , and thus stereotypes, are updated according to Equation (6).

Given the Bayesian update rule implicit in equations (4) and (6), we now analyze how information arrival affects the DM's beliefs. Proposition 3 considers the effect of new information about the types that come to mind, allowing us to characterize when and how stereotypes change. By stereotype change we mean a change in the set of types included in the stereotype (e.g., whether the original exemplar is discarded). Proposition 4 considers the effect of information on probability assessments.

**Proposition 3** *Suppose that the DM observes the same number of realizations from both groups, formally  $\sum_u n_{u,S} = \sum_u n_{u,-S} = n$ . Then:*

*i) If for both groups all observations occur on the same type  $x$  that is initially non-representative for  $S$ , then this type does not become representative for  $S$ . Formally, if  $n_{x,S} = n_{x,-S} = n$  for a type  $x$  such that  $\alpha_{x,S}/\alpha_{x,-S} < 1$ , then  $\Pr(X = x|\alpha_W, n_S)/\Pr(X = x|\alpha_W, n_{-S}) < 1$  for all  $n$ .*

*ii) If all observations for  $S$  occur in a non representative type for  $S$ , while those for  $-S$  occur in a type that is representative for  $S$ , then for a sufficiently large number of observations the stereotype for  $S$  changes. Formally, if  $n_{x,S} = n$  for a type  $x$  such that  $\alpha_{x,S}/\alpha_{x,-S} < 1$ , while  $n_{x',-S} = n$  for a type  $x'$  such that  $\alpha_{x',S}/\alpha_{x',-S} > 1$ , then for  $n$  sufficiently large  $\Pr(X = x'|\alpha_W, n_S)/\Pr(X = x'|\alpha_W, n_{-S}) < 1 < \Pr(X = x|\alpha_W, n_S)/\Pr(X = x|\alpha_W, n_{-S})$ .*

Intuitively, when the DM observes a sample from both groups, his stereotype for  $S$  changes only if the new observations are sufficiently contrary to the initial stereotype. In fact, only such contrarian data can reduce the likelihood ratio of the initial stereotype while boosting the likelihood ratio of previously neglected types.

To see this, consider first case i), in which data uniformly accrue in the two groups  $S$  and  $-S$ . In particular, the  $n$  observations occur – for both groups – in a type that is non-

representative for  $S$ . In this case, the non-representative type never becomes representative for  $S$  despite the fact that the data consistently point to its relevance. Reductions in the overall incidence of crime do not debunk a negative stereotype about a given group if a majority of criminals still come from that same group. A process of economic development that improves the livelihoods of all groups in a population does not improve the stereotype of a group that continues to include a disproportionately high share of underdogs. The intuition for this result comes from diminishing sensitivity of the likelihood ratio (Remark 1): types that are highly likely to occur in both groups are *ceteris paribus* less representative.

Although stereotypes do not change when information is symmetric across groups, they can change quickly when information is asymmetric. In case ii), the  $n$  observations for  $S$  occur in a non-representative type  $x$  for  $S$ , while the  $n$  observations for  $-S$  occur in a representative type  $x'$  for  $S$ . In this case, for  $n$  sufficiently large,  $x$  becomes representative for  $S$  while  $x'$  becomes unrepresentative for  $S$ . One intuitive process captured by this case is the asymmetric reduction in the incidence of tail (but highly representative) events in a group. Reducing crime in certain high-incidence neighborhoods (i.e., ghettos), but not overall, decreases the association between the population of those neighborhoods and crime, debunking the group's crime-based stereotype. The rapid rise of a new commercial class out of an underdog group creates a new stereotype for that group. Some periods of above market performance turns an uninteresting company into a growth stock. The arrival of new information, while beneficial for a rational agent, may not change stereotypes for the better: in the case of the listed company, its recent above average performance may be due to noise. But the investor leaves little room for noise. He looks for causal patterns and quickly jumps to conclusions, even if the informativeness of stereotype-changing information is low. After all, he thinks, above average performance is the distinctive mark of great companies.

We now consider how the initial stereotype for group  $S$  (formally, the priors over  $S$  and  $-S$ ) affects the way in which the DM processes new information about  $S$ . We only consider information concerning  $S$ : since the set of types included in the stereotype is assumed to be constant, information about  $-S$  is irrelevant.

**Proposition 4** *Let  $d > 1$ . Suppose that one observation about type  $x$  is received in group  $S$  (formally,  $n = n_{x,S} = 1$ ). Then:*

*i) If  $x$  belongs to the stereotype of  $S$  and its probability is sufficiently low, the DM over-reacts (relative to the Bayesian) in revising upward his assessment of  $x$ 's probability. Formally, there is a threshold  $\nu \in (0, 1/2)$  such that the DM's assessment of  $x$  over-reacts if and only if  $\alpha_{x,S} / \sum_u a_{u,S} < \nu$ .*

*ii) If  $x$  does not belong to the stereotype of  $S$ , the DM does not update its probability at all (so he under-reacts relative to the Bayesian DM).*

Proposition 4 indicates that stereotypes can both over and under-react to information. In case i), the DM strongly over-reacts to information confirming the stereotype. Intuitively, because the DM neglects non-representative types, he does not fully account the current observation may be due to sampling variability. As a consequence, his beliefs overreact when a type he does attend to is confirmed by the data. If criminal activity is part of a group's stereotype, the DM over-reacts to seeing a criminal from that group and his judgments become even more biased against the group. If a growth company generates surprisingly positive earnings, investors drastically upgrade their belief that the stock is a good investment, because they neglect that an extreme observation can be due to noise.

At the same time, case ii) shows that the DM under-reacts (relative to a Bayesian) to information inconsistent with the stereotype. This is because insofar as the stereotype is unaffected, the probability of a non-stereotypical type is not upgraded, as the type remains neglected in the assessment of the group. Upon observing a highly successful member of a group stereotyped as the underdog, DMs code the occurrence as an "anomaly" and continue to believe that the group at large should be viewed through the lens of the negative stereotype. There are many examples of people who espouse racist views and yet are friendly with individual members of the group they discriminate against. However, as shown in Proposition 4, non-stereotypical information is often ineffective at changing beliefs even if it swamps the few instances underlying the stereotype.

Putting the two cases together, Proposition 4 implies that the DM exhibits a type of confirmation bias (Nickerson, 1998). Faced with two observations of different types from group  $S$  (formally,  $n_{x,S} = n_{x',S} = 1$  and  $n = 2$ ), such that  $x$  belongs to the stereotype of  $S$  but  $x'$  does not, the DM over-reacts to information consistent with the stereotype and ignores information inconsistent with it. Corollary 4 and Proposition 4 show that our

approach provides a unified model of both base rate neglect and confirmation bias: base rate neglect arises when representative types are unlikely, while confirmation bias arises when new information does not change representativeness and allows stereotypes to persist. These biases are the two sides of the same coin of representativeness based recall.

Our model has potentially significant implications for the analysis of financial markets. First, at the broadest level, our analysis suggests that the formation of stereotypes leads to disregard or neglect of certain types or events in the formation of the DM's beliefs. These types or events are in the DM's long term memory, but are not retrieved when he forms his model or stereotype of the world. Gennaioli, Shleifer, and Vishny (2012, 2013), without the benefit of this model of stereotypes, argue that neglect of some risks is in fact central to understanding the behavior of financial markets, particularly around the 2008 financial crisis.<sup>14</sup> GSV present some evidence that investors in mortgage backed securities did not consider the possibility of extremely sharp declines in housing prices among the scenarios they entertained.

Second, case ii) of Proposition 4 yields the further critical implication that, once DM's have formed their stereotypical future, they do not react initially to information inconsistent with the stereotype. One of the great puzzles in thinking about the crisis is why it occurred so late: home prices started declining in late 2006 or 2007, several banks went bust in 2007, and there was a run on asset based commercial paper. Yet the crisis did not occur until a year after some fairly dramatic negative news. The model of stereotypes suggests a possible answer to this puzzle: once the stereotype is formed, and until it changes, investors interpret negative news as anomalous, and do not change their model of the future. It is difficult to obtain this prediction from more standard models.

## 5 Stereotypes for Continuous Distributions

Our model of stereotypes can be extended to cover the case of continuous probability distributions. Let  $X$  be a continuous variable defined on the support  $\bar{X} \subset \mathbb{R}^k$ . Denote by  $x \in \bar{X}$

---

<sup>14</sup>This logic may also provide a foundation for the investor sentiment model of asset pricing (Barberis, Shleifer, Vishny, 1998), a topic we leave to future study.

a realization of  $X$ , which is distributed according to a density function  $f(x) : X \rightarrow \mathbb{R}_+$ . Denote by  $f(x|S)$  and  $f(x|-S)$ , the distributions of  $x$  in  $S$  and  $-S$ , respectively. In line with Definition 1, we define representativeness as:

**Definition 3** *The representativeness of  $x \in \overline{X}$  for group  $S$  is measured by the ratio of the probability of  $S$  and  $-S$  at  $X = x$ , where  $-S = \Omega/S$ . Using Bayes' rule, this implies that representativeness increases in the likelihood ratio  $f(x|S)/f(x|-S)$ .*

In the continuous case, the exemplar for  $S$  is the realization  $x$  that is most informative about  $S$ . Consider the case of one-dimensional variables, namely  $X \subset \mathbb{R}$ . In this case, as in Corollary 2, if the distribution satisfies MLRP, the exemplar for  $S$ , which we denote  $x'_E$ , is  $\sup(X)$  if the likelihood ratio is increasing, or  $\inf(X)$  if the likelihood ratio is decreasing.

The DM constructs the stereotype, which includes additional realizations of  $x$  beyond the exemplar, by recalling the most representative values of  $x$  until the recalled probability mass is equal to the bounded memory parameter  $\delta \in [0, 1]$ . When  $\delta = 0$ , the DM only recalls the exemplar. When  $\delta = 1$  the agent recalls the entire support  $X$  and his beliefs are accurate. When  $\delta$  is between 0 and 1, we are in an intermediate case.

**Definition 4** *Given a group  $S$ , define the set  $X_S(t) = \{x \in X | f(x|S)/f(x|-S) \geq t\}$ . The agent forms his beliefs using a truncated distribution in  $X_S(t(\delta))$  where  $t(\delta)$  solves the equation:*

$$\int_{x \in X(t(\delta))} f(x|S) dx = \delta.$$

The logic is similar to that of Definition 2. The only difference is that now the bounded memory constraint acts on the recalled probability mass and not on the measure of states, which would be problematic to compute when distributions have unbounded support. This feature yields the new implication that changes in the distribution typically change also the support of the stereotype by triggering the agent to recall or forget some states, even if the states' relative representativeness did not change.

## 5.1 The Normal Distribution

Consider the case in which  $f(x|S)$  and  $f(x|-S)$  are univariate normal. In this case, the stereotype of  $S$  is easy to characterize in closed form.



**Proposition 5** *In the normal case, the stereotype works as follows:*

*i) Suppose  $\sigma_S = \sigma_{-S} = \sigma$ . Then, if  $\mu_S > \mu_{-S}$  the stereotype for  $S$  is  $X_S = [x_S, +\infty)$ , where  $x_S$  is implicitly identified by  $\int_{x_S}^{\sup(X)} f(x|S)dx = \delta$ . If instead  $\mu_S < \mu_{-S}$ , the stereotype for  $S$  is  $X_S = (-\infty, x_S]$ , where  $x_S$  is implicitly identified by  $\int_{\inf(X)}^{x_S} f(x|S)dx = \delta$ .*

*ii) Suppose that  $\sigma_S < \sigma_{-S}$ . Then, the stereotype for  $S$  is  $X_S = [\underline{x}_S, \bar{x}_S]$  where  $\underline{x}_S, \bar{x}_S$  are implicitly identified by the system of equations (in  $\underline{x}_S(k), \bar{x}_S(k)$  and  $k$ ):*

$$(x - \mu_{-S})^2 \cdot \sigma_S^2 - (x - \mu_S)^2 \cdot \sigma_{-S}^2 = 2k \cdot \sigma_S^2 \cdot \sigma_{-S}^2, \quad \text{for } x = \underline{x}_S(k), \bar{x}_S(k)$$

and

$$\int_{\underline{x}_S(k)}^{\bar{x}_S(k)} f(x|S)dx = \delta.$$

*iii) Suppose that  $\sigma_S > \sigma_{-S}$ . Then, the stereotype for  $S$  is  $X_S = (-\infty, \underline{x}_S] \cup [\bar{x}_S, +\infty)$  where  $\underline{x}_S, \bar{x}_S$  are implicitly identified by the system of equations (in  $\underline{x}_S(k), \bar{x}_S(k)$  and  $k$ ):*

$$(x - \mu_{-S})^2 \cdot \sigma_S^2 - (x - \mu_S)^2 \cdot \sigma_{-S}^2 = 2k \cdot \sigma_S^2 \cdot \sigma_{-S}^2, \quad \text{for } x = \underline{x}_S(k), \bar{x}_S(k)$$

and

$$\int_{-\infty}^{\underline{x}_S(k)} f(x|S)dx + \int_{\bar{x}_S(k)}^{+\infty} f(x|S)dx = \delta.$$

When the two distributions have the same variance, the stereotype is formed by truncating from the original distribution the least representative tail. This parallels the findings of Section 3. If groups differ only in their means, as in case i), then MLRP holds. When the mean in  $S$  is above the mean in  $-S$ , then MLRP is increasing and the exemplar for the former group is  $+\infty$ ; otherwise it is  $-\infty$ . The stereotype then truncates the “non-exemplar tail.” In both cases the exemplar is bad because it relies on a highly infrequent realization.

Figure 3 represents the distribution considered by the agent for the high mean group when traits are normally distributed with the same variance across groups.

In this example, the true mean  $\mu_S$  is included in the support, which in turn means that the value of  $\delta$  is above .5, e.g.  $\delta = .7$ . It is evident that in this case both the assessed mean is above  $\mu_S$  and the assessed variance is below the true variance  $\sigma_S$ . Both features are due to the fact that the distribution is distorted towards the group exemplar at  $+\infty$ .

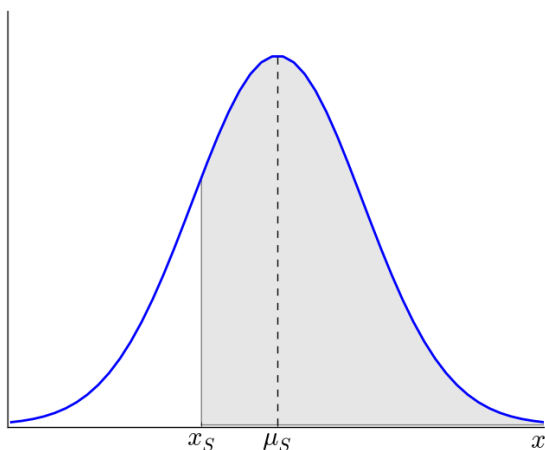


Figure 3: Stereotype of a Normal distribution when  $\sigma_S = \sigma_{-S}$  and  $\mu_S > \mu_{-S}$ .

Consider now case ii), where the variance of  $S$  is below that of  $-S$ . The stereotype consists of an interval around an intermediate exemplar. In particular, the exemplar for group  $S$  is now equal to  $\mu_S \cdot [(\sigma_{-S} - (\mu_{-S}/\mu_S) \cdot \sigma_S)/(\sigma_{-S} - \sigma_S)]$ . As in Proposition 1, when the distribution in  $S$  is more concentrated than that in  $-S$ , the exemplar is good, in the sense that it captures a relatively frequent, intermediate event. It is however somewhat distorted, because  $\hat{x}_S$  lies below the group's true mean  $\mu_S$  if and only if  $\mu_S < \mu_{-S}$ . Interestingly, when the mean in the two groups is the same, the low variability group is represented by its correct mean, namely  $\mu_S$ .

Finally, consider case iii). Now the variance in  $S$  is higher than that in  $-S$ . As a consequence, both tails are exemplars and the stereotype includes both tails, truncating away an intermediate Section of  $R$ . As we saw in Section 3, this representation enhances volatility and thus captures the distinctive trait of  $S$ , which is precisely its high variability.

## 5.2 Reaction to Information in the Normal Case

Suppose that starting from the initial date  $t = 1$ , in each period  $t$  a sample  $(x_{S,t}, x_{-S,t})$  of outcomes is observed, drawn from the two normal groups. The history of observations up to period  $\tau$  is denoted by the vector  $\mathbf{x}^\tau = (x_{S,t}, x_{-S,t})_{t=1, \dots, \tau}$ .

Based on  $\mathbf{x}^\tau$ , and thus on the conditional distributions  $f(x|W, \mathbf{x}^\tau)$  for  $W = S, -S$ , the

agent updates stereotypes and beliefs. In one tractable case, the  $t = 0$  initial distribution  $f(x|W)$  is normal for  $W = S, -S$ . Formally, suppose that  $x_W = \theta_W + \epsilon_W$  where  $\epsilon_W$  is i.i.d. normally distributed with mean 0 and variance  $v$ , and  $\theta_W$  is the group specific mean. Initially, the two groups are believed to be identical, in the sense that  $\theta_W$  is normally distributed with mean 0 and variance  $\gamma$ . After observing  $(x_{S,1}, x_{-S,1})$ , the distribution of  $\theta_W$  is updated according to Bayesian learning. Updating continues as progressively more observations are learned.

Bayesian learning implies that at time  $\tau$ , after observing the sample  $x^\tau$ , we have:

$$f(x|W, \mathbf{x}^\tau) = \mathcal{N}\left(\frac{\gamma \cdot \tau}{v + \gamma \cdot \tau} \cdot \frac{\sum x_{W,t}}{\tau}; v \cdot \frac{v + \gamma \cdot (\tau + 1)}{v + \gamma \cdot \tau}\right). \quad (7)$$

The posterior mean for group  $W$  is an increasing function of the sample mean  $\sum x_{W,t}/\tau$  for the same group. The variance of the posterior declines in sample size  $\tau$ , because the building of progressively more observations reduces the variance of  $\theta_W$ , in turn reducing the variability of possible outcomes. Equation (7) yields the following result.

**Proposition 6** *At time  $\tau$ , the exemplar for group  $S$  is equal to  $+\infty$  if  $\sum x_{S,t} > \sum x_{-S,t}$  and to  $-\infty$  if  $\sum x_{S,t} < \sum x_{-S,t}$ . As a result:*

*i) Gradual improvement of the performance of group  $S$  does not improve that group's exemplar (and only marginally affects its stereotype) provided  $\sum x_{S,t}$  stays below  $\sum x_{-S,t}$ . In particular, common improvements in the performance of  $S$  and  $-S$  (which leave  $\sum x_{S,t} - \sum x_{-S,t}$  constant) leave stereotypes unaffected.*

*ii) Small improvements in the relative performance of  $S$  that switch the sign of  $\sum x_{S,t} - \sum x_{-S,t}$  have a drastic effect on exemplars and stereotypes.*

Even in the normal case, the process of stereotyping suffers from both under- and over-reaction to information. If the ranking between groups stays constant, exemplars do not change. Thus, even if a group gradually improves, its stereotype may remain bad. Once more, this is because stereotypes depend on relative, not on absolute, differences between groups. On the other hand, even small pieces of information can cause a strong over-reaction if they reverse the ranking between groups.

## 6 Conclusion

We have presented a model of stereotypes, in which decision makers assessing a group recall only a limited range of its types or features from memory. Recall is limited but also selective: the recalled types are not the most likely ones given the DM's data, but rather the most representative ones in the sense of being the most ex post informative about the group relative to other groups.

We use this approach to describe how stereotypes form. Because less frequent types are *ceteris paribus* more representative, the model generates a bias towards unlikely stereotypes. Thus even if a majority of poor people are white, decision makers might predict that a poor person is more likely to be black, because poverty is higher among blacks than among whites. We also examine which features of groups are most likely to be picked for stereotyping, and how DMs update stereotypes in reaction to new information. The model generates a number of predictions that are consistent with available evidence and intuition.

## References

- Arrow, Kenneth. 1973. "The Theory of Discrimination." In Orley Ashenfelter and Albert Rees, eds. *Discrimination in Labor Markets*. Princeton, N.J.: Princeton University Press: 3 - 33.
- Barberis, Nicholas, Andrei Shleifer, and Robert Vishny. 1998. "A Model of Investor Sentiment." *Journal of Financial Economics* 49 (3): 307 – 343.
- Bodoh-Creed, Aaron, Dan Benjamin and Matthew Rabin. 2013. "The Dynamics of Base Rate Neglect." Mimeo Haas Business School.
- Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice under Risk." *Quarterly Journal of Economics* 127 (3): 1243 – 1285.
- Bordalo, Pedro, Nicola Gennaioli and Andrei Shleifer. 2013. "Salience and Consumer Choice." *Journal of Political Economy* 121 (5): 42.
- Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2013b. "Salience and Asset Prices," *American Economic Review, Papers & Proceedings*, 623 – 628.
- Fryer, Roland, and Matthew Jackson. "Categorical Cognition: A Psychological Model of Categories and Identification in Decision Making." In Proceedings of the 9th conference on Theoretical aspects of rationality and knowledge, pp. 29 – 34. ACM, 2007.
- Gennaioli, Nicola, and Andrei Shleifer. 2010. "What Comes to Mind." *Quarterly Journal of Economics* 125 (4): 1399 – 1433.
- Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny. 2012. "Neglected Risks, Financial Innovation, and Financial Fragility." *Journal of Financial Economics* 104 (3): 452 – 468.
- Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny. 2013. "A Model of Shadow Banking." *Journal of Finance* 68 (4): 1331 – 1363.
- Grether, David. 1980. "Bayes Rule as a Descriptive Model: The Representativeness Heuristic." *Quarterly Journal of Economics* 95 (3):537 – 557.
- Hilton, James, and William Von Hippel. 1996. "Stereotypes." *Annual Review of Psychology* 47 (1): 237 – 271.

- Kahneman, Daniel, and Amos Tversky. 1972. "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology* 3 (3): 430 – 454.
- Lakonishok, Josef, Andrei Shleifer, and Robert Vishny. "Contrarian Investment, Extrapolation, and Risk." *Journal of Finance* 49 (5): 1541 – 1578.
- Mullainathan, Sendhil. 2002. "Thinking through Categories", Mimeo Harvard University.
- Nickerson, Raymond. 1998. "Confirmation Bias: a Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2(2): 175 – 220.
- Phelps, Edmund. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review* 62 (4): 659 – 661.
- Schwartzstein, Joshua. 2012. "Selective Attention and Learning." Unpublished Manuscript, Harvard University.
- Tajfel, Henri. 1982. "Social Psychology of Intergroup Relations." *Annual Review of Psychology* 33 (1): 1 – 39.
- Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185 (4157): 1124 – 1131.

## Proofs

**Remark 1.** We first establish that representativeness  $R(x, S)$  of a type  $x$  for group  $S$  increases in the likelihood ratio  $Pr(X = x|S)/Pr(X = x|-S)$ . From Definition 1,  $R(x, S) = Pr(S|X = x)/Pr(-S|X = x)$ . Using Bayes' rule,  $Pr(S|X = x) = Pr(X = x|S) \cdot Pr(S)/Pr(X = x)$  and similarly for  $-S$ , leading to

$$R(x, S) = \frac{Pr(X = x|S)}{Pr(X = x|-S)} \cdot \frac{Pr(S)}{Pr(-S)}$$

Given that  $Pr(S)$  and  $Pr(-S)$  do not depend on  $x$ , the result follows.

Denote the conditional probability  $Pr(X = x|W)$  by  $\pi_{x,W}$ . Rewrite the likelihood ratio as

$$\frac{\pi_{x,S}}{\pi_{x,-S}} = \frac{(\pi_{x,S} - \pi_{x,-S}) + \pi_{x,-S}}{\pi_{x,-S}}$$

from which result i) follows immediately. Moreover, it is clear that keeping the difference  $\pi_{x,S} - \pi_{x,-S}$  fixed, the likelihood ratio decreases with the baseline probability  $\pi_{x,-S}$  if and only if the difference is positive. ■

**Proposition 1.** The likelihood ratio of type  $x$  is (up to a normalizing constant) equal to  $\pi_{x,S}^{1-\alpha}$ . If  $\alpha < 1$ , then this ratio increases with the probability  $\pi_{x,S}$  of  $x$ , so that the exemplar is the most likely type,  $\operatorname{argmax}_x \pi_{x,S}$ . If  $\alpha > 1$ , then the ratio decreases with  $\pi_{x,S}$ , so that the exemplar is the least likely type,  $\operatorname{argmin}_x \pi_{x,S}$ . When  $\alpha = 1$ , then the groups  $S$  and  $-S$  have identical distributions and all types are equally representative. ■

**Corollary 1.** From Proposition 1, when  $\alpha \leq 0$  the exemplar of population  $S$  is its most likely type. For population  $-S$ , the relevant likelihood ratio is  $\pi^* \pi_{x,S}^{\alpha-1}$ . So the exemplar of  $-S$  is population  $S$ 's least likely type,  $\operatorname{argmin}_x \pi_{x,S}$ , which (because  $\alpha \leq 0$ ) is (one of)  $-S$ 's most likely type,  $\operatorname{argmax}_x \pi_{x,-S}$ . This shows case i). Case ii) follows from Proposition 1. ■

**Corollary 2.** Follows from Definition 1. ■

**Corollary 3.** From Proposition 1, we have  $\pi_{x,S}/\pi_{x,-S} \propto \pi_{x,S}^{1-\alpha}$ . Thus, if  $\alpha < 1$ , the ranking of types by increasing representativeness or by increasing likelihood is identical. Therefore

the stereotype for  $S$  consists of its  $d$  most likely types.

If instead  $\alpha > 1$ , the ranking of types by increasing representativeness or by *decreasing* likelihood is identical. Therefore the stereotype for  $S$  consists of its  $d$  *least* likely types. ■

**Corollary 4.** Index the types  $x \in \{1, \dots, N\}$  according to the “natural” ordering relation (e.g. type 1 is on the left and type  $N$  is on the right). Suppose the likelihood ratio is monotonically decreasing in  $x$ . Then the ordering of types by representativeness coincides with the natural ordering of types, so that the stereotype consists of types 1 through  $d$ . By truncating the upper tail, it follows that  $\mathbb{E}^{st}(x|S) < \mathbb{E}(x|S)$ .

If the the likelihood ratio is monotonically increasing in  $x$ , then the ordering of types by representativeness coincides with the inverse of the natural ordering of types, so that the stereotype consists of types  $N - d + 1$  through  $N$ . By truncating the lower tail, it follows that  $\mathbb{E}^{st}(x|S) > \mathbb{E}(x|S)$ .

Suppose that the likelihood ratio is U-shaped and symmetric in the natural ordering of types. Then the ranking of types by representativeness decreases with the distance to the extremes. Therefore, supposing for simplicity that  $d$  is even, the stereotype will consist of types 1 through  $d/2$  and  $N - d/2 + 1$  through  $N$ . ■

**Proposition 2.** Following the assumptions of the proposition, write  $Pr(z|y, S) = Pr(z|S) \cdot \phi(x, y)$  and  $Pr(z|y, -S) = Pr(z|-S) \cdot \phi(x, y)$ . We now describe the most extreme way that the stereotype may be organised along dimension  $Y$ , in which all variation along dimension  $Z$  is taken into account, namely  $d_Z = |N_Z|$  (maximal) and  $d_Y = d/|N_Z|$  (minimal). The representativeness of type  $(y, z)$  is given by

$$\frac{Pr(z|y, S)}{Pr(z|y, -S)} \cdot \frac{Pr(y|S)}{Pr(y|-S)} = \frac{Pr(z|S)}{Pr(z|-S)} \cdot \frac{Pr(y|S)}{Pr(y|-S)}$$

Because the representativeness of type  $(y, z)$  increases in the representativeness of  $y$  keeping  $z$  fixed (and vice versa), it is useful to consider the ranking of (unconditional) types  $y \in Y$  and  $z \in Z$ . Let  $y_i$  (resp.  $z_i$ ) denote the  $i$ -th most representative type in  $Y$  (resp.  $Z$ ). Then, intuitively, the stereotype organises around  $Y$  if the variation in representativeness along the entire  $Z$  dimension is smaller than the variation in representativeness between any two types



in  $Y$ . Formally, the representativeness ranking is lexicography if and only if

$$\frac{Pr(z_1|S)}{Pr(z_1|-S)} \Big/ \frac{Pr(z_{|C_Z|}|S)}{Pr(z_{|C_Z|}|-S)} < \min_r \frac{Pr(y_r|S)}{Pr(y_r|-S)} \Big/ \frac{Pr(y_{r+1}|S)}{Pr(y_{r+1}|-S)}.$$

■

**Proposition 3.** We assume that the same number of observations are received at each stage of the learning process for both groups  $S$  and  $-S$ . This assumption is not restrictive, since only the relative frequency of observations matter. In particular, all probabilities remained unchanged if the sample size of one group is scaled up relative to the sample size of the other. Thus we can set  $\sum_{x'} a_{x',S} = \sum_{x'} a_{x',-S} = a$  and  $\sum_{x'} n_{x',S} = \sum_{x'} n_{x',-S} = n$ .

Representativeness of a type  $x$  is now measured by the ratio

$$\frac{Pr(X = x|\alpha_S, \mathbf{n}_S)}{Pr(X = x|\alpha_{-S}, \mathbf{n}_{-S})} = \frac{\alpha_{x,S} + n_{x,S}}{\alpha_{x,-S} + n_{x,-S}}$$

Consider case i) where all observations occur in type  $x$ , so that  $n_{x,S} = n$  and  $n_{x',S} = 0$  for  $x' \neq x$ , and similarly for  $-S$ . Then the representativeness of types other than  $x$  do not change, while the representativeness of  $x$  is  $(\alpha_{x,S} + n)/(\alpha_{x,-S} + n_{x,-S})$ . This tends to one monotonically as  $n$  increases. Therefore, if  $a_{x,S}/a_{x,-S} < 1$  then  $(a_{x,S} + n)/(a_{x,-S} + n) < 1$  for all  $n$ : namely, if  $x$  is non-representative to begin with, then no amount of observations of  $x$  in population  $S$  (when accompanied by observations of  $x$  in population  $-S$ ) will make  $x$  representative for  $S$ .

Consider now case ii), where all observations in  $S$  occur in a non-representative type  $x$  while all observations in  $-S$  occur in a representative (for  $S$ ) type  $x'$ . In that case, the representativeness of  $x$  for group  $S$  increases as  $(a_{x,S} + n)/(a_{x,-S})$ , while the representativeness of  $x'$  for group  $S$  decreases as  $(a_{x',S} + n)/(a_{x',-S} + n)$ . The result follows. ■

**Proposition 4.** Consider the case where a single observation of group  $S$  occurring in type  $x$  does not change the representativeness ranking of types – and thus the stereotype – for  $S$ .

If  $x$  is in the stereotype of  $S$ , then its estimated probability is  $a_{x,S}/\sum_{x'=1}^d a_{x',S}$ , which is boosted by a factor of  $\sum_{x'=1}^N a_{x,S}/\sum_{x'=1}^d a_{x',S} > 1$ , where  $d$  is the number of types in the

stereotype. Suppose an observation occurs in type  $x$ . Its representativeness for  $S$  increases, and its assessed probability jumps to  $(a_{x,S} + 1)/(\sum_{x'=1}^d a_{x',S} + 1)$ . This corresponds to a larger increase of assessed probability than that done by a Bayesian whenever

$$\frac{a_{x,S} + 1}{\sum_{x'=1}^d a_{x',S} + 1} - \frac{a_{x,S}}{\sum_{x'=1}^d a_{x',S}} > \frac{a_{x,S} + 1}{\sum_{x'=1}^N a_{x',S} + 1} - \frac{a_{x,S}}{\sum_{x'=1}^N a_{x',S}}$$

namely when

$$\frac{a_{x,S}}{\sum_{x'=1}^N a_{x',S}} < \frac{\sum_{x'=1}^d a_{x',S}}{1 + \sum_{x'=1}^d a_{x',S} + \sum_{x'=1}^N a_{x',S}} < \frac{1}{2}$$

The intuition is that the stereotype ignores some observations, it is as though the probability is being updated over a smaller sample size. Therefore, as long as the prior of  $x$  (in the stereotype) is not too large, the DM boosts it more than the Bayesian.

If  $x$  is not in the stereotype, then – given that the stereotype does not change – it does not become representative. Its assessed probability stays at zero, so the decision maker under-reacts to this observation relative to a Bayesian. ■

**Proposition 5.** Let  $\rho_{\mu,\sigma}$  denote the probability density of  $\mathcal{N}(\mu, \sigma)$ , namely  $\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ . The exemplar  $\hat{x}_S$  of  $S \equiv \mathcal{N}(\mu_S, \sigma_S)$  relative to  $-S \equiv \mathcal{N}(\mu_{-S}, \sigma_{-S})$  satisfies  $\hat{x}_E = \operatorname{argmax}_x \frac{\rho_{\mu_S, \sigma_S}}{\rho_{\mu_{-S}, \sigma_{-S}}}$  where

$$\frac{\rho_{\mu_S, \sigma_S}}{\rho_{\mu_{-S}, \sigma_{-S}}} = \frac{\sigma_{-S}}{\sigma_S} \cdot \exp \left\{ -x^2 \left( \frac{1}{2\sigma_S^2} - \frac{1}{2\sigma_{-S}^2} \right) + x \left( \frac{\mu_S}{\sigma_S^2} - \frac{\mu_{-S}}{\sigma_{-S}^2} \right) - \left( \frac{\mu_S^2}{2\sigma_S^2} - \frac{\mu_{-S}^2}{2\sigma_{-S}^2} \right) \right\}$$

When  $\sigma_S < \sigma_{-S}$ , the function above has a single maximum in  $x$ , namely that which maximizes the parabola in the exponent,  $\hat{x}_E = \frac{\frac{\mu_S}{\sigma_S^2} - \frac{\mu_{-S}}{\sigma_{-S}^2}}{\frac{1}{\sigma_S^2} - \frac{1}{\sigma_{-S}^2}}$  from which the result follows.

When  $\sigma_S > \sigma_{-S}$ , the function above is grows without bounds with  $|x|$ , so that  $\hat{x}_S \in \{-\infty, +\infty\}$ .

When  $\sigma_S = \sigma_{-S} = \sigma$ , the exemplar  $\hat{x}_S$  of  $S \equiv \mathcal{N}(\mu_S, \sigma)$  relative to  $-S \equiv \mathcal{N}(\mu_{-S}, \sigma)$  satisfies

$$\hat{x}_S = \operatorname{argmax}_x e^{-\frac{\mu_S^2 - \mu_{-S}^2}{2\sigma^2}} \cdot e^{\frac{x}{2\sigma^2}(\mu_S - \mu_{-S})}$$

so that  $\hat{x}_S = -\infty$  if  $\mu_S < \mu_{-S}$  and  $\hat{x}_S = +\infty$  otherwise. If  $\mu_S < \mu_{-S}$  all values of  $x$  are

equally representative. ■

**Corollary 5.** Consider case i): because  $\sigma_S = \sigma_{-S} = \sigma$ , MLRP holds. The likelihood ratio for  $S$  is increasing in  $x$  if  $\mu_S > \mu_{-S}$ . In this case, it follows from Definition 2' that the stereotype for  $S$  consists of the right tail, down to the threshold  $x_S$  that satisfies  $\int_{x_S}^{sup(X)} f(x|S)dx = \delta$ . A similar reasoning applies in the case where  $\mu_S < \mu_{-S}$ .

Consider now case ii). When  $\sigma_S < \sigma_{-S}$ , population  $S$  has thinner tails than  $-S$  so that its most representative types lie in an intermediate range of  $x$ . Formally, we know from Proposition 5 that  $R(x, S)$  is given by  $\frac{\rho_{\mu_S, \sigma_S}}{\rho_{\mu_{-S}, \sigma_{-S}}}$  which is (the exponential of) a quadratic function of  $x$ , and that because  $\sigma_S < \sigma_{-S}$  this function is inverse-U shaped. The stereotype selects the types  $x$  whose representativeness is greater or equal to  $\frac{\sigma_{-S}}{\sigma_S} e^k$  for some  $k$ . Because representativeness is quadratic, this constraint selects two types,  $\underline{x}_S(k), \bar{x}_S(k)$  and a constant  $k$  that simultaneously satisfy  $\frac{\rho_{\mu_S, \sigma_S}}{\rho_{\mu_{-S}, \sigma_{-S}}}(\underline{x}_S(k)) = \frac{\rho_{\mu_S, \sigma_S}}{\rho_{\mu_{-S}, \sigma_{-S}}}(\bar{x}_S(k)) = \frac{\sigma_{-S}}{\sigma_S} e^k$  and  $\int_{\underline{x}_S(k)}^{\bar{x}_S(k)} f(x, S)dx = \delta$ .

A similar reasoning applies in the case where  $\sigma_S > \sigma_{-S}$ . ■

**Proposition 6.** Since the variances of the sample populations  $S$  and  $-S$  are equal, the stereotypes are fully determined by the sample means. From Proposition 5, if  $\sum_t x_{S,t} > \sum_t x_{-S,t}$ , then the sample mean of  $S$  is larger than that of  $-S$ , so that its exemplar is  $\hat{x}_S = +\infty$ . If instead  $\sum_t x_{S,t} < \sum_t x_{-S,t}$ , the exemplar of  $S$  is  $\hat{x}_S = -\infty$ . ■