Three papers in neuroeconomics

Thesis by

Meghana Bhatt

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2007
(Defended May 7, 2007)

Acknowledgements

First and foremost I would like to thank my advisor, Colin Camerer, for giving me the unique opportunity to work in an exciting new field, encouraging me to explore new ideas and teaching me a about how to be a scientist. I'd also like to thank Jacob Goeree, Antonio Rangel, and Jernej Copic for enduring some of the early versions of this thesis. Without them, no one would ever be able to read the first chapter of this dissertation. I would also like to thank my co-authors Terry Lohrenz and Read Montague for their contributions on the third chapter.

I've also gotten many comments and suggestions from friends and colleagues. I'd like to thank Alex Brown, Jessica Edwards, Ming Hsu, Asha Iyer, Ian Krabich, Galen Loram, Maggie McConnell, Debrah Meloso, Lauren Munyan, Hilke Plassmann, Elena Reutskaja, Ryan Riegg, Michael Spezio, Tomomi Tanaka, Joseph Wang, and Kathering Werner; and participants of seminars at the ACR and Neuroeconomics Conferences.

More generally I'd like to thank Laurel Auchampaugh and the rest of the HSS staff for always helping me out despite my tendency to forget all things administrative. I'd also like to thank the women of Vox Femina for four years of great singing and always reminding me that there is a world outside of academia; my parents for always supporting me in whatever I try to do; and finally, Jacob Berlin, I can't imagine the last 5 years without him.

Abstract

I consider the role of automatic psychological and neural processes in different settings. First, how does advertising affect consumer perceptions of a product? I assert that one of the major mechanisms used in marketing is the creation of implicit associations among concepts. A neural-network framework is adapted to model how these associations evolve and interact. Use of this formal model allows us to consider some of the indirect effects of advertising, particularly "spillover" and "dilution" effects where the advertisements for one product can help or harm ther perceptions of another. Second, I use fMRI to consider neural activation patterns during strategic thinking. Two studies reveal the possible importance of various neural areas to belief formation, strategic deception, and suspicion. The first study focuses on the neural correlates of belief formation, particularly the difference between equilibrium and out-of-equilibrium decisions. We find both neural activations and behavioral evidence that subjects have relatively shallow belief processes, often apparently assigning too much agency to themselves, when they are out of equilibrium. The second study focuses on a bargaining interaction where a "buyer" has incentive to deceive a "seller" in order to get a low price for a hypothetical product. We find neural correlates to strategic deception in the dorsal striatum and to suspicion of deception in the ACC.

Contents

Contents

Contents

Chapter 1

Evaluation and associations: a network model of advertising and consumer choice

1.1 Introduction

The existence of advertising is one of the major open problems in economics. It presents difficulties in traditional economic models because advertisements are meant to change the decisions of consumers. The only way this should be possible in a model with stable preferences is if (a) advertisements change the information available to the consumer, or (b) consumption of the advertisement acts as a complement to the product. However, uninformative advertising is commonplace. It has been suggested that uninformative ads may be a form of "money-burning" (59; 55): firms conspicuously spend money to signal that its products are of high quality and will attract repeat consumers, i.e., only companies with high quality products can afford to advertise. But if this were the case, we would expect that (a) all uninformative advertising must be expensive and (b) all such advertising would prominently display its cost; however, neither of these implications hold. The complementary view seems more plausible, but in order to understand the wider effects of advertising, it would be useful to understand the mechanisms underlying these complementarities.

The empirical results from the marketing literature yield very few consistent results. For example, a successful brand extension[1] may help or harm sales for the brand's other products (27; 56; 18; 49; 40). Conversely, using a brand extension as opposed to creating a new brand may or may not increase the probability that a product will succeed. Such contradictory findings indicate the existence of important underlying variables that determine the existence and direction of the effects of advertising.

One proposed candidate for these underlying variables is the set of attribute-specific associations

[1] A brand extension is a new product that is introduced under an established brand name

generated by a given brand or product (49; 38). In this paper I take the view that advertising affects the associations among possible product attributes and that these associations in turn mediate how we perceive any given product. The seemingly contradictory effects of common marketing strategies such as brand extensions can be understood better by considering how (a) these associations evolve in response to advertising and (b) how these associations interact to affect a consumer's perception of any given product thereby affecting the consumer's willingness to pay for the product. I use a neurally inspired network model address these questions. This model is largely based on one simple principle — the strength of the association between two attributes is proportional to the *observed* correlation of those attributes in the experience of the consumer where this experience includes advertisements.

Marketing professionals take for granted the assumption that perception, distinguished from fact or rational beliefs in the economic sense, are the most important factor in marketing. This view is nicely summarized in a quote by Greg Stine:

> Every day, buyers like you and I make judgments about quality. But how much of our assessment is due to the actual quality of the item? Not much, I'm afraid. It's not all about quality, but the perception of quality that really counts. – Greg Stine, CEO Polaris Branding Solutions (71)

This raises the question: when do perceptions really diverge from rational beliefs? People choose one option over another for a variety of reasons. Sometimes these reasons are easily identified and quantified. One car may have a larger trunk or a faster engine than another. However, just as often decisions are made based on more abstract and difficult-to-quantify factors - like "luxury" or "fashion." These factors often depend heavily on the associations people have between concepts. The effects of these associations can take a variety of forms. For example, certain products, or even simply attributes of products, can become associated with social traits, celebrities, or emotionally loaded concepts. These associations can in turn affect a decision-maker's evaluation of a product. A prominent example is the phenomenon of celebrity endorsements: firms hope to improve consumer perceptions of their product by associating it with a well-known personality.

On the other hand, consumers' associations can cause cognitive dissonance often leading to more negative evaluations of products. This can often cause problems for firms when they attempt to enter new markets. A firm well known for producing mini-vans might have a difficult time expanding to selling motorcycles because the associations with these products are in conflict. This can occur even though they may exist useful supply-side scale economies allowing the firm to produce motorcycles very efficiently.

There are some difficult-to-quantify characteristics that are almost universally seen as good — for example, attractiveness — but depend on culturally determined, and often varying quantifiable physical characteristics. Advertisements figure prominently in determining *which* physicial characteristics are associated with these abstract characteristics. For example, in any given year, advertisers will help determine what makes a skirt fashionable, including its length, fabric and pattern. In essence, advertisements create complementarities among sets of physical attributes, allowing the combination to be "attractive," or "fashionable." I call these sorts of attributes "social characteristics" in this paper.

The importance of associations for evaluating goods is apparent in the marketing literature. In fact, Kevin Keller defines "brand image" as the "perceptions about a brand as reflected by the brand associations held in consumer memory." (38) Using this definition, in order for brand image to translate into brand *equity*, i.e., increase the sales of a particular product, the brand's associations must increase the utility of the product for some portion of the population. Similarly, the existence of industry-wide advertising campaigns ("Got Milk?") indicate that firms believe that creating positive associations with an entire product category will increase sales for that industry generally. The effectiveness of both brand specific and industry-wide advertising can be explained by saying that the advertisements provide information about the quality of a brand (Hondas are well-made cars) and the usefulness of a product (milk is healthy). But, when these advertisements pertain to social characteristics they go beyond sending information and actually help determine the observable traits that constitute the social characteristics.

However, as noted above, implicit associations can behave in surprising ways leading to the contradictory results from the marketing literature. A major focus of this paper is to outline a framework that can predict this behavior. I will consider two general classes of effects. The first, which I will call "spillover" effects, refers to the positive effects an advertisement on products other than the advertised product. These include the industry-wide positive effects of a single firm's advertising campaign, the benefits of the association with a parent brand for a brand extension, the reciprocal benefits to the parent brand from association with a successful extension, and the benefits of firm associations with other firms. The second class, which I will call "dilution" effects, refers to the negative externalities of advertising. These include the possibility of decreasing brand equity due to a failed or incongruous brand extension, as well as the two forms of legally actionable trademark dilution: The first is called "tarnishing" in legal parlance, and refers to any advertising that introduces negative associations with a famous trademark; the second is called "blurring", and refers to advertising that weakens the existing association of a famous mark.

The legal standard in trademark law is that the mark must have "acquired secondary meaning" in order to be protected under the law. The law also states that firms may not trademark any "functional" (i.e., utility bearing) part of a product. These laws can apply broadly, not only to a brand logo or name, but to distinctive packaging or design features such as color (referred to as "trade dress"). This raises the questions: how do these marks acquire meaning and exactly how are

are they affected by advertisements for both the brand owning the trademark *and* other brands? In some ways trademarks act like patents, they protect a firm's investment in its brand image the same way a patent protects the firm's investment in research and development. In order to address how this investment can be harmed through dilution or infringement we need a model of exactly *how* the mark acquires or loses secondary meaning.

In this paper I will introduce a neural-network framework that not only models these associations and how they interact, but how these associations evolve over time in response to the personal and cultural experiences of an individual. The structure of these neural-networks are based on existing models in computational neuroscience and cognitive psychology that are meant to reflect the way the human mind actually processes information. The interconnected nature of the network structure allows us to make predictions about how an advertisement that is meant to create a specific association will affect the other associations in the network, producing the externalities mentioned above.

Finally, by modeling the underlying mechanisms of how associations are formed and how they interact I can generate predictions on the firm level about both advertising levels and different advertising strategies.

1.2 Motivation and Background

Until relatively recently, economists have viewed utility as purely a function of observed measurable characteristics, treating the mind as a black-box. The viewpoint that economics is purely concerned with choices parallels in some ways the behaviorist school of psychology, which similarly focuses purely on behavior and pays no attention to the underlying cognitive processes.

The reduction of the choice problem to a simple optimization has been incredibly useful for making predictions not only about individual decisions but also about larger systems like markets. It is especially useful because it allows economists to use well-defined mathematical principles. Utility maximization has the additional property of being extremely flexible. Economists have been able to describe a number of behavioral phenomena by introducing parameters into the utility function to account for social preferences including fairness and jealousy (22; 63), discrepancies in inter-temporal choice (44), and reference dependence (35).

However, there are cases where the underlying cognitive process, such as memory and attention, are not only relevant, but also essential to understanding the decision-making process. For example, our limited attention may not constrain our overall choice set, but will constrain the choices that are salient at any given moment. These constraints are illustrated by the popular example of the American tourist crossing a street in London. The American crossing the street mistakenly looks left instead of right not because looking right is not feasible, but because the experiences of the

American make looking left more automated and salient. Our experiences, both explicit memories of specific events and implicit associations among concepts, affect the way we perceive things, from people to consumer products. And since we are constantly experiencing the world these memories and associations are always changing.

Consider the following examples:

- In 2003 the lead singer of the Dixie Chicks, a popular country music trio, made a controversial statement denouncing President George W. Bush: "Just so you know, we're ashamed the President of the United States is from Texas." As a country group, much of their core audience were in favor of the President and had a very negative reaction to the remark. They were so angered by the remark that some fans publicly threw out Dixie Chicks CD's and concert tickets (20). While the music itself had not changed, the perception of the group as "unpatriotic" caused a drastic preference reversal for these fans. Borrowing the language of trademark dilution, this remark "tarnished" the Dixie Chicks image by associating them with anti-president and anti-war views.

  However, for listeners who were already opposed to Bush, the same statement had a positive effect on their view of the group. One fan wrote on a message board:

  > Hell, I'm ashamed [Bush is] American and I am doubly ashamed that the people of this country were stupid enough to elect him for a second term. I will support the Dixie Chicks in any way possible including attending their concert here in Dallas. YOU GO NATALIE!!!! (21)

  This fan actually states an extra willingness to *purchase* products from the group because of their statement. For these fans, the criticism of Bush implied the same associations, but for them these associations were positive.

  "Consuming" (hearing) the statement acted as a complement to consumption of the groups CD's and concert by changing the groups image. Depending on the personal preferences of listeners it could be highly negative or highly positive. We assert that this change in image is the result of shifting associations. Even though this statement was not an advertisement, it changed perceptions and choices in the same ways that ads do.

- In 1993, Heinz began selling an "All-Natural Cleaning Vinegar." The company already successfully sold a line of edible vinegar that consumers often used as a cleaning solution. The new product had twice as much acid and was more effective than its edible counterpart. The new product failed despite the facts that (a) Heinz had a well-established reputation for making vinegar, (b) Heinz kitchen vinegar was often used as a cleaning solution, and (c) the new product was specifically designed for cleaning and would have been more effective for the

> task. BusinessWeek explained the failure, saying that the product failed “in part because the packaging was too similar to the original Heinz vinegar; consumers didn’t like buying nearly identical bottles of very different products, despite the clear labeling” (33).
>
> As in the first example, we can model this outcome by saying that the kitchen variety of vinegar, or just the advertisements for it, were negative complements to the new product. But in this instance arguments could be made a priori that it should be a positive complement for most consumers. The negative effect was the result of a set of interacting associations: Heinz with food, cleaners with poisons, vinegar with cleaners, vinegar with food. The net result of these associations is unclear without some more well-defined framework in which to consider them.
>
> The introduction of new products is big business. Brand extension, the use of an *existing* brand name with a new product, is an extremely common method for introducing products to the market. Companies stand to lose large amounts of money for each failed extension. The frequent failure of brand extensions imply that the interactions of existing brand associations are complicated and may be poorly understood even by experts in the field, such as business executives and advertising professionals. These complexities suggest that it would be be useful to have a better model of *how* their existing products may affect extensions and vice versa.

These two examples illustrate the different effects that associations can have on evaluation. The Dixie Chicks example deals with the simple negative association of a good with the highly controversial figure of George W. Bush, while the vinegar example illustrates a concrete example of how conflicting associations negatively affected the sales of a product.

In both examples, events in the consumer’s life pose positive or negative externalities for the consumption of a good. But in both cases we need a more detailed model to predict the nature of the externality. To do this, revealed preference alone may not suffice. In its purest form, revealed preference can only make predictions about existing goods. Without imposing some extra structure, revealed preference on its own has little to say about new products.

1.2.1 Hedonic pricing with “perceived attributes”

One way to impose such structure is to represent possible products in such a way that they are easy to relate to each other. Hedonic pricing models are a common way to decompose the utility of various goods into the utilities of their attributes (46; 66; 50). Most of this research has focused on goods with many objectively measurable characteristics. For example, a house can be described as a combination of its size, location, number of bedrooms, number of bathrooms, etc. The decomposition of a product into its component characteristics allows the set of possible products to all be represented in a common space: for example, when characteristics can be given well-defined cardinal values, products would be represented as vectors in $\mathbb{R}^n$, where n is the number of attributes under consideration. Given this representation we can define a utility function $u : \mathbb{R}^n \to \mathbb{R}$ rather than treating each possible product as a distinct atomic unit. The new utility function, defined over a more general space, in turn allows us to predict the utility of novel combinations of attributes.

I apply this same basic structure with the caveat that the attributes of the product are a function of both its objective characteristics and the associations in the mind of the consumer. In the first example above, the objective characteristics of the music recordings remain the same, but the remark changes the consumers' associations with those recordings, thus changing their subjective evaluation. Similarly in the second example, the objective measures of the product, i.e., cleaning ability, should make it preferred as a cleaning product to the existing kitchen vinegar; however, the association of the cleaning product with a food product caused dissonance in consumers' minds, and they would not buy it.

For any particular product there are two important questions: What characteristics do decision makers perceive as associated with the product, and how are these perceptions guided by their experience (in this paper, advertisements)? I address this issue by introducing the concept of the "perceived characteristics" of a product as distinct from the initial, quantifiable, characteristics of the product. I assume here that consumers derive their utility for a product from the characteristics they perceive rather than the objectively assessed characteristics they initially see. Sometimes perceptions will be close to objective facts and standard hedonic pricing models are appropriate and effective. In cases where this is not true, we need to formulate a model of how the characteristics presented to a decision maker are related to their perceptions via associations.

In this paper, I add one layer of cognitive processing that transforms the initial characteristics of a product into its perceived characteristics. I am primarily interested in how experiences modulate the implicit associations of an individual, and how these in turn affect a decision maker's perception of abstract and difficult-to-quantify characteristics. I use a neural-network model of pattern recognition which makes perceived characteristics a function of the initial objective characteristics and the experiences of the consumer (see Figure 1.1).

I build on the basic construction of hedonic pricing models by assuming that decision makers have utility over bundles of characteristics, defining characteristics very broadly. While anything generally classified as a product attribute is a characteristic, so are features such as the product type itself, and the "brand" of a product.

In my network model each characteristic will be represented by a node in the network while the associations between characteristics will be represented by the connections between these nodes. Unlike the standard hedonic pricing model where characteristics are independent from each other, this network structure will model the way characteristics interact to change the decision maker's eventual perception of a product. In the Heinz cleaning vinegar example above, the consumers'

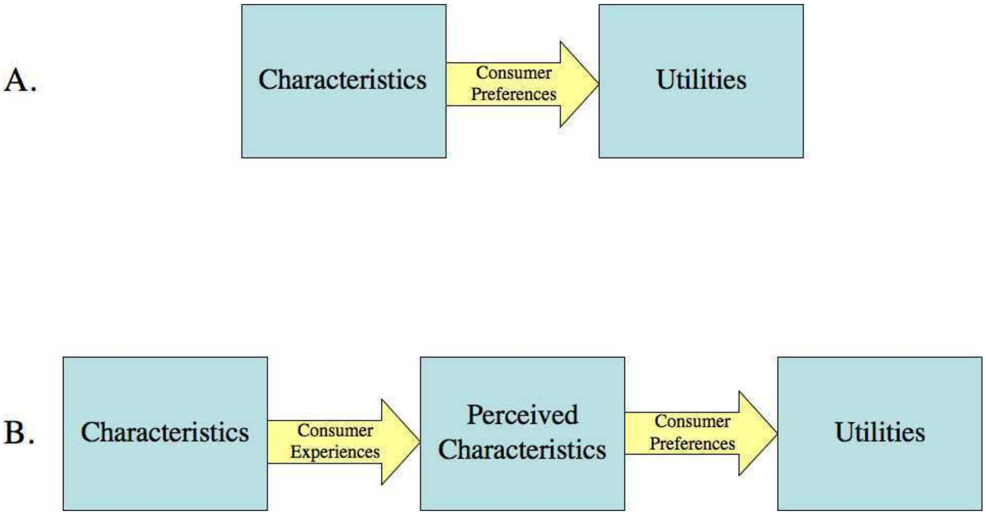


Figure 1.1: A) Standard hedonic pricing models use a function to take sets of attributes directly to utilities. B) In this model I add an extra level of processing. First the initial set of characteristics is transformed to a "perceived" set of characteristics based on the experiences of the consumer. Then I use a utility function to take the perceived attributes to utilities.

experiences with the edible version of Heinz vinegar and with cleaning products created associations between Heinz and food, and cleaning products and poison, respectively. These associations then changed the way consumers perceived the new product, Heinz Cleaning Vinegar. Similarly, in the Dixie Chicks example, the singer's remark cause fans to associate Dixie Chicks music with anti-Bush sentiment and more importantly, a host of other implied political beliefs changing the music's utility for listeners.

Combining the assertions that (a) a decision maker's experiences determine his associations among characteristics and (b) these associations interact to affect the way he perceives the characteristics of goods implies that the experiences of the decision maker will constrain the set of possible *perceived* bundles of characteristics. Notice that the first of these assertions deals with dynamic properties of cognition: experiences change associations. While the second deals with static properties: existing associations modulate perceptions. The contribution of this model is to:

- Specify these constraints on the perceived bundles of characteristics.

- Specify a tractable, formal way to represent the decision maker's experiences.

- Describe how these experiences will affect the perception of the decision maker.

- Derive optimal firm behavior given the effect of advertisements on consumer perceptions.

I will use the network model mentioned above to formalize these processes. It is important to note that while this is a neurally inspired network model, and it follows some very basic neural facts, it is not meant to be a literal interpretation of what is going on in the brain. Such a model would quickly become intractable. In this paper I try to include just enough neural detail to capture some of the major characteristics of perception while still using a relatively low-dimensional abstract representation of what we think is going on in the mind. In addition, this simplified model of cognition allows us to describe a consumer's experiences using a single symmetric matrix: the connection weights of the network summarize the associations built up over a consumer's lifetime. Furthermore, there is a straightforward way to derive these weights from a given set of experiences for the purposes of prediction.

This model draws on research in economics, marketing, cognitive psychology, and neuroscience. To understand where this model is coming from I first look at how economics has treated the problem of advertising historically. I then outline some of the relevant areas of research from the marketing side, including empirical phenomena. Finally I review the roots of the model in cognitive psychology and computational neuroscience.

1.2.2 Economic Models of Advertising

Economic models of advertising fall into three basic categories: informative, persuasive, and complementary (4). Each of these types propose different mechanisms through which advertising works and have differing positive and normative consequences. In addition to these three basic categories I will describe two more recent papers that attribute advertising effects to bounded rationality.

Informative models all assume that any changes in demand caused by advertising are rational in the sense that the advertisement conveys useful information either directly or as a signal of quality. Some models assert that only firms with products of high quality can be confident of repeat purchases. Therefore, firms that advertise are conspicuously spending money, showing that they are confident of high future profits as a result of the quality of the product. According to these and other informative models, advertising serves the positive role of informing consumers and enabling competition. For example, when competing stores use advertisements to inform consumers of their prices it reduces the search costs of the consumer, allowing them to more easily purchase a good at the lowest existing price.

Persuasive models, on the other hand, take the view that while advertising does shift the demand curve for a product outward. In their view the creation of brand loyalty distorts preferences and creates barriers to entry (it makes it more expensive for new firms to enter the market and compete with established brands). Unlike the informative models, these models predict that advertising actually impedes competition.

Finally, complementary models, largely following a model set forth by Becker and Murphy, treat advertisements themselves as “goods” or “bads.” This means that the advertisements themselves enter the decision maker's utility function: the decision maker consumes advertisements the way they consume anything else. While the marginal utility of the advertisements themselves may be positive or negative, the power of these models comes from the effect consumption of advertisements has on the marginal utility of the *advertised good.*[2] In these models advertisements shift the demand curve for a product by entering directly into the utility function. While they may not contribute directly to a consumer's utility, they can effectively change the marginal utility of the product. These models have the benefit of using a stable underlying utility function while still allowing advertising to be persuasive.[3]

1.2.2.1 Informative models

As the name suggests, informative models focus on the role of advertising as a means to convey information. Advertisements first and foremost inform consumers as to the existence of the product and its basic features. But informative models have also held that advertising is in itself a signal of quality. Nelson writes about the function of advertising as "money-burning," a costly signal for producers to show that they are efficient, otherwise they would not have the money to advertise (59).

One particularly useful distinction Nelson makes is that between "search" products and "experience" products. Search products are products that can be evaluated before purchase, whereas experience products must actually be used before a consumer can evaluate them. Since search products can be evaluated before purchase, the signaling power of advertising less important. However, since consumers need to actually use an experience good in order to evaluate its quality, advertising signals are far more relevant to these goods.

While the function of advertising to inform customers of a product's existence is indisputable, the other "money-burning" function of advertising, where ads are simply a costly signal, is problematic. There is an entire industry based around manipulating the *content* of ads. This money-burning account does not allow for differences in skill at creating advertisements. According to this view an ad shown during the Super Bowl would be more effective than any large scale viral marketing campaign[4] regardless of content. These models also imply that firms should actually report the costs of expensive ads, which is almost never done.

1.2.2.2 Persuasive models

Persuasive models of advertising focus on the anti-competitive effects of advertising. In these models advertising creates brand-loyalty and creates a barrier to entry for new firms. At the heart of most

[2] The marginal utility of a discrete good y_0 in some consumption bundle Y is the difference $u(Y) - u(Y - y_0)$. When goods are continuous it is the partial derivative of the utility function $\frac{\partial u(Y)}{\partial y_0}$.

[3] For a more detailed review of this literature, see Bagwell 2005 (4), on which much of this subsection is based.

[4] A viral marketing campaign is one where consumers themselves do the advertising. These campaigns consumers are encouraged to pass along advertisements in the form of amusing video clips, images or a variety of other forms.

of these models is the assumption that advertising somehow changes or distorts a consumer's utility function. Unlike the informative models, persuasive models implicitly assume that the uninformative content of an advertisement is important. In 1933, Joan Robinson stated that "the customer will be influenced by advertisement, which plays upon his mind with studied skill, and makes him prefer the goods of one producer to those of another because they are brought to his notice in a more pleasing and forceful manner" (65). Her reference to the "studied skill" of the advertiser implies that there are hidden mechanisms exploited by advertisers to persuade consumers to buy one product over the other, even if those mechanisms are not well understood on a formal level.

In addition, in these models advertising tends to act as a barrier to entry to a market (5; 36). Advertising allows a manufacturer to differentiate his product from his competitors so that the products of a new entrant are seen as distinct from the established brand. This perceived distinctiveness is crucial to eliciting brand loyalty from consumers; if all brands are the same, there is no reason to prefer one to the other.

Another difference between informative and persuasive models are the effects of the frequency of advertising.[5] Comanor and Wilson assert that advertisements "reinforce the experience that consumers have with established products" (17). This implicitly assumes that there are cognitive processes that are somehow effected by repeated exposure. The model laid out in section 1.3 can account for the effects of repeated exposure to a pattern. These effects are captured by the connection strengths of the neural network. The more frequently two characteristics are associated in the consumer's experience the stronger the connection between the nodes corresponding to those characteristics.

The two major problems with persuasive models are that they do not outline any formal mechanisms through which advertising can shift consumer demand, and they make the normative assumption that any such shifts are distortions.

1.2.2.3 Complementary models

The approach in this paper is closest to the complementary approach to advertising. In complementary models advertisements enter directly into the utility function. While Kaldor first proposed that advertisements be treated as goods in 1950 (36), it was not until recently that Becker and Murphy used this approach to explore the complementary effects of advertisements on consumption (8). Specifically, they asserted that advertisements could change the *marginal* utility of an advertised good, thereby increasing demand. However, since the important effect is on the marginal utility, the advertisement itself may be have positive or negative utility, i.e., it may be either a good or a bad.

[5] Frequency effects may be captured indirectly in Nelson's money burning model since more frequent ads imply greater advertising expenditure to first time users. But the persuasive view, and the model laid out in this paper, assert that frequency will have an effect even if advertising is free. For example, the more often you look at a single print ad, the more the pattern shown in the ad is reinforced even though it has only been paid for once.

This approach has two major benefits over the informative and persuasive views. First, these analyses assume a stable underlying utility function, so welfare can be measured with respect to a stable function. This avoids the problems created by persuasive models where utility is changed thus rendering the choice of which utility function to use for welfare analysis unclear. Second, it allows for a great deal of flexibility as to why advertising might affect demand, encompassing aspects of both the informative and persuasive views.

One major advance made by this model is that it allowed economists to start considering the effect of advertising on the *social* value of products, such as prestige. Further, it doesn't assume that this utility is "false" or a distortion. However, as asserted in the introduction, the very flexibility of this model may be a problem in that it does not impose enough structure on the effects of advertising to make specific predictions.

1.2.2.4 Bounded rationality

Another way to explain the effects of advertising is to assume that they cause consumers to make systematic "mistakes" in how they update their beliefs about a product. Shapiro recently developed a model of advertising that is based on the assumption that consumers will not always remember the source of a positive memory about a product. In this model, advertisements are persuasive because consumers will sometimes mistake a positive advertisement for a positive consumption experience (68). The consumer uses these mistaken experiences to update his beliefs about the quality of the product. Unlike the persuasive models outlined above, the variable affected by advertising is not brand loyalty, but the consumer's beliefs about the quality of the product.

Another similar model by Mullainathan, Schwartzstein and Shleifer relies on what they call "coarse thinking" (58). They assume that consumers divide situations into coarse categories so that a situation where a particular type of advertising message, in their example a celebrity spokesperson, is informative (the show they star in or the sport they play) is pooled with the situation of buying the advertised product. Because of this pooling, the celebrity spokesperson is considered useful information for updating beliefs about the quality of the product. Here the authors assume that rationality is limited by the number of categories consumers have at their disposal rather than by their memory. Like the model described in this paper, they assert that associations are a key factor for persuasion by determining how categorization will take place. However, in their analysis they take categories as exogenous rather than describing exactly how associations might be involved in the process.

These models bear the most resemblance to persuasive models since advertising effects are in essence "mistakes." However, the effects are the result of a rational process, Bayesian updating, with flawed or limited information.

To sum up, all these models assert that advertisements help firms to differentiate their products and increase consumer beliefs about the quality of their products. Informative models would hold that this differentiation is rational, while persuasive models view this differentiation as a mistake. Complementary models on the other hand, show that product differentiation is brought about by the specific effects of advertisements on the marginal utility of a good. Finally, bounded rationality models assert that an increase in the perceived quality of a product is the result of misapplied rational processes. A common thread among the first three approaches is the focus on product *differentiation*. In fact, Bain showed empirically that product differentiation was the single most important factor for firm profitability (5).

The model set forth in this paper also asserts that advertisements can create brand differentiation. However, this model allows for the opposite to occur. For example, we will show that under certain circumstances small producers will be able to free-ride on the advertising of larger firms. This model also shows that differentiation on irrelevant (non-utility bearing) attributes can support differentiation on important attributes. Another major difference is that rather than depending on a single catch-all variable, quality, the model is built on top of the multivariate hedonic pricing structure. Most importantly this model sets forth a concrete mechanism whereby advertising can bring about these effects, allowing us to discuss what distinguishes successful advertising campaigns from unsuccessful campaigns.

1.2.3 Marketing

The importance of associations in the marketing literature is captured by the concept of brand image. Associative memory features prominently in the marketing literature (38; 49; 57). Brand equity is thought to rely not only on the general evaluation of brand quality, but also on attribute specific beliefs about the brand (49). In fact, network models have been used extensively to understand how brand image is affected by a variety of factors, especially brand extensions.

1.2.3.1 Brand and Trademark Dilution

Brand and trademark dilution can be thought of as “weakening a famous brand’s propensity to bring to mind relevant associations” (57). While the marketing literature draws a distinction between trademark dilution, the result of advertisements from another firm, and brand dilution, the result of advertisements from the original firm, both phenomena arise from the same root: weakening associations with a brand.

Many early studies on brand dilutions focused on the effect of a brand extension failure on the general evaluation of a brand (39), and found no effect. However, work that focuses instead on how extensions may effect specific attribute beliefs about a brand (49), show experimentally that extensions that are incongruent with the parent brand on a particular attribute will dilute consumer beliefs about that attribute. These beliefs may, in turn, effect general product evaluation and brand extension success. From this perspective, the success or failure of the brand extension is less important that its congruence with the parent brand.

Keller and Sood assert that brand dilution as a result of extension failure "requires a strong experience with a brand extension – one deemed both diagnostic of and inconsistent with the parent brand experience" (40). In this paper we focus on one of these three factors, the consistency of the extension.[6]

Morrin and Jacoby found experimentally that trademark dilution as measured by product recognition increased when the diluted brand was less familiar, and when the diluting ads were for a dissimilar product category. However, the second-user (purveyor of the diluting ads) benefited more in terms of recall when products were similar. In these cases, the second-user seemed to be able to free-ride off of the associations generated by the first-user, while this is an example of dilution for the first-user, it also counts as an example of spillover for the second-user.

1.2.3.2 Spillover

Spillovers occur whenever the positive associations of one attribute rub off on several products. The best examples of these are once again brand extensions. In a successful brand extension, a firm leverages the positive associations with the parent brand the improve the evaluation of the extension. However, spillovers also result from alliances between brands and from trivial or zero-utility attributes (spillovers from trivial attributes are discussed in the next subsection).

As noted above there are studies showing that brand extensions can potentially harm brand equity, yet it remains an extremely popular strategy for introducing new products. Under the right conditions brand extensions can both benefit from a parent brand and actually increase the parent brand's equity. Morrin showed in a series of experiments that brand extensions can enhance both the recall and recognition of parent brands using congruent brand extensions (eg., Motrin Sinus Pain Formula as extension to Motrin Pain Reliever, Jergens soap as and extension to Jergens lotion) (56). She also shows that this reciprocal effect is correlated with the "fit" between a parent brand and a brand extension.

Other studies also indicate that once again, one of the most important factors for brand extension success in this sense appears to be the the "fit" between the parent brand and its extensions (1; 11). Brand extensions that share more attributes with the parent brand are more likely to be favorably evaluated. Implicitly these studies show that successful brand extensions are able to gain from a consumer's existing associations with the parent brand, i.e., advertising for the parent brand in

[6] Since we are only considering a one-period model in this paper, "strength" of experience will correspond to the multiplicity of advertisements seen for the extension. In a multi-period model, experiences will essentially be weighted by their relevance, e.g., consumption experiences will count more than seeing advertisements. Keller and Sood's point about diagnosticity should also fall under the heading of relevance, but the current model is not well-suited to address when an inconsistent extension will be diagnostic.

the past indirectly benefits the brand extension. Most of the literature focuses on how a generally favorable rating of the parent brand can be transfered to an extensions, but in this paper we take the view that all such spillovers are the result of the transfer of specific attribute association.

The importance of "fit" for the success of brand extensions seems to decrease when the parent brand is already associated with a wide variety of products, e.g. Disney (37; 18). Parent brands with many extensions are likely to be associated with abstract attributes such as quality or luxury. If these attributes are kept consistent, variation among other, concrete attributes appear to make very little difference since the association with the abstract but utility-bearing attribute is continually reinforced.

In addition to spillovers within a brand, from parent brand to extension and vice-versa, there are instances where advertising for one brand will benefit the market for an entire product category. For example, Seldon found empirically that while advertising in the cigarette industry was partially predatory (one firm gains at the expense of another), any given firm's advertising also increased overall demand for cigarettes (67).

Finally, Simonin and Ruth show that brand alliances or co-branded products (for example, a Dell computer with "Intel Inside") will have positive spillover effects for the brands involved. These effects are once again modulated by the consistency of attribute between the two brands and the two product categories. The co-branded products are positively affected by the "brand fit" and "product fit" of allied brands. Positive evaluations of the co-branded products then spillover into more positive evaluations of the individual brands involved.

The common thread in both the dilution and spillover literature is consistency. Consistent associations create spillover effects, while inconsistent associations generate dilution.

1.2.3.3 Trivial Attributes

Trivial attributes are characteristics of a product that have no direct effect on utility. The marketing literature on trivial attributes focuses on attributes that "[appear] valuable but, on closer examination, is irrelevant to creating the implied benefit." (15) However, studies have shown that even when these attributes are revealed to be trivial, they have a significant impact on brand evaluation (12; 15). This effect appears to come from two distinct mechanisms. First, even obviously trivial attributes, such as the color of a soda can (Coke's signature red for example), serve to differentiate a brand and help facilitate positive brand associations. This is particularly useful when considering trademark and trade-dress issues. Second, trivial attributes that appear valuable usually do so because they are already associated with positive relevant attributes, for example, silk and pearls are associated with luster and shine, so they sometimes appear as ingredients in hair products. This second mechanism, is just another form or spillover.

In this paper we consider trivial attributes more broadly to include zero-utility attributes[7] of all kinds. We focus specifically on highly salient physical, but irrelevant, attributes such as color and shape. As noted above, these are the types of attributes that can be legally protected via trademark and are there for useful for building brand-specific associations.

However, associations alone cannot explain all the evidence concerning trivial attributes. Brown and Carpenter show that when a trivial attribute is common, its absence will differentiate a product and serve as a reason for purchase. On the other hand, when the attribute is rare, its presence serves as a reason for purchase. (13) In this study it appears that the trivial attribute's presence or absence is only important as a means to attract attention to products that are different. Unlike the other studies, here the attribute's associations are irrelevant.

1.2.3.4 Network models from marketing

The model developed in this paper is not the first use of an associative network to understand advertising. Over a decade ago, Keller proposed the use of spreading activation networks (described in more detail in the next subsection) as a conceptual framework to understand brand image and equity (38). These networks differ significantly from the networks developed in section 1.3 of this paper, but share two important qualities: nodes in the network represent concepts, and the connections between them represent how these concepts are associated.

Network models of brand extension and dilution currently assume that there is a limited amount of "activation energy" available to retrieve concepts from memory. The introduction of extensions therefore dilutes the amount of activation energy available to concepts associated with the brand's original product. These models have also been applied to trademark dilution by other firm's (57). In these models, another firm using a similar mark a significantly different product class (for example "Hyatt Legal Services") introduces another node representing the brand name, activation energy is then split between these two nodes, diluting the strength of the first user (Hyatt Hotels).

In the model discussed in the paper we instead focus on the effects of brand extensions, and other firms using a similar mark, on the *strength* of the association between the mark and specific characteristics including the product class and favorable product attributes. Another important difference is that while most of the papers think of each brand as having its own network, in our model brand associations are all stored in a single network, allowing us to explore spillover and dilution effect between brands as well as within them.

[7] More formally, a trivial attribute is any attribute such that any two good that are identical on all *perceived* attributes other than the trivial attribute are prefect substitutes

1.2.4 Psychology and Neuroscience

The contribution of this model to the existing literature is the use of a formal model of associative memory to transform exogenously given information and internally generated perceptions of products. Formalizing this transformation requires that we specify a process through which experiences and information interact to create perception. To do this we build on intuitions gained from the cognitive psychology literature and modeling techniques borrowed from the computational neuroscience literature.

1.2.4.1 Cognitive Psychology and Priming

Any model where experience affects perception must include some concept of memory. However, there are many different kinds of memory to consider, we can roughly divide these different types of memory into three broad classes: episodic memory, implicit and semantic memory, and affective memory. The kind of memory we most often think of is "episodic memory." Episodic memory is our memory for specific events: a person's memory of their first day of school years ago or a conversation with a friend a few days ago would both fall into this category. However, for the purposes of this paper we are more concerned with semantic and implicit memory. These two types of memory encapsulate any memory that does not require conscious recollection of when or how the information is acquired. For example, people know their own name, but they don't necessarily remember *learning* their name. Semantic memory specifically includes our memory for concepts and their relationships. These memories are more robust than episodic memory. They survive even when episodic memories are lost because of Alzheimer's disease or severe brain damage (48; 28). A final relevant type of memory is affective memory, which describes our *emotional* associations with various objects or concepts.

The model proposed in this paper takes the view that semantic memory is the most important for understanding perception. Implicit in this view is the assumption that semantic memory influences affective memory as opposed to the converse. While some earlier papers assert that affective processing precedes the semantic processing of stimuli (6), more recent studies indicate that affective processing is highly context sensitive (41). This context sensitivity indicates that the semantic content of a situation is modulating its affective processing. If subjects were making a fast affective assessment before processing the semantic content, this would not occur. Indeed, it has been shown that in some situations semantic processing precedes affective processing (72). Since the affective judgments about a stimulus depend on its semantic content as perceived by the decision maker, we focus on semantic processing and consider affective judgments of a good as *consequences* of the stimulus's perceived semantic content. Referring back to Figure 1.1, this model collapses affective processing into the function taking perceived characteristics to utility in the second stage of processing.

The model exploits the *associative* nature of semantic knowledge: the idea that concepts will potentiate related concepts and interfere with negatively related concepts. Cognitive evidence for this comes largely from the priming literature. In these studies psychologists examine the effect of some stimulus or "prime" on the processing of another target stimulus. Subjects are asked to perform some simple task, such as reading a word, after being exposed to a priming stimulus. When the priming stimulus is related either semantically or affectively to the subject of the subsequent task, people are able to perform the task more quickly. The most common hypothesis to explain this phenomenon is that the priming stimulus somehow "activates" certain pathways in the brain, decreasing response time. Psychologists have been able to show these priming effects across many different conditions. For example, subjects are faster at reading target words after being primed by words (7), music (42) or pictures (72) that are semantically related to the target words. The generality of these results suggest that these associations occur at a very basic level of cognitive processing.

The associative nature of semantic memory was first modeled by Quillian in the 1960s (61; 62). He proposed that semantic memory could be modeled as a network of nodes, each representing a concept. Concepts were related to each other by different sorts of links. Links could be any of five different kinds of links representing different possible relationships between concepts. When one or more of the nodes were activated, the activation would spread along these links to related nodes. These types of networks are often referred to as semantic networks since the nodes are are semantic (they refer to concepts) and the connections describe semantic relationships. Quillian's model was the original "spreading-activation" model of semantic memory. The model was originally conceived to allow computers to simulate language comprehension in humans, but Collins and Loftus adapted it as a more general cognitive theory of semantic memory (16). Since that time, spreading activation in semantic networks has become one of the leading theories to describe priming effects.

Although these semantic networks were developed to explain and model priming effects, they can also be used to model the more general effects of associative memory. In the Dixie Chicks example, the groups statements caused fans to associate them with liberal political views. This association had either negative or positive emotional connotations depending on the consumer. This association in turn changed the consumers' utilities for Dixie Chicks music.

However, Quillian's spreading activation networks are only one of may ways to formalize the these associations. Spreading activation models have some drawbacks for the purposes of modeling evaluation. First, these theories do not formally specify exactly how the connections among concepts are determined. Second, in most of these theories activation can only start at one node at a time. This makes it difficult to model how *dissonant* concepts interact, which is one of the main goals of this paper, as illustrated by the Heinz vinegar example. The problem in that example was that people had conflicting association with the product, and spreading activation models are poorly equipped to model how such a conflict my be cognitively resolved. Finally, these models can become very complicated. Links between concepts can have many types *and* have varying strengths, yet, as noted above, there are no formal specifications as to *how* these links are determined.

The model in this paper retains some aspects of Quillian's semantic networks: nodes still represent concepts, specifically characteristics, and connections still reflect semantic associations. But the *dynamics* of the network are based on Hopfield networks, described in the next subsection. These dynamics are more suited to modeling the way semantic associations *interact* than spreading-activation models. Modeling these interactions is important for understanding situations like the Heinz Vinegar example, where conflicting associations come into play.

1.2.4.2 Hopfield Networks

The semantic networks described in the previous subsection were developed to better understand language comprehension and semantic priming. Coming from a different direction, Hopfield (30) proposed a neural-network model of pattern recognition in 1979. Like Quillian's spreading-activation model, these structures were originally formulated so that they could be implemented by computers. However, unlike Quillian, Hopfield-networks were designed to learn and store patterns. In these networks, the patterns in question are patterns of activation for the network: a vector of elements describing the state of each node in the network. Hopfield networks are able to store specific patterns of activation in the following sense: given a corrupted version of a stored pattern the network is able to recognize and restore the original.

Hopfield networks also differ significantly from Quillian's spreading activation model in their basic structure. In the spreading activation model, stimulating one node will spread activation slowly through the network, but until the node is stimulated the network is essentially neutral. In contrast, any given node in a Hopfield-network is always sending input to its connected nodes.

In these networks, nodes could be in one of two states, 1 or -1 and these nodes will affect connected nodes in *both* states. The state of each node is dynamic: the input it receives from connected nodes can change its state. For example, if a node is currently in the -1 state but is receiving a great deal of positive input from other nodes, its state will flip, going to 1.

Definition 1.2.1 *A node is* stable *if the input received does not change its state, the entire network is stable if each individual node is stable. The patterns of activation that make the network stable are the stable* states *of the network.*

Hopfield's key insight was that these networks will always converge to a stable state. Convergence can be shown using a Lyapunov function to describe the "energy" of a network. This energy function is strictly decreasing under the dynamics of the network, so the minima of the function correspond to stable states of the network. The network "remembers" a pattern of activation if it is a stable state. The only parameters in these networks are the connection strengths between the nodes, so these connections contain the information about all the patterns remembered by the network.

The way these networks are trained is particularly interesting because it corresponds roughly to the way synapse strength is modulated in the brain. The network is given a set of patterns P, its training set, and the connection strength between two neurons is proportional to the number of patterns in which they are both in the same state. This is similar to Hebb's law, which states that neurons that fire together become more strongly connected (29). Conversely, it has been found that the inhibitory connections between neurons grow stronger when their firing is negatively correlated (60). Although much of the work on these sorts of network structures have focused on how to modify connections between nodes making them better and more efficient at recognizing set target patterns, this paper focuses largely on how a given rule for generating connection weights affects the stable states of the network.

After training, the network should remember all the patterns in the training set, but not exclusively. The stable-states of the network will often contain new patterns that were not in its training set. For example, as you will see, if $p \in \mathcal{S}$ is a stable state of the network so is $-p$, even if $-p$ was not in the training set.

Recall that one of the goals of the model is to create a plausible theory of how competing associations interact when we evaluate a good. Using a Hebbian rule to determine connection strengths not only reflects neural processing, it means that the connections between nodes actually reflect their degree of association. Conflicting information can be modeled in these networks as creating unstable nodes. Applying the Hopfield framework to the vinegar example, note that the connection between the node representing food and the one representing cleaners should be highly negative, reflecting their negative association with each other. However, both nodes are activated by "Heinz All-Natural Cleaning Vinegar." The convergence properties of Hopfield networks imply that the network will change to *resolve* this conflict in a definitive way.

The major problem with applying this model directly to economic problems is that it requires complete information. Since the nodes are always in an active state (1 or -1), it is impossible to describe patterns where one or more of the nodes is neutral. If we think of nodes as the characteristics of a good, we must have information about every possible characteristic. Furthermore, every characteristic is relevant since each node is always sending input to its connected nodes. It is easy to envision situations where characteristics are either unknown, or irrelevant to the good in question. The network model described in this paper addresses this problem by modifying the standard Hopfield-network to allow for missing information while retaining its convergence properties.

1.3 Network model for determining "perceived characteristics"

In this section, we introduce a formal, abstract framework to model the way experiences affect perceived characteristics. First, we specify a way to encode both goods and experiences in the same space, allowing us to easily relate experiences to goods. We then formalize the way that experiences change the associations among characteristics. Finally, we describe how these associations determine the perceived characteristics of a good. Table 1.1 briefly summarizes the notation and terms used in this section.

Notation	Description
$\mathcal{S}$	$\{-1,0,1\}^D$: Characteristic space. Both goods and experiences are encoded in this space. This space also describes the possible states of the networks defined in this section.
D	Number of possible characteristics, also the number of nodes in the network
n	Number of goods in the choice set.
$X \in \mathcal{S}$	Initial characteristics — perceived characteristics are derived from this vector.
$V \in \mathcal{S}$	A state of the network
P	A set of elements drawn with replacement from $\mathcal{S}$ representing the *experiences* of the decision maker. In this paper this will be the set of all advertisements seen by the consumer.
E	The space of all possible finite sets P defined as above
$A(P,X) : E \times \mathcal{S} \to \mathcal{S}$	The function taking initial characteristic vector X and experience set P to the perceived characteristics, V.
w_{ij}	The *connection weight* between two nodes i and j. These weights are derived from the experience set P.
w_{ii}	The connection weight of a node to itself. This is exogenously given.
W	A matrix containing all the connection weights
$\alpha \in \mathbb{R}_+^D$	The vector of *activation thresholds* for the nodes of the network
I	The vector of *inputs* to each node. When the network is in state V, $I = W \cdot V$.
$T^i : \mathcal{S} \to \mathcal{S}$	A transition function of the network that updates the ith node of the network. There are D such functions, one corresponding to each characteristic. These functions are defined by the the weight matrix W and the activation thresholds α.

Table 1.1: Notation Table for Section 1.3

1.3.1 Encoding goods and experiences

As discussed above, the hedonic pricing literature already describes a good way to encode multiple goods in the same space. Here we expand on this idea by encoding decision-maker's experiences in characteristic space as well. This type of encoding will lose the detail of an experience, but since this model is primarily concerned with the associations among characteristics rather than the effects of episodic memory, the simplification is appropriate. Both goods and experiences are viewed as vectors describing their component characteristics.

We start with a simple decomposition of goods into discrete characteristics. There is a space $\mathcal{S} = \{-1, 0, 1\}^D$ where D is the number of possible characteristics. Vectors V in this space are interpreted as follows:

$$V_i = \begin{cases} 1 & \textit{V has characteristic i} \\ -1 & \textit{V does not have i} \\ 0 & \textit{i is not salient to V} \end{cases} .$$

We will refer to objects in this space as "patterns," and both experiences and products are represented by these patterns. Suppose that a decision maker is evaluating a pizza and ith dimension represents the characteristic "hot." "Hot" is always salient to pizzas so if V is the decision maker's representation of a pizza, V_i will always be non-zero. If the pizza is fresh out of the oven, $V_i = 1$ since the pizza is, in fact, hot. However, if the pizza is cold or even merely room temperature then $V_i = -1$ since the pizza is not hot *and* that information is salient to the decision maker. If, on the other hand, the decision maker were evaluating a painting, V_i should always be 0 except in extreme circumstances — for example if the painting were on fire. From here on subscripts will denote indices and superscripts will denote distinct elements of $\mathcal{S}$.

Any good or experience is encoded in this way. Possible characteristics in $\mathcal{S}$ might be something like:

(Domino's, Pizza Hut, hot, cheese, pepperoni, peppers, mushrooms, onions).

In this space a "hot Domino's pizza with onions" will be encoded as $(1, 0, 1, 1, 0, 0, 0, 1)$. Notice that the *salience* of certain characteristics allows for some wiggle room in how any good or experience is encoded. For example, if the decision maker's favorite topping is pepperoni, the presence or absence of pepperoni will always be salient to him. In this case the pizza will instead be encoded as $(1, 0, 1, 1, -1, 0, 0, 1)$.

1.3.2 "Perceived characteristics" and experiences

Let E be the space of all possible sets of experiences. That is, E is the space of all finite random samples drawn from $\mathcal{S}$ with replacement. An element $P \in E$ will be a set of patterns $p \in \mathcal{S}$ and any decision maker will have some specific set of experiences described by an element $P \in E$. It is important to note that any patterns "seen" multiple times by the consumer will also appear multiple times in P. For example, if you consumed the same kind of pizza five times, the pattern describing the pizza will occur five times in P.[8]

The patterns in the consumer's experience set P affect the way the consumer processes information when they make purchasing decisions. Specifically, there is some function

$$A : E \times \mathcal{S} \to \mathcal{S}$$

taking an experience set $P \in E$ and some initial pattern $X^i \in \mathcal{S}$ to another pattern $A(P, X^i) \in \mathcal{S}$.

We can now reformulate the decision problem. A decision maker with options described by initial information X^i will choose to maximize his dollar utility given the *perceived* attributes of each good i based on the initial information vectors, X^i. Formally the decision maker chooses

$$argmax_{x \in \mathbb{R}^n}\{u(x, (A(P, X^1), \ldots, A(P, X^n))) - q' \cdot x\},$$

where q is the vector of prices for the goods.

1.3.3 Constructing $A(P, X^i)$ using neural-networks

The function $A(P, X)$ is constructed using a neural-network model of pattern recognition where there are D nodes, each representing a possible characteristic of a product. These networks are extremely similar to Hopfield networks and share many of their properties. The major modification made in order to apply the networks to economic problems is the addition of an extra possible state for each node of the network: the 0, inactive state. It turns out that the networks retain their convergence properties with this modification so that they will still resolve conflicting associations, but it is now possible to represent characteristics as unknown or irrelevant.

The experiences of the consumer are represented by the connections *between* the nodes. These connection weights are numbers in $[-1, 1]$ that represent the degree to which two characteristics are related in the experience of the consumer. For example, for many people the characteristics "red" and "hot" are very related. The hot water knob on sinks is usually labeled in red, as metal is heated it glows red, so people will often have many experiences associating these two characteristics. This association will be reflected by the weight between them, $w_{red,hot}$, which will be positive and relatively close to 1. On the other hand, most people would consider "blue" and "hot" to be negatively related and the weight $w_{blue,hot}$ should be negative and relatively close to -1.

More formally,

$$i \neq j,\ w_{ij} = \frac{\sum_{p \in P} p_i p_j}{|P|}.$$

Recall that elements $p \in P$ are drawn from $\mathcal{S}$, and let p_i refer to the value of the pattern p on

[8] This is important because repetition of patterns will have a significant effect on how information is processed in the model. The more often you see one pattern relative to another, the greater effect it will have on your evaluations.

characteristic i. This formula means that the more often nodes i and j are in the same state, the higher the weight w_{ij}. To see this, fix P so that $|P| = n$. Now add a new pattern, p^* such that $p_i = p_j \neq 0 \Rightarrow p_ip_j = 1$. Adding the new pattern will change the weight:

$$\begin{array}{rcl}
w_{ij} = \dfrac{\sum_{p\in P} p_ip_j}{n} & \underbrace{\longrightarrow}_{add\ p^*} & w^*_{ij} = \dfrac{\sum_{p\in P} p_ip_j + 1}{n+1} \\
\displaystyle\sum_{p\in P} p_ip_j & \leq & n \Rightarrow \\
\dfrac{(n+1)\sum_{p\in P} p_ip_j}{n(n+1)} & \leq & \dfrac{n\sum_{p\in P} p_ip_j + n}{n(n+1)} \Rightarrow \\
w_{ij} & \leq & w^*_{ij}
\end{array}$$

Returning to the example above, any time the decision maker sees “hot” and “red” in conjunction, it will be encoded as a pattern $p \in P$ where $p_{hot} = p_{red} = 1$. The presence of this pattern in the experience set P will contribute to a higher value of $w_{red,hot}$. Conversely, for any $p \in P$ such that the nodes are in opposite states, $p_i = -p_j \neq 0 \Rightarrow p_ip_j = -1$, the pattern will contribute negatively to the weight w_{ij}.

As in the spreading-activation and Hopfield network models discussed in section 1.2.4, the state of each node evolves as a function of the *input* it receives from other nodes. If node i is in state v_i it generates an input of $w_{ij} \cdot v_i$ to node j. This input is aggregated with input from other nodes to create the total input to node j. The total input in turn affects the state of node j.

So far we have only defined the relationship *between* nodes. Each node also has two other parameters that describe qualities of the characteristic itself. First, there is the weight from a node to itself, $w_{ii} \geq 0$, its self-connection. The self-connection can be thought of as how “persistent” the characteristic is since once the node is in a non-zero state, $V_i \neq 0$, this is the magnitude of the input the node sends itself. Since $w_{ii} \geq 0$ this input will always have the same sign as V_i, i.e., the input will be congruent with the current state of the node. Second, each node has an activation threshold α_i. As the name suggests, the higher this threshold, the more input is required to put the node in a non-zero state. In practice this means that the higher the threshold, the harder it will be to associate the characteristic with any other set of characteristics. Both the node's self-connection w_{ii} and activation threshold α_i are exogenous.

Now that we have defined w_{ij} for all pairs of nodes i and j with respect to the experience set P, we can represent the weights of the network as a matrix $W(P)$ where the diagonal elements are the exogenously defined self-connections and the off-diagonal terms are the experience defined connections between nodes. Since $w_{ij} = w_{ji}$ when using the formula above, the matrix will be symmetric. After the weights have been fixed the network is completely characterized by the $D \times D$ matrix $W(P)$ and the vector of activation thresholds α.

Recall that any node of the network can be in one of three states, $\{1, 0, -1\}$. The state of the network is simply the vector containing the state of each node. So the elements $V \in \mathcal{S}$ describe the possible states of the network. We can use the matrix $W(P)$ and the state of the network $V \in \mathcal{S}$ to define the inputs to each node. Specifically, when the network is in state V, $I = W(P) \cdot V$ in $\mathbb{R}^D$ is the vector describing the *inputs* to all of the nodes of the network.

Intuitively, we can link the input to node i, I_i, to the amount of information *consonant* or *dissonant* to the characteristic i. This is because each salient (i.e., non-zero) node of the network in state V send its own input to node i. Think about our "red" and "hot" example again. If $V_{red} = 1$, the red node sends the positive input $w_{red,hot}$ to the hot node, which will tend to activate it. However, suppose that the blue node is also active, $V_{blue} = 1$. This will send *negative* input $w_{blue,hot}$ to the hot node, tending to deactivate it. The actual effect on the hot node depends on (a) the relative magnitudes of $w_{red,hot}$ and $w_{blue,hot}$ and (b) the input received from other nodes in the network.

More generally, we can break the input I_i into the individual summands $w_{ij}V_j$. When the jth node is positively associated with the ith node, $w_{ij} > 0$, then the term $w_{ij}V_j$ will have the same sign as V_j. This means that if $V_j = 1$, the jth node will send positive input to the ith node, helping to "activate" the ith node (change its state to 1). Conversely, if $V_j = -1$ the jth nodes send negative input to the ith node and tend to change its state to -1. The total input, I_i simply aggregates these effects.

We can now use the input vector I to describe the dynamics of these networks. The inputs to each node will change their state; however, in this model we do not change the states of all the nodes at once. Instead we *update* the states of nodes one at a time based on the updated node's input I_i. We describe this updating process using transition functions. Since there are D nodes, we define D transition functions $T^i : \mathcal{S} \to \mathcal{S}$, one for each node. Each T^i depends on the input to the relevant node, I_i, and that node's activation threshold α_i:

$$T^i_i(V) = \begin{cases} -1, & I_i = \sum_j w_{ij}V_j \quad < \quad -\alpha_i \\ 0, & \qquad\qquad\qquad -\alpha_i \quad \leq \quad I_i \quad \leq \quad \alpha_i \\ 1, & \qquad\qquad\qquad\qquad\qquad\qquad\quad \alpha_i \quad < \quad I_i \end{cases}$$

$$T^i_j(V) = V_j, j \neq i.$$

To use these networks we need to consider time on a macro as well as micro level. On the macro-level we consider how experiences build our associations. Formally, in this model, this means looking at what patterns make up the set P and constructing the weight matrix W. P can be built up over months or years. For example, in the pizza example, P would consist of all the experiences you have had with both Domino's and Pizza Hut including consumption experiences, advertisements, and word of mouth descriptions.

However, when we are actually making a decision, $A(P, X)$ acts almost instantaneously, giving us our perceived characteristics, $V = A(P, X)$, of the good. We model this by taking P and fixing the weights of the network so we can consider time on the micro-level. In each micro-time period here we choose some $i \leq D$ and apply T^i to the network.[9] If a node i is stable under its updating function T^i then $T^i_i(V) = V_i$ and the node is said to be stable. We repeat this procedure until we reach a pattern $V \in \mathcal{S}$ so that $T^i(V) = V$ for all i — in other words, a pattern where all the nodes are individually stable. Patterns that satisfy this condition are called stable patterns.

Claim 1.3.1 *A network with transition functions T^i, as described above, will converge to a stable pattern in finite time.*

We show convergence by defining an energy function over all possible states of the network.

$$E(V) = -\sum_i w_{ii} V_i(t) - \frac{1}{2} \sum_{i \neq j} w_{ij} V_i(t) V_j(t) + \sum_i V_i(t) V_i(t) \alpha_i$$

The energy of the network is a measure of how unstable the network is. It takes into account whether the input to a node I_i is congruent with its state V_i. To see this, examine the terms $-w_{ij} V_i V_j$. These will be positive whenever w_{ij} does not have the same sign as $V_i V_j$. For example, suppose i and j are usually positively (negatively) related yielding a positive (negative) weight w_{ij}. If these nodes are in opposite (the same) state, the term $-w_{ij} V_i V_j$ will be positive, so the incongruence of the states of these nodes with their connection weight increases the energy of the network. Returning to the "red," "hot" example yet again, if $V_{red} = 1$ and $V_{hot} = -1$ the term $-w_{ij} V_i V_j > 0$. That is to say, these nodes contribute positively to the energy of the network and are a source of instability in the network. These incongruities may exist in a stable state of the network only if there are stabilizing inputs to both i and j coming from other nodes. The other terms in this energy function correct for the influence of the self-connection and activation thresholds of the nodes.

In fact we can show that the energy of the network strictly decreases under all of the transition functions T^i unless the node i is itself stable (implying that $T^i(V) = V$). We can use this fact to show that the network must always converge to a stable state and that stable states of the network correspond to the local minima of the network's energy function (see Appendix A, for details).

The convergence property of these networks divides the space of possible states into "basins of attraction" around each stable pattern. Think of a surface with lots of divots. When you drop a ball anywhere on the surface, it will come to rest at a low point. Notice that the order in which we update the nodes of the network is random, so although it is assured that the network will converge,

[9] This sort of updating rule is referred to in the neural-network literature as an *asynchronous* updating rule, since nodes are updated one at a time. Networks with *synchronous* updating rules, i.e., all nodes are updated simultaneously, do not necessarily converge to a single stable state and are biologically less plausible than those with asynchronous rules.

there are cases where it could converge to any one of a number of different stable states. We define the "basin of attraction" of a stable state as the set of network states that will converge to the stable state with certainty. Any state of the network that falls outside of these basins might possibly go to any one of at least two different stable states. We say that states that might converge to one of a number of stables states are on the "edge" between the basins of attraction for those states.

When we start inside a basin of attraction, the order in which we update the nodes of the network does not effect the final stable state. However, when we are on the edge between basins, the pattern the network converges to is path dependent, as illustrated in the next subsection. These stable patterns serve as restrictions over what bundles of characteristics a decision maker will ever actually evaluate given a fixed experience set P. When the weights have been set, no matter what initial information the consumer receives, they will process the information as one of these patterns. Notice that, using the transition function above, the trivial state of the network, $V = \vec{0}$ is always stable regardless of the weight matrix W. So the trivial state will always be in the set of patterns that a decision maker may evaluate.

We can now define a (possibly stochastic) function $\mathcal{T}(W, X)$ which is the stable state resulting from an initial state of X and a weight matrix W. The resulting stable state, $\mathcal{T}(W, X)$, will be deterministic for any X within the basins of attraction and stochastic for patterns on the edges. The initial information X of products will be within the basins of attraction for the cases examined in this paper.[10] So we use this function to define the perceived characteristics of X:

$$A(P, X) = \mathcal{T}(W(P), X)$$

That is, the perceived characteristics of a good, given initial information X, are the characteristics described by the stable state that the network converges to when it starts in state X, and the network in question is defined by the experiences of the decision maker, the set P.

1.3.4 Example network

To gain a better understanding of these network dynamics, consider the dynamics of a simple three node network as shown in Figure 1.2 and defined by

$$W = \begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \end{pmatrix} \ \& \ \alpha = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

That is, nodes have no self-connection, all activation thresholds are 0, and connection weights

[10] The path dependence of these networks for borderline initial states could be interesting for cases when the order of characteristic considerations can be controlled. Specifically, cases where information is fed sequentially to a decision maker, as it is on a resume.

are as labeled in Figure 1.2. When you start with $X = (1, 1, 0)$, as shown on the left, the pattern immediately stabilizes to $T^3(X) = \mathcal{T}(X) = (1, 1, -1)$. Since the first and second nodes are highly associated with each other, they are both stable. However, updating the third node leads to $V = (1, 1, -1)$, which is a stable state of the network. To see this, notice that, when the network is in state V, the vector of inputs is:

$$I = W \cdot \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ -2 \end{pmatrix}$$

The inputs to nodes 1 and 2 are both $2 > 0$, so $T^1(V) = T^2(V) = V$, and the input to node three is $-2 < 0$ so $T^3(V) = V$. Since all the nodes are individually stable, the network is stable.

On the other hand, starting with $V = (1, 0, 1)$ the outcome will depend on which node you update first. Nodes 1 and 3 are negatively related, but they are both active. This constitutes conflicting, dissonant information, and the dynamics of the network will resolve the pattern so that it agrees with one of two pieces of information, $V_1 = 1$ or $V_3 = 1$.[11]

The second node is stable in the initial configuration, but if we start by updating the first node, the network will converge to $(-1, -1, 1)$, and if we start by updating the third node, the network converges to $(1, 1, -1)$. In other words if we start with conflicting information, the network will stochastically go to a stable state that is consistent with part of the initial pattern. In this case, the first node, V_1, is consistent with stable state $(1, 1, -1)$, while the third node V_3 is consistent with stable state $(-1, -1, 1)$, and the network goes to one of these two stable states with equal probability.

In this section we have laid out a very general formal framework for assigning perceived characteristics to some initial set of characteristics X. So far there have been no constraints on either the exogenous parameters (α, the w_{ii}s) or the initial information X. In the next section we address what these parameters should be and what we mean by initial information with respect to the specific economic problem of advertising. We use the framework developed in this section to discuss the purpose of advertisements: they serve to associate abstract or difficult-to-assess traits of a product with physical, easy-to-asses traits of a product. For example, associating status with a particular brand of shoe or car. In this setting the meaning of the initial vector X becomes clear. X is the vector describing the easily assessed characteristics of a product while $A(P, X)$ describes how these initial characteristics are perceived, including the derived abstract characteristics discussed in the introduction.

[11] For now we take this order of update to be random. However, as we will discuss in the extensions section, this order might be determined by a number of factors including endogenous interest (self-guided consideration of a particular attribute), or exogenous information (you may receive information sequentially instead of all at once, like reading a list of product attributes in a catalog).

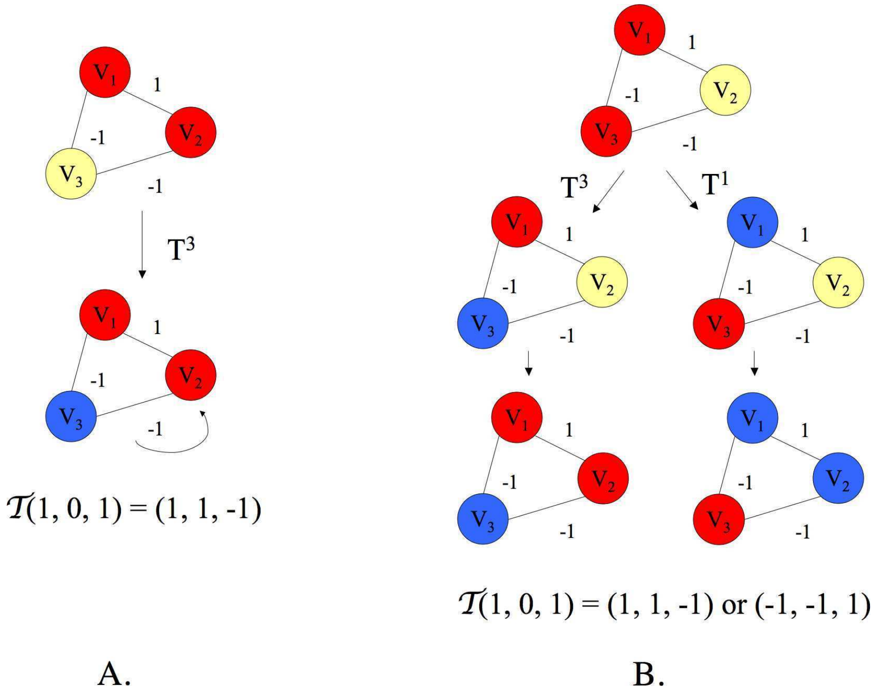


Figure 1.2: The dynamics of a simple three-node network with no self-connection and activation thresholds of 0. Connections are labeled with their corresponding weights. The color of each node denotes its state: Red = 1, Yellow = 0, Blue = -1.
A) On the left the network starts at an initial state $X = (1, 1, 0)$ that is within the basin of attraction for the stable state $(1, 1, -1)$ so there is only one possible outcome. Specifically, the initial state is stable under the two transition functions T^1 and T^2, while T^3 takes it to the stable state $(1, 1, -1)$.
B) On the right the network starts at state $(1, 0, 1)$ which is on the "edge" between the attractor basins for $(1, 1, -1)$ and $(-1, -1, 1)$. This initial state is stable under T^2, but T^1 and T^3 take it into the attractor basins for $(-1, -1, 1)$ and $(1, 1, -1)$, respectively. The value of $\mathcal{T}(W, (1, 0, 1))$ is therefore stochastic, and depends on the order in which the transition functions are applied.

Associative networks have been proposed as framework for understanding brand image in the marketing literature (38; 57). But these have all followed the spreading-activation threshold. The main differences between the model developed in this paper and this framework are that (a) the structure of the network is formalized, including an account of how connections between nodes are formed, (b) concepts can only be either positively or negatively associated with each other, and (c) in this model there is a natural stopping point for the spread of activation, namely a stable state of the network.

1.4 Consumer utility, advertisements, and the decision problem

In this section we relate the networks described in section 1.3 to advertisements and the hedonic pricing model. We consider the simple one period model where consumers are exposed to a set of advertisements, P, at the beginning of the period and then make their consumption decisions at the end. We make the simplifying assumptions that P is composed entirely of advertisements.

First, we relate parameters and variables in the model such as the activation thresholds α_i, and the experience set P, to concrete ideas about product attributes and advertising. We then incorporate our network model into a hedonic pricing model. We define what an advertisement is and how firms change the experience set P. And finally, we consider the implications of this new model with respect to findings in the marketing literature.

1.4.1 Characteristic types

To apply this neural-network framework to advertising, we will place some more structure on the space $\mathcal{S}$ by categorizing characteristics into three classes: search, experience, and social. The differences among these categories are then described with respect to the characteristic level parameters of self-connection w_{ii}, and activation threshold α_i.

Following Nelson's distinction between search and experience goods, I define "*search characteristics*" as all those characteristics that are easily evaluated before consumption, and "*experience characteristics*" as those requiring consumption of the good to assess (59). Finally, I add one more category: "*social characteristics.*" These are the more abstract features of a product that rely not only on physically verifiable traits, but social norms, personal experience, and context.

1.4.1.1 Search characteristics

All easily assessed characteristics of a product are included in the search group. These include traits traditionally thought of as attributes, such as color or size, but also characteristics such as the product type itself (is it a hat or a car?), and the product's brand (is it a Honda or a Toyota?).

Notice that all of these are tangible features of the product. They are both easily defined and easily assessed. Because of this, the nodes corresponding to search characteristics in the network will have high degrees of self-connection. The main property of these nodes is that once they are in a non-zero state they are extremely difficult to change.

Mathematically, we express this by saying that the self-connection weight $w_{ii} \gg 0$, but we also require that $w_{ii} \gg \alpha_i$. This states that once the node is in a non-zero state it will create self-reinforcing input with absolute value w_{ii}. Since this input is in excess of the activation threshold,

α_i, of the node, the node is capable of staying in either non-zero state without any other input. Furthermore, since $w_{ii} \gg \alpha_i$, the node with persist in a given non-zero state even in the face of a large amount of input in the opposite direction. For example, if a node is in state 1, it would take a large amount of negative input from other nodes to change its state to 0 or -1. This is an important trait to have since we generally do not want the perception of a search attribute to change under the dynamics of the network.[12]

1.4.1.2 Experience characteristics

Experience characteristics require the consumer to use the product to ascertain their value. One example would be the effectiveness of a stain remover. Looking at the product in the store, you cannot directly assess how well it will work; you can only observe the product's brand, packaging, price, and sometimes its ingredients. A first-time user has very little information about the actual cleaning power of the product, but repeat users will learn to associate the search attributes of the product, like its brand, with its cleaning power.

With respect to the neural-network model, experience attributes are characterized by low levels of self-connection and moderate activation thresholds. This means that simply including a non-zero value for an experience characteristic in a consumer's initial information will not be enough to sustain it in its non-zero state. However, in the presence of supporting associated characteristics it will be included in the final pattern. Mathematically, we express this by saying $\alpha_i > w_{ii}$.

1.4.1.3 Social characteristics

Unlike search and experience characteristics, social characteristics are not objectively measurable. Instead, these characteristics are socially defined and subject to change over time. The quintessential example of a social attribute is fashion. People consider a piece of clothing to be fashionable based on both its search characteristics and some set of conditions taken as social convention.

These are the characteristics that are most sensitive to culture and time period. As time passes, the physical trappings of fashion or wealth may change, but the core idea remains the same. Traditional models of advertising have trouble addressing these characteristics since they are abstract and subjective concepts. But since these are exactly those attributes that are most sensitive to culture and a consumer's experiences, they will be a focus of our model.

The most important feature of a social characteristic in this model is that it has absolutely no self-connection. Unlike search attributes, these nodes can change state easily depending on the states

[12] There are, however, notable exceptions where contradictory sensory information will actually change a persons perception of what we would generally think of as a completely objective trait. One famous example is the McGurk effect: when people are given auditory input of a man saying "ba," but visual input of the man saying "ga," they actually *hear* a sound like "da." This is particularly interesting, since knowing about the effect does not allow a person to correct his perception (52).

of associated nodes. For these characteristics $w_{ii} = 0$. Restrictions on the parameter values for each node are summarized in Table 1.2.

Type	Self Connection (w_{ii})	Activation Threshold (α_i)	Reason
Search	$w_{ii} \gg 0$	$w_{ii} \gg \alpha_i$	Search characteristics are very persistent: if $V_i \neq 0$, $T_i^i(V) = V_i$ unless there is a large amount of information to the contrary, i.e., $\lvert I_i \rvert \gg 0$ and $sign(I_i) \neq sign(V_i)$.
Experience	Intermediate	$\alpha_i > w_{ii}$	These require the activation of associated characteristics to become and remain active. So, $\lvert I_i - V_i w_{ii} \rvert > \alpha_i$.
Social	None, $w_{ii} = 0$	Moderate to high	The state of a social characteristic is completely dependent on input from other characteristics since I_i depends only on input from other nodes.

Table 1.2: The three categories of characteristics and their associated parameters, self-connection, the weight of a node to itself, and activation threshold, the net input required to obtain a non-zero state. These correspond intuitively to the "persistence" of a characteristic, its tendency to stay in some non-zero state once there, and the "difficulty" of a characteristic, how much information is required to make the characteristic salient.

In a world where there are only search attributes and these search attributes are always salient (i.e. non-zero), this model is equivalent to Lancaster's hedonic pricing model. Since search attributes are highly self-connected they will tend to be invariant regardless of the states of other nodes in the network, making the function A the identity function. In fact, that model has been most notably applied to markets like the housing market, where all relevant aspects of a good are search attributes (square feet, number of bedrooms, etc.) and in these models prices are always determined using *all* of these search attributes.

However, it is rarely the case that consumers only care about the search attributes of a good. We assert in this paper that advertisements help develop associations between the search attributes of a product (especially the brand) and the experience and social attributes of a product. The neural-network model laid out in section 1.3 can be used to describe how advertisements and other experiences can actually affect a consumer's perception of a product and thereby affect the consumer's evaluation of the product. This is especially true of the consumer's perception of the social characteristics of a good. For example, teenagers in the inner-city think about athletic shoes as status symbols in a way that most suburban teens do not. The goal of many advertisements is to change the way consumers think about products, often shifting from a simple utilitarian evaluation to one where the *social* value of the product is relevant.

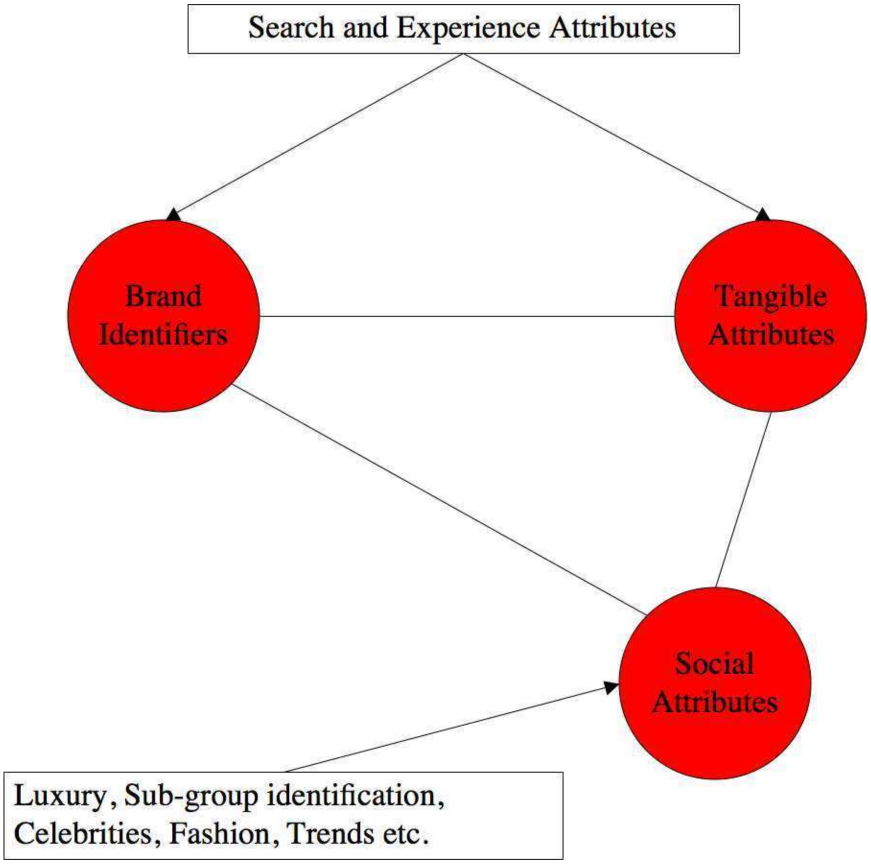

Search and Experience Attributes
Brand Identifiers
Tangible Attributes
Social Attributes
Luxury, Sub-group identification, Celebrities, Fashion, Trends etc.

Figure 1.3: The goal of marketing is to associate social value with tangible physical attributes, preferably those associated with their own brand. Advertisements help to strengthen the links between the brand identifiers of a product and a set of desirable tangible and social characteristics. Because of the connectivity of the network the tangible attributes associated with the brand reinforce the brand's connection to desired social attributes.

1.4.2 Products and utility

Given the characteristics space $\mathcal{S}$, there exist *in potentia* 3^D possible products. In full generality, decision makers have a utility function defined over all possible bundles of products in this space, i.e., there exists some $\bar{u} : \mathbb{R}^{3^D} \to \mathbb{R}$. However, in any given market we only care about utility over products that are actually offered. The function $A(P, X)$ developed in section 1.3.3 will tell us which products are actually offered.

We define the set $X_i \in \mathcal{S}$ as the vectors describing the salient *search* characteristics of each of the offered products. We qualify that the search characteristics must be *salient* because for any product there will be a number of search characteristics that the consumer will find irrelevant. For example, while the smell of an article of clothing is a search attribute, it usually will not be salient, so it should not be included in X. However, for borderline search characteristics, like the texture of clothing, a retailer may be able to exogenously draw attention to or away from it by manipulating

$X^i \in \mathcal{S}$	Vector describing the *salient* search attributes of product i.
P	Advertisements seen by the consumer.
$V^i = A(P, X^i)$	Perceived characteristics of product i.
$\vec{V} = \left(V^1 \cdots V^n\right)$	A $D \times n$ matrix where column i is the vector describing the perceived characteristics of product i.
$\bar{u} : \mathbb{R}^{3^D} \to \mathbb{R}$	Utility function over bundles of *all potential* products.
$u : R^n \to \mathbb{R}\|\vec{V}$	Utility function over bundler of *existing products* as they are perceived.
$\vec{\gamma} = (\gamma_1, \ldots, \gamma_D)$	Characteristic utility coefficients used in parametric utility function example.

Table 1.3: Notation Table for Section 1.4.2

the lighting in the store.[13] Salience may also be determined by an endogenous manipulation of attention. If a customer goes into a store with a particular search characteristic in mind, it will be salient regardless of the manipulations by a retailer.

Given X^i, we let $V^i = A(P, X^i)$ be the perceived attributes of these products, notice that these are *fixed* and *independent* of the other products offered. We apply A to each product separately and then fix the perceived attributes for tractability.[14]

Once the perceived characteristics of each product have been fixed we can get a reduced form of of the utility function: $u : \mathbb{R}^n \to \mathbb{R}|\vec{V}$ where n is the number of products in the market and $\vec{V}$ are the n elements of $\mathcal{S}$ describing their perceived characteristics. This function is simply $\hat{u}$ taken on the n-dimensional subspace corresponding to $\vec{V}$. To see how this function might actually work we will define a parametric example:

1. Let $\vec{V}$ be the $D \times n$ matrix holding the perceived characteristics of each product in the columns.

2. Let $\vec{\gamma}$ be a vector in $\mathbb{R}^D$. This is a vector of attribute utility coefficients.

3. Define the reduced form utility function as $u(x) = M\vec{\gamma}' \cdot g\left(\vec{V} \cdot x\right)$ where M is money and $g(y) = \begin{pmatrix} y_i^{1/3} \\ \vdots \\ y_D^{1/3} \end{pmatrix}$. Notice that for functions of this form the overall utility function is simply $\bar{u}(x) = u(x)|\vec{V}_A$ where $\vec{V}_A$ is the D^D matrix containing all potential products.

Notice that this function is separable w.r.t. product attributes rather than the products themselves. The vector $\vec{V} \cdot x \in \mathbb{R}^D$ is the aggregate amount of each attribute. The coefficients γ_i denote the strength of preference for each attribute. If $\gamma_i > 0$, utility is increasing in the amount of attribute

[13] In some ways this is very similar to the problem of bringing an advertisement to the attention of a consumer. Visual tricks like lighting, or even shelf placement, are relevant; so are harder-to-define factors, like context or even the order in which choices are evaluated. I talk about the latter in section 1.7.4.

[14] Allowing the attributes of different products in a bundle to interact with each other raises the question of what really constitutes consumption. In order for these interactions to make sense the products must not only by purchased at the same time, but consumed at the same time spatially and temporally. We discuss the possibilities of introducing these interactions to understand product complementarities in section 1.7.4.

i, i.e., the consumer "likes" attribute i. On the other hand, when $\gamma_i < 0$ the consumer dislikes the attribute (they like the salient lack of the attribute). We will continue using this type of utility function to illustrate the effects of advertising for the rest of this section.

1.4.3 Trivial Attributes and Substitution

Products with similar attributes should be substitutes. In fact we assert that products that differ only on inconsequential characteristics will be perfect substitutes. We define an attribute as *trivial* if products that differ only on those attributes are always perfect substitutes.

Definition 1.4.1 *An attribute i is* trivial *if for any V, $V^{'}$ in $\mathcal{S}$ such that $V_j = V_j^{'}$ for all $j \neq i$, the product denoted by V and $V^{'}$ are substitutes with respect to $\bar{u}(x)$.*

In our parametric utility function an attribute i is trivial if $\gamma_i = 0$.

A trivial attribute has no direct impact on utility. If the available products V^i were set exogenously, they would be, as their name implies, trivial. However, in this framework the V^i are endogenously generated based on the set of advertisements seen by the consumer P. Even though the trivial attribute does not actually enter into the utility function, its value in X^i can change V^i and thereby indirectly affect a product's utility.

We assume in this paper that firm and brand identifiers are always trivial attributes. Under this assumption, a brand's equity is completely a function of the strength of its association with utility bearing attributes. Given the network structure from the previous section, we can now formally address the affect of different types of advertisements on the connections between nodes and how those in turn affect brand equity.

We now turn our attention to the the experience set P.

Definition 1.4.2 *A pattern or set of patterns p^* is added to the experience set P_0, $P_1 = P_0 \uplus p^*$, when you concatenate the set P with p^*. For example, if $P_0 = \{x, y, z, x, z, x\}$ and we add the pattern x to P_0 we get $P_1 = P_0 \uplus x = \{x, y, z, x, z, x, x\}$. Notice that even though $x \in P_0$, adding x still changes the set by increasing the* number *of times it appears in the new experience set P_1.*

Definition 1.4.3 *An advertisement is a pattern $A \in \mathcal{S}$ that firms pay some cost, c_A, to add to the experience set P for some or all consumers.*

We make the very strong assumption in this paper that firms are actually able to develop advertisements that will be encoded perfectly. This means that A) consumers must pay attention to the ads, and B) they must interpret them as intended. While these are both difficult to ensure, their variability is beyond the scope of this paper.[15]

[15] Ways to make people pay attention to ads range from using bright colors in print ads to using humor. Psychology and neuroscience studies of attention have thus far looked at phenomena like "pop-out" effects. In these situations,

Now we can look at the consumer's decision problem once again. Recall that there is some utility function, depending on $\vec{V}$ taking bundles of products to dollar utility. Let $q \in \mathbb{R}^n$ be the vector of prices for the goods in the market. The decision maker's problem is now simply to choose $x \in \mathbb{R}^n$

$$x = argmax\{u(x|\vec{V}) - x \cdot q\}$$

Example: Given our parametric utility function we can look at a numerical example of how this might work. Consider a two product, two non-trivial attribute world with attribute coefficients $\gamma_1 = 1$, $\gamma_2 = .5$. The first non-trivial attribute is the product category Z, while the second is a positive social characteristic Y. There are also two nodes in the network representing two different brands. These are assumed to be trivial so $\gamma_{F^1} = \gamma_{F^2} = 0$. We can ignore the trivial nodes and consider a utility function $\bar{u} : \mathbb{R}^9 \to \mathbb{R}$ over all meaningful combinations of the non-trivial attributes $\vec{V}$ (we leave the values for the brand identifiers in the third and fourth rows blank, all nine patterns associated with the initial pair are perfect substitutes so we pool them into a single column).

$$\begin{array}{cccccccccccc} & & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 & \\ \vec{V}^* = & \left(\begin{array}{c} \\ \\ \\ \\ \end{array}\right. & \begin{array}{c} -1 \\ -1 \\ - \\ - \end{array} & \begin{array}{c} -1 \\ 0 \\ - \\ - \end{array} & \begin{array}{c} -1 \\ 1 \\ - \\ - \end{array} & \begin{array}{c} 0 \\ -1 \\ - \\ - \end{array} & \begin{array}{c} 0 \\ 0 \\ - \\ - \end{array} & \begin{array}{c} 0 \\ 1 \\ - \\ - \end{array} & \begin{array}{c} 1 \\ -1 \\ - \\ - \end{array} & \begin{array}{c} 1 \\ 0 \\ - \\ - \end{array} & \begin{array}{c} 1 \\ 1 \\ - \\ - \end{array} & \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right) \begin{array}{l} attribute\ 1 \\ attribute\ 2 \\ F^1 \\ F^2 \end{array} \end{array}$$

Given these we know that

$$\begin{aligned} \bar{u}(x) &= \gamma_1 g((x_7 + x_8 + x_9) - (x_1 + x_2 + x_3)) + \gamma_2 g((x_3 + x_6 + x_9) - (x_1 + x_4 + x_7)) \\ &= ((x_7 + x_8 + x_9) - (x_1 + x_2 + x_3))^{1/3} + .5((x_3 + x_6 + x_9) - (x_1 + x_4 + x_7))^{1/3} \end{aligned}$$

However, at most two of these products will actually exist. First consider the case where neither firm advertises, i.e., neither firm is associated with the social attribute Y. This means that, ignoring the trivial brand identifiers, $V^1 = V^2 = (1, 0)$ and $u(x_1, x_2) = (x_1 + x_2)^{1/3}$. Notice that since both firms' products are identical except for the brand identifiers, they are perfect substitutes, so the consumer will have straight indifference curves.

Now consider the case where F^1 advertises the pattern $(1, 1, 1, 0)$, so $P = \{(1, 1, 1, 0)\}$. Let $\alpha_Y \in [1, 2)$, so the activation threshold for Y is between 1 and 2. Given the initial search characteristics for F^1, $(1, 0, 1, 0)$, the input to Y will now be above the threshold and $A(P, (1, 0, 1, 0)) = (1, 1, 1, 0)$. On the other hand $A(P, (1, 0, 0, 1)) = (1, 0, 0, 1)$, i.e., firm 2's product is not associated with Y. The

elements that are very different from the background immediately draw people's attention. For example, an "O" will stand out in a field of "T"s since it contains curves, but an "L" might be hard to find. These processes explain the use of color in ads, but the effect of emotional responses like humor are more poorly understood.

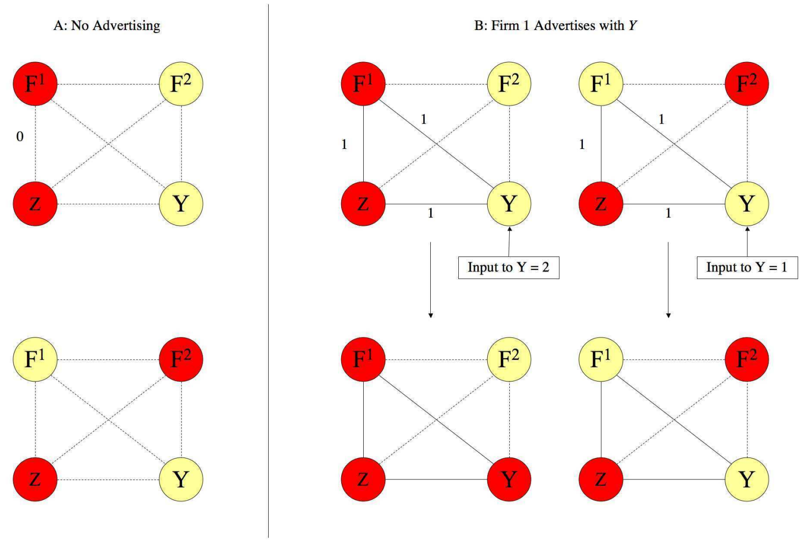


Figure 1.4: A) Network when neither firm advertises. Dotted lines represent connections with 0-weight. B) Network when firm 1 advertises. Notice that the search characteristics F^1 and Z are sufficient to activate Y, but Z alone is not.

new utility function is $u'(x_1, x_2) = (x_1 + x_2)^{1/3} + x_1^{1/3}$. This will raise the absolute and marginal utility of extra units of x_1 and give us convex indifference curves. The networks for both cases and the related indifference curves are shown in Figure 1.4. Indifference curves are shown in Figure 1.5.

Y is a social attribute, so it is culturally determined rather than objectively measurable. In this one period model, the consumption of the advertisement is actually creating extra utility. This fits with the complementary view of advertising discussed in section 1.2.2.[16]

1.5 Common effects at the consumer level

1.5.1 Dilution

Recall that the weight between nodes i and j is

$$w_{ij} = \frac{\sum_{p \in P} p_i p_j}{|P|}$$

The possible sources of brand dilution discussed in section 1.2.3 all involved pairing a brand identifier B with an incongruent product. Incongruence simply means that the original brand differs from the

[16] In a multi-period model where the the experience set P includes consumption experiences as well as advertisements the connection strengths will reflect a mixture of informative experiences and advertisements.

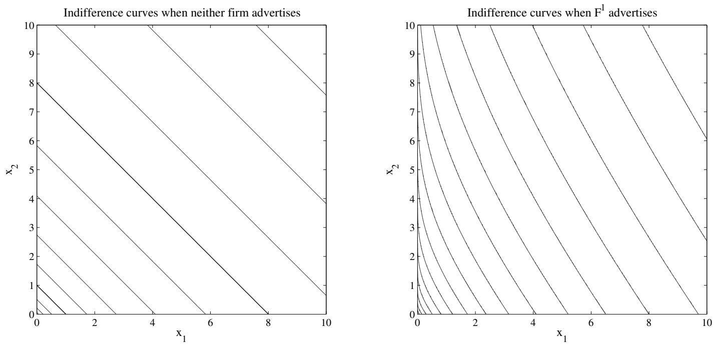


Figure 1.5: Indifference curves when neither firm advertises (left), and when firm 1 advertises (right).

other product on one or more attributes. We can quantify dilution on this level easily by considering the marginal effect of an incongruous advertisement on the connection strength between the brand name and the differing attribute. Without loss of generality, assume that the original brand, B is always paired with some attribute Y. The new product advertising the brand name B with either $Y = 0$ or $Y = -1$. Let P^0 be the set of advertisements for the original brand, and let a be an advertisement for the new product. Assume that for all other patterns in the original experience set, $p \in P$, $p \notin P^0$, the state of the brand node is 0

$$\Delta w_{BY} = \frac{\sum_{p \in P \uplus a} p_B p_Y}{|P| + 1} - \frac{\sum_{p \in P} p_B p_Y}{|P|} = \frac{|P^0| + a_B a_Y}{|P| + 1} - \frac{|P^0|}{|P|}$$

Notice that this effect is larger when

1. $a_Y = -1$ instead of 0
2. The size of P^0 is small (when there hasn't been as much advertising for the original brand)

The second implication is consistent with Morrin and Jacoby's finding that trademark dilution is more pronounced for less familiar (read less advertised) brands. Since the size of the effect on the connection weight is inversely proportional to the number of advertisements for the original brand, products with fewer advertisements are more vulnerable to trademark dilution.

But brand image is not simply a brand's association with one attribute, but an entire pattern of attributes. Dilution will occur for every attribute of the original product that differs from the new product. In addition, the association between the brand name and the attribute isn't the only one being diluted, connections among the attributes themselves will also be diluted. Since multiple connections are diluted, the aggregate effect of the new ad on the *input* to any given node is proportional to the number of characteristics where the two products differ. This, possibly large, decrease in input to the node may change the state of attribute Y in V^B, the perceived characteristics of the original product.

Assuming that the state of Y in V^B goes from 1 to 0, the change in marginal utility of product i given some fixed endowment will be:

$$\Delta \frac{\partial u}{\partial x_i} = -\frac{\gamma_Y}{3}(x_i + e_Y)^{-2/3} < 0$$

where e_Y is the total amount of attribute Y from the initial endowment and $\gamma_Y > 0$ is the attribute utility coefficient for Y.

This decrease in marginal utility is possible for every diluted positive association and is cumulative. If several positive attributes, $Y_1, \ldots, Y_k$ are no longer active in V^B the change in utility is

$$\Delta \frac{\partial u}{\partial x_i} = -\sum_{j=1}^{k} \frac{\gamma_{Y_j}}{3}(x_i + e_{Y_j})^{-2/3}$$

Since this decrease occurs at all levels of consumption, the consumer's demand for product i is decreasing in the number of diluted positive attributes.[17]

These implications are supported by the findings in marketing that trademark and brand dilution is worse when the products sharing the brand or trademark are more dissimilar.

Notice that connections strengths are only diluted for the changed attributes. The connections between any consistent set of attributes will be reinforced by ads for both the parent brand and the extension. This is consistent with Dacin and Smith's finding that brands associated with a variety of products, but consistent on quality, are less vulnerable to dilution. Here, every advertisement for the brand reinforces connections between the brand identifiers and quality, so the dilution of the other connections is less relevant.

1.5.2 Spillover

In the context of this model, a spillover is simply the activation of a positive attribute due to advertising for a different product. Every advertisement affects *all* of the network connections, so advertisements associating a particular brand and product with a social attribute Y will strengthen the relationship between the brand and Y, but it will also strengthen the association between the product category itself. This product category spillover is consistent with Seldon's finding that cigarette advertisements positively affected the demand curve for the entire industry as well as the specific demand for the advertising brand.

Applying the model to brand extensions again, notice that the marginal effect of advertisements for both the parent brand and the extensions is to *increase* the connection strength between the brand name and positive shared attributes. Using the same notation used in the section on dilution

[17] Conversely, the consumer's demand for product i is increasing in the number of diluted negative attributes.

notice that the marginal effect of a single advertisement for the extension on the association between the brand and a positive social attribute Y is once again

$$\Delta w_{BY} = \frac{\sum_{p \in P \uplus a} p_B p_Y}{|P|+1} - \frac{\sum_{p \in P} p_B p_Y}{|P|} = \frac{|P^0| + a_B a_Y}{|P|+1} - \frac{|P^0|}{|P|}$$

Only in this case $a_B a_Y = 1$ so the effect is positive. Notice that once again, this effect is decreasing in the size of P^0, predicting that the reciprocal spillover effect from the extension to the parent brand is decreasing in the size of P^0. This finding is consistent with studies that show that reciprocal spillover is is stronger for *less* familiar brands. (56)

Using exactly the same math as above, we see that advertising that enforces the connection between the brand and a positive attribute will weakly increase the marginal utility of both the original and extended products at all consumption levels. In general, the model agrees with findings in the marketing literature that consistency between a brand and its extension will facilitate spillover effects in both directions.

The model also helps to explain spillover effects from trivial attributes when their irrelevance to the product is known. Supporting the model prediction that mere association is enough to affect evaluation is Gilovich's finding that sportswriters' evaluations of college football players are affected by irrelevant information such as the fact that the college player was from the same hometown as a famous NFL player. (24) Here, the trivial attribute (hometown) establishes an indirect association with the athletic success of the professional player, i.e., the association of the hometown with success spills over to college player.

1.5.3 Trivial attributes, trademarks, and negative associations

As noted above, advertisements by a firm can create spillovers for other firms by increasing the connection between a product *itself* and some positive social or experience attribute. This is exactly what many generic products rely on. If you walk around a drugstore you'll notice that most of the store-brand items are packaged to look as similar to name-brand products as possible to facilitate free-riding off the associations of the search attributes of those brands.

There are legal remedies for this situation in the form of trademark and trade dress laws. Trademark laws protect actual brand names and logos, while trade dress applies more generally to distinctive packaging and design features. Interestingly, the standard for anything to be a trademark in the legal sense is that it must have "acquired secondary meaning"(Lanham Act 1946). In this model, "acquiring secondary meaning" means that the mark is highly associated with some positive social or experience attribute, i.e. $w_{Mi} \gg 0$ for some positive non-search attribute i.

Trademark and trade dress cases focus on two phenomena. The first is infringement, which addresses the free-riding issue: these laws attempt to prevent firms from benefiting from the secondary meaning attached to competitor brands. The second is brand dilution.

We discussed the problem of brand dilution with respect to brand extension in the previous section. In these cases firms dilute their own associations by creating incongruent brand extensions. There are also cases where competing or unrelated firms are accused of brand dilution for using the distinguishing physical elements of a famous brand for their products. According to the Federal Trademark Dilution Act (1995), dilution is defined as "the lessening of the capacity of a famous mark to identify and distinguish goods or services."

Trademark dilution is often divided into two categories. The first is "blurring," where a trademark becomes (a) less-distinctly associated with the particular products advertising under the mark and (b) associated with unrelated products. This dilution is no different that the brand dilution created by a firm's own brand extension. The second is "tarnishing," where a trademark is associated with *negative* characteristics.

A famous case of tarnishing was a case where Victoria's Secret sued a store called "Victor's Little Secret" which sold adult toys. Victor's Little Secret was capitalizing on the existing associations between Victoria's Secret and sexuality; however, Victoria's Secret claimed that the adult store created new, undesirable associations with its trademark.

In this model, the secondary meaning of a search attribute, or set of search attributes, defining a brand is simply the set of social and experience attributes in $A(P, X^B)$, where X^B is a vector representing those search attributes. In addition a company is not allowed to trademark any "functional" aspect of a product. In other words, characteristics protected under trademark or trade dress are of no direct utility to the population by themselves: they must be trivial. However, their ability to activate other intangible attributes allows them to add utility to a brand.

Considering trivial attributes as search characteristics in this network model allows for any number of effects, but we will concentrate on the way these attributes can prevent free-riding.

The simple existence of extra search attributes would not prevent free-riding by itself. The key to preventing free-riding is to provide a search attribute that is always salient. In other words, its absence from a competitor's product must be noteworthy. For example attributes like color are actually pre-attentive, meaning that they are processed without any sort of directed attention. These are perfect for the purpose of distinguishing one brand from the rest, as are other visual features such as shape. In fact the Supreme Court held that a firm could actually protect the color of their product under trade dress law in *Qualitex vs. Jacobson, 1995.*[18]

Apple's iPod campaign provides a more familiar example of the importance of trade dress. The iPod's trademark white ear-buds are a salient centerpiece of most iPod advertisements. The ads even place the white ear buds against black backgrounds to draw extra attention to them. When

[18] In this case, Qualitex was a major producer of dry-cleaning pads. Qualitex pads were a distinctive green-gold color that Jacobson, a competing firm, started using.

the iPod started selling players with different colors they made sure to keep the white ear buds since the iPod's distinctiveness is what makes it more desirable than any of the many other mp3 players on the market.

Because of its pre-attentive nature, Al and Laura Ries includes the "Law of Color" as one of their "The 22 Immutable Laws of Branding" (64). They address the importance of picking a color to represent your brand that contrast with those of your competitors helping to create differentiation.

> Pepsi-Cola made a poor choice. It picked red and blue as the brand's colors. Red to symbolize cola and blue to differentiate the brand from Coca-Cola. For years Pepsi has struggled with a less-than-ideal response to Coke's color strategy.
>
> Be honest. In your mind's eye, doesn't the world seem to be awash in Coca-Cola signs? An isn't it hard to picture many Pepsi-Cola signs? Pepsi is out there, but the lack of a unique differentiating color tends to make Pepsi invisible in a sea of Coca-Cola red.
>
> Recently Pepsi-Cola has seen the light, or rather the color. It is doing what it should have done more than 50 years ago. Make the brand's color the opposite of its major competitor's color.
>
> Pepsi-Cola is going blue.— Al and Laura Ries, "The 22 Immutable Laws of Branding"

In this case color is definitely a trivial attribute. While one can argue that consumers have a taste for color in a portable music player, the color of your drink container will not directly effect your utility from the drink. However, Ries asserts that color choice is one of the most important choices these companies make. Why? Because color is a salient trait that absolutely everyone notices and because of this it is one of the most important factor in creating brand differentiation.

This is another place where this network model differs significantly from traditional hedonic pricing. Zero-utility attributes are truly trivial in the traditional model since they should have no *direct* effect on consumer evaluations. However, using networks, we can explore how inherently valueless attributes can still have significant effects on consumer decisions.

All these examples suggest the importance of trivial attributes, especially attention-grabbing physical characteristics, for product differentiation. We address this problem by using a neural-network that includes a node to represent the trivial attribute.

Consider the simple two firm network in Figure 1.6. Let F^2 be a generic brand free-riding on the advertising of F^1. Until now we have ignored the possibility that a brand can actually be activated by other characteristics such as the product type, but in the case of a generic this could actually be a strategy by itself, especially if the generic wishes to benefit from "hard" social characteristics with high activation thresholds.[19]

[19] Famous examples of brands activated by the product type itself are Kleenex, Xerox, and Coke.

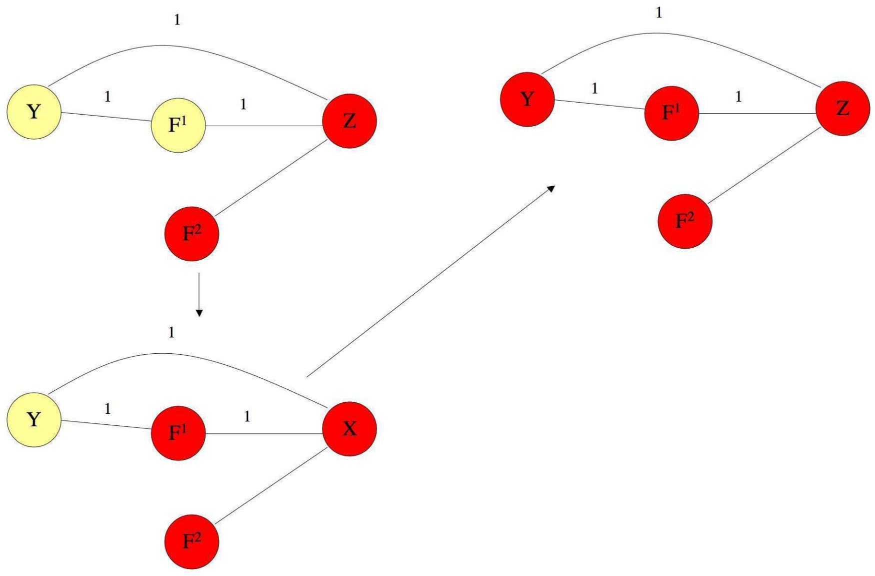


Figure 1.6: The generic marketing strategy: use the associations of a “name”-brand to increase the decision utility of their product. The generic product does not advertise itself at all, instead the input from the product node Z to the name-brand node F^1 is sufficient to activate F^1 which in turn activates the utility bearing characteristic Y.

In this network, the product type Z actually activates the name-brand node F^1. Once F^1 is active, the associated positive characteristic Y can be activated by input from both the product type Z and the name-brand F^1. This way the generic product F^2 benefits from the reflected glow generated by advertising for F^1. A real world example of this strategy is shown in Figure 1.7. The generic cereal brand “Crisp Crunch” is not only the same kind of cereal as “Cap’n Crunch,” having the same ingredients, shape and texture, but goes further by using a cartoon picture of a captain (a trivial attribute) to gain association with the name-brand cereal “Cap’n Crunch.”

Then name-brand F_1 can prevent this free-riding and retain its market-share by introducing a highly salient trivial attribute T, such as a distinctive color, that it can legally protect. The crucial point is that the trivial attribute T is in state -1 when a consumer assesses the competitor brand F_2. Because of this, it will actually *inhibit* the activation of the social attribute Y in $A(P, X_2)$. This model implies that elements of trademarks and trade dress not only help a product by helping to activate non-search characteristics, but actually can prevent free-riding by competitor firms.[20] The new network, with T included is shown in Figure 1.8.

While tarnishing a trademark creates interesting legal implications, advertising can create negative associations with other aspects of a product. An interesting, though not legally actionable example of tarnishing, was a 2005 Dasani commercial where a bear disparages the virtues of natural

[20] Examining the cereal boxes above, even though “Crisp Crunch” uses a cartoon captain on their box, they are not allowed to use the distinctive, trademarked, image of the “Cap’n Crunch” captain. The differences between the two captains may be enough to actually highlight the distinctiveness of the name-brand cereal depending on the consumer.

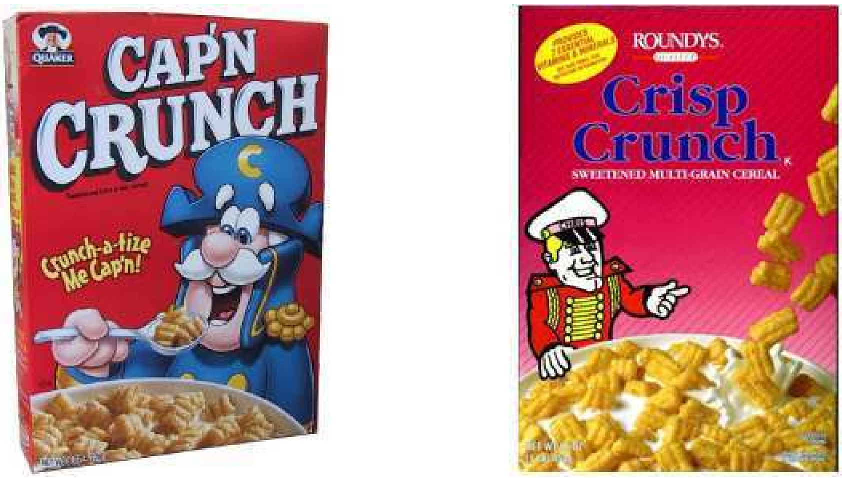


Figure 1.7: Generic brand attempts to associate themselves with their name-brand competitors can be extremely obvious. In this example "Crisp Crunch" actually uses a cartoon captain in uniform to elicit associations with "Cap'n Crunch."

spring water.

> The whole natural mountain spring thing is fine. But I was just in there, and you do not want any of that. There are salmon in there. And do you know what they're doing? Spawning, in the mountain stream. Way too natural for me. —Dasani Advertisement, 2005

This advertisement in many ways better illustrates tarnishing than the legal case above: the firm employs a national advertising campaign to attack its competitors by associating their most salient feature, natural — usually associated with purity — with its exact opposite – impurity. Applying the neural network model, the advertisement places the node corresponding to purity in the -1 state while placing "natural" in the 1 state. This serves the purpose of quickly diluting the positive association between "natural" and "pure" and possibly, given heavy exposure to the ads, create a negative association between the two characteristics.

Another example of advertising that "tarnishes" the attributes of a competitor is the recent Jack in the Box campaign indirectly attacking Angus burgers, offered by competing fast food chains Carl's Jr, McDonald's, and Burger King. Even though the Angus burger are not protected by intellectual property laws, the owners of Carl's Jr. are suing Jack in the Box.

> CKE Restaurants Inc. sued Jack In The Box in U.S. District Court on Friday over an ad in which executives laugh hysterically at the word "Angus" and another where the chain's pingpong ball-headed mascot, Jack, is asked to point to a diagram of a cow and show where Angus meat comes from.

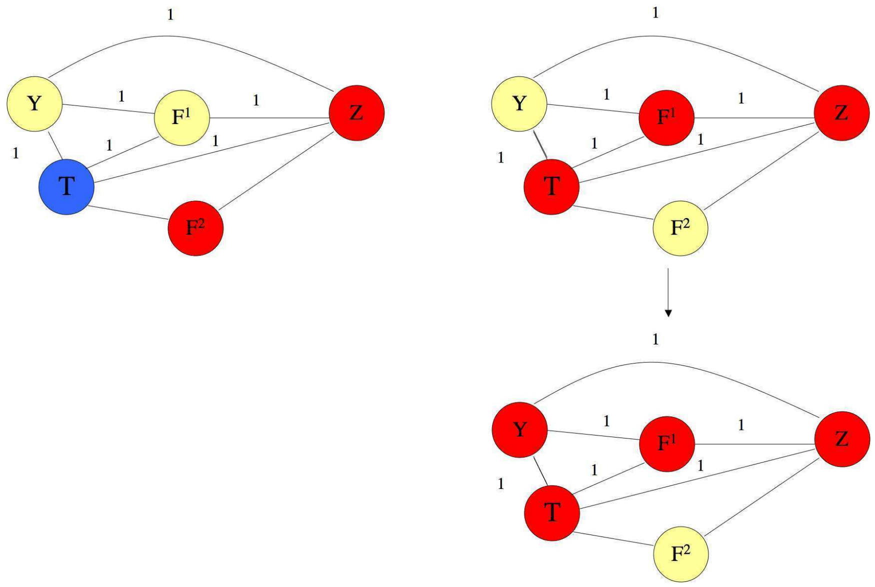


Figure 1.8: When a consumer evaluates F_s under this network structure, the salient absence of the trivial attribute T inhibits the activation of Y.

"I'd rather not," the pointy-nosed Jack replies. (Associated Press, May 29, 2007) (14)

Like the Dasani commercial, this advertisement aims to change the positive consumer associations with Angus beef by associating it with something negative. The major difference is that, in this case, rather than reversing an existing positive association, the new commercial simply seeks to associate competitor products with a previously unrelated, negative attribute.

In summary, advertisements can serve to differentiate a product from its competitors by creating not only positive associations to desirable social and experience attributes, but also by using negative associations. This is a particularly important role for the distinctive and salient but non-functional aspects of a product that can be legally protected under trademark laws. These trivial attributes allow firms to protect their advertising investment by not only uniquely identifying their brand as opposed to others (a simple name does that) but by creating negative input to positive advertised traits in their salient absence.

1.6 Implications for Firms

Given the implied effects of advertising on consumers' evaluations of products, what might be the optimal advertising strategy for a firm in different situations? In this section we will consider two common situations. First, what are the optimal advertising levels in a simple oligopoly with one social attribute. And second, what is the optimal advertising strategy for a firm that wishes to sell products that are differentiated over a social attribute.

As in the previous section we consider a one period model. At the beginning of the period the firms simultaneously choose their advertising and price strategies. We then assume that all consumers see all advertisements, so the set P is the same for each consumer. At the end of the period consumers make their purchasing decisions and firms produce and sell the demanded quantities (Bertrand oligopoly).

For the purposes of this paper we will constrain the set of allowable advertisements. First, firms cannot refer to other firms' brands in their advertisements (no negative advertising).[21] Second, we will only consider advertisements that include a product category Z since the vast majority of advertisements actually seen are for a specific brand and product category.[22]

1.6.1 Application to a market with a single good

Now we are ready to work out a simple example of this model. Assume that there is only one product type Z and one possible, universally positive, social attribute Y. This means that there is a "basic" version of the product, which would consist of Z alone, and a preferred version that would have $Z = 1$ and $Y = 1$, which we will refer to as the high-Y good from here on. The space of possible network states is $\mathcal{S} = (F^1, \ldots, F^n, Z, Y)$, where the F^is represent the firms in the market. In this example, each firm has only one brand, so we can treat brands and firms as synonymous.

Given the assumptions above we can simplify the consumer's utility function substantially. There are essentially only three possibly goods: high-Y, basic, and negative-Y.

$$\hat{u} : \mathbb{R}^3 \to \mathbb{R}$$

takes bundles of these goods to their dollar utility.

We consider the implications of this model in monopoly and oligopoly situations. The structure of the oligopoly market changes radically depending on the value of the activation threshold for Y, α_Y. For low values of α_Y firms are able to free-ride on the advertising of their competitors. On the other hand, for high values of α_Y, not only is free-riding impossible, but only a limited number of firms will be able to sell the high-Y good. Heuristically this means that for less common, "harder" Y attributes (like luxury) only a few brands will actually be able to participate in the high-Y market of a good, regardless of advertising levels. We call the number of firms that may actually participate in the high-Y market the *capacity* of the high-Y market and show that this capacity is decreasing

[21] Negative advertising could be introduced by simply allowing the brand nodes referring to competitor brands to be in the -1 state. This would allow a firm to create a *negative* association between the competitor and themselves, along with negative associations between the competitor and the favorable characteristics they attempt to associate with themselves. In reality, negative advertising can occur but is impeded by the presence of trademark laws.

[22] The actual reason for including product categories in advertisements has more to do with brand recall than brand evaluation. In the current model, recall of all competitors is assumed. In this case it is not obviously useful to associate a brand with a particular product category. Indeed as shown in this section, it may allow competitors to free-ride on the firm's advertisements. But in reality, brands generally must be associated with the product category to be included in the consideration set. We discuss this aspect of advertising in the extensions section of this paper.

the social characteristic's activation threshold α_Y.

We make the following simplifying assumptions:

1. Firms produce a good Z. We assume that consumers always consider all firms in the market. Think of the situation where you are in a store with all the brands of a product in front of you, so brand recall is not relevant.

2. Recall that advertisements are *patterns*, $p \in \mathcal{S}$, that firms pay to add to the set P. We make the further simplifying assumption that P only contains advertisements. In addition, we assume that every consumer sees every advertisement, so P is the same for all consumers. Let A^i be the set of advertisements purchased by firm F^i. This assumption states that for all consumers

$$P = \biguplus_i A^i.$$

3. The activation threshold for the social characteristic Y, α_Y, is constant over the population. This along with (2) mean that perceptions are shared across consumers: $A(P, X)$ is the same for all consumers for any starting information X.

4. The only firm that is initially salient to the consumer is the firm being evaluated, i.e., $X^i_{F^j} = 0$ for all $j \neq i$.

5. All firms have the same cost curve $c(x)$, and there is a fixed cost per advertisement, c_A.

6. Y is universally positive for all consumers. Specifically, the marginal utility for the high-Y good is always greater than the utility for the basic good, and the marginal utility for a saliently *low*-Y good is less than zero:

$$u_1(x) > u_2(x) \;\&\; u_3(x) < 0 \;\forall\; x \in \mathbb{R}^3_+$$

This corresponds to the attribute utility vectors satisfying $\gamma_Y > \gamma_Z > 0$ for all consumers if we use utility functions of the form described in section 1.4.2.

In this model, the initial information, X^i, for each good will be the vector that is 1 at the indices representing the manufacturing firm F^i, and the product type Z, and 0 everywhere else. Notice that the only tangible difference between products from different firms is the firm identifier itself, and the only difference relevant to utility must come from the consumer perceptions of Y.

Since the experience set P is the same for all consumers, perceptions will be the same across consumers. Formally, for each i $A(P, X^i)$ is the same for every consumer.

Let $V^i = A(P, X^i)$. The demand any given firm, F^i, faces is given by a function

$$D_i(\vec{q}, \vec{V})$$

Symbol	Description
Z	Product type, e.g. running shoe, car, or hat
Y	Universally positive social characteristic, e.g. status
n	Number of firms selling Z
F^i	Firm i (in this section firms and brands are synonymous)
A^i	Advertisements purchased by F^i
N_i	Number of advertisements purchased by F^i: $N_i = \lvert A^i \rvert$
X^i	Search characteristics of the the product sold by firm i. In this section X^i is 1 at Z and the firm index F^i, and 0 everywhere else.
$\vec{q}$	Vector of prices charged. q_i is the price charged by F^i.
$\vec{V}$	Perceived characteristics of the products for each firm, $V^i = A(X^i, P)$
$D_i(\vec{q}, \vec{V})$	The demand for goods produced by F^i given prices $\vec{q}$ and perceived characteristics $(\vec{V})$ for all firms
$c(x)$	The cost function for production of Z. All firms face the same cost curve.
c_A	The cost of one advertisement

Table 1.4: Notation Table for Section 1.6

where $\vec{q}$ are the prices charged by each firm in the market and $\vec{V}$ are the perceived characteristics of each product in the market. These perceived attributes are in turn a function of the set of advertisements purchased by each firm, so we can write $\vec{V}$ as a function of all the firms' advertising campaigns.

$$\vec{V} = V(P) = V\left(\biguplus_j A^j\right)$$

A firm F^i's problem is then to choose a price q_i and advertising strategy, A^i to maximize the profit function,

$$\pi^i(q_i, A^i) = q_i \cdot D_i\left(\vec{q}, V\left(\biguplus_j A^j\right)\right) - c(D_i) - c_A \cdot |A^i|$$

given the other firms' prices and advertising strategies.

1.6.1.1 Monopoly

Assume that there is only one firm, F^1. This will lead to the three-node network shown in Figure 1.9, and the state space is $\mathcal{S} = (F^1, Z, Y)$. Recall that since Z and F^1 are search characteristics, they are both highly self-connected (see Table 1.2), so that once these nodes are activated they will stay activated. This means that there are only three possible $A(P, X^1)$:

$$A(P, X^1) = \begin{cases} (1,1,-1) \\ (1,1,0) \\ or\ (1,1,1) \end{cases}$$

Furthermore, in this example note that $A(P, X^1)$ is deterministic because the states of Z and F^1 are stable and, since Y is a social attribute, $w_{YY} = 0$ and the state of Y is completely dependent on input from Z and F^1, i.e., the network will always be stable under transition functions T^Z and T^{F^1}, so $A(P, X^1) = T^Y(X^1)$.

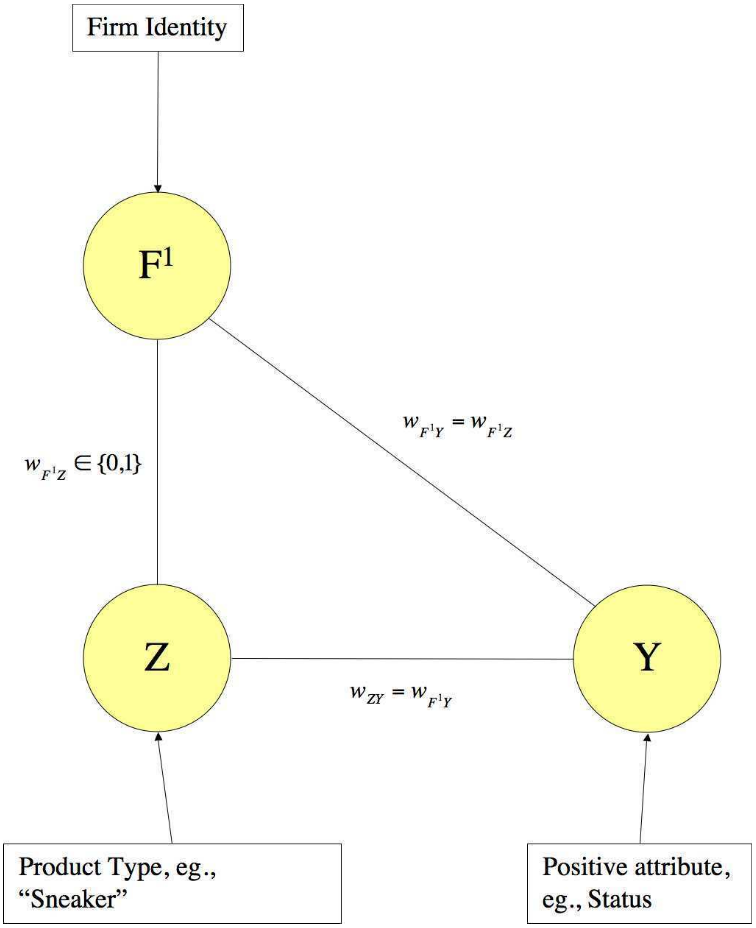


Figure 1.9: The monopoly case with one product Z and one social attribute Y

Claim 1.6.1 *F^1 will purchase exactly one advertisement $\mathcal{A} = (1,1,1)$ if and only if $c_A < q^h_m D(q^h, (1,1,1)) - c(D(q^h_m, (1,1,1))) - q^l_m D(q^l_m, (1,1,0)) + c(D(q^l_m, (1,1,0)))$ and $\alpha_Y < 2$, where q^h_m and q^l_m maximize $qD(q, (1,1,1)) - c(D(q, (1,1,1)))$ and $qD(q, (1,1,0)) - c(D(q, (1,1,0)))$, respectively.*

Proof:

In a monopoly setting, the firm will never advertise $(1,1,-1)$ because $u_3(x) < 0$ for all consumers, and therefore $D(q,(1,1,-1)) = 0$ for all $q > 0$. Since recall is ensured in the current example, they will never advertise $(1,1,0)$ either. This is because $X^1 = (1,1,0)$ and $A(P,(1,1,0)) = (1,1,0)$ even when $P = \emptyset$.[23] If the firm purchases no advertisements, $w_{F^1Y} = w_{ZY} = 0$, they will face demand curve $D(q,(1,1,0))$ and produce $D(q^l_m,(1,1,0))$.

If they do advertise, they only advertise $A = (1,1,1)$, which means that for all $p \in P$, $p = (1,1,1)$. In the monopoly case they will only need to purchase one advertisement since

$$w_{ZY} = \frac{\sum_{p \in P} p_Z p_Y}{|P|} = \frac{|P|}{|P|} = 1$$

as long as $|P| \geq 1$, and the same is true for w_{F^1Y}.

When the firm advertises, the matrix of weights is

$$W = \begin{pmatrix} w_{F^1F^1} & 1 & 1 \\ 1 & w_{ZZ} & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

so the inputs to each node when the network is in state X^1 are

$$I^T = (1,1,0) \cdot W = (w_{F^1F^1} + 1,\ w_{ZZ} + 1,\ 2).$$

F^1 and Z are search characteristics so $(1,1,0)$ is stable under the transition functions T^{F^1} and T^Z. However, if $\alpha_Y < 2$, then $T^3(1,1,0) = (1,1,1)$, which is then stable under all T^i, implying that $A(P,X^1) = (1,1,1)$. The firm will face $D(q,(1,1,1))$ and produce $D(q^h_m,(1,1,1))$.

Therefore, the firm will either choose not to advertise at all, yielding a maximum profit of

$$q^l_m D(q^l_m,(1,1,0)) - c(D(q^l_m,(1,1,0))),$$

or purchase exactly one advertisement $a = (1,1,1)$, yielding a maximum profit of

$$q^h_m D(q^h_m,(1,1,1)) - c(D(q^h_m,(1,1,1))) - c_A.$$

This means that the firm will advertise if and only if the gains of facing the higher demand curve outweigh the cost of one advertisement c_A.

Finally if $\alpha_Y \geq 2$ there is no way for the network to produce enough input to surpass the activation threshold for Y; therefore, there is no incentive for the firm to advertise. This proves the

[23] Stability of the pattern $(1,1,0)$ is ensured because F^1 and Z are both search attributes.

claim. *Q.E.D.*

1.6.1.2 Oligopoly

What happens if we add other firms $(F_2, \ldots F_n)$? Now, instead of the three-node network from the monopoly example, we have a $(n{+}2)$-node network with the state space described as $(F_1, \ldots, F_n, Z, Y)$. This network is illustrated in Figure 1.10.

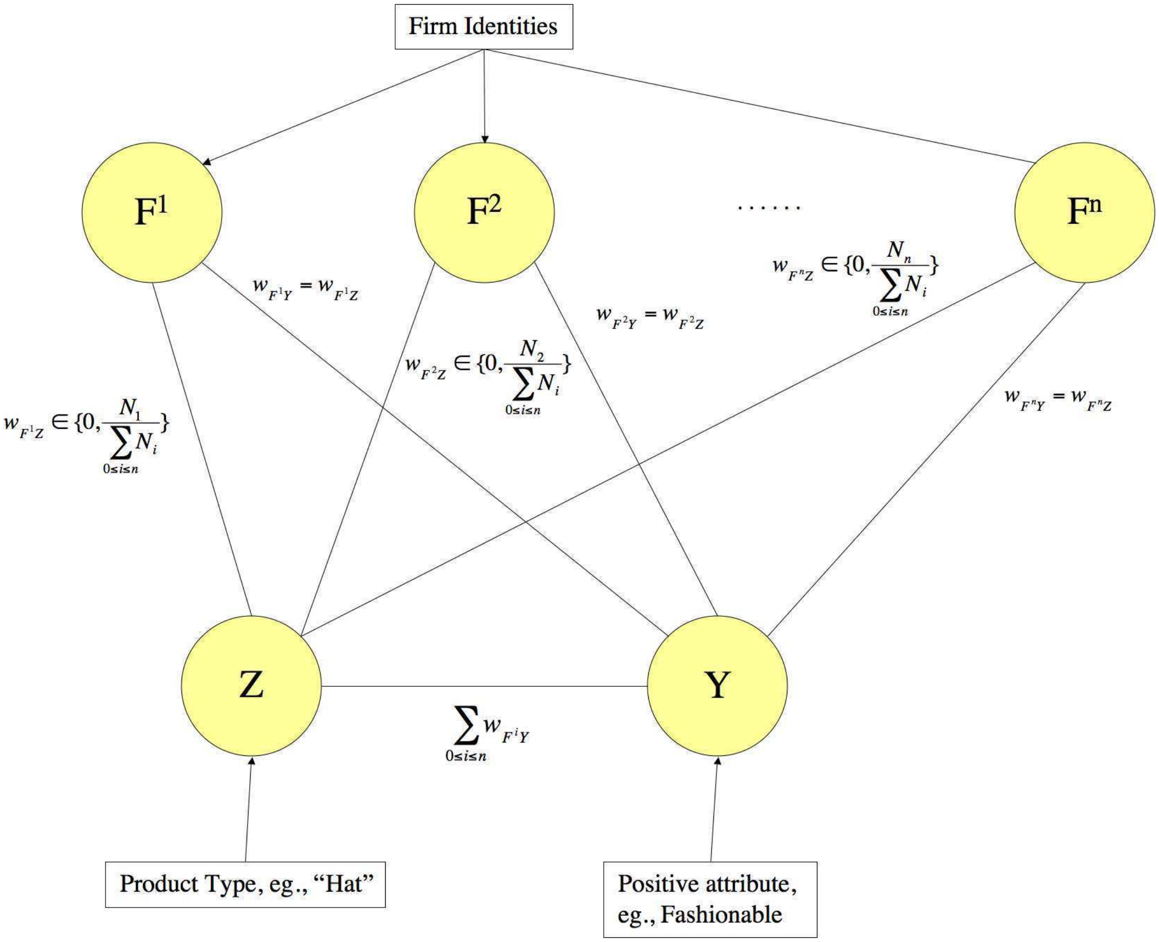

Firm Identities
F1
F2
......
Fn
$w_{F^1Y} = w_{F^1Z}$
$w_{F^2Y} = w_{F^2Z}$
$w_{F^nZ} \in \{0, \frac{N_n}{\sum_{0 \le i \le n} N_i}\}$
$w_{F^2Z} \in \{0, \frac{N_2}{\sum_{0 \le i \le n} N_i}\}$
$w_{F^nY} = w_{F^nZ}$
$w_{F^1Z} \in \{0, \frac{N_1}{\sum_{0 \le i \le n} N_i}\}$
Z
Y
$\sum_{0 \le i \le n} w_{F^iY}$
Product Type, eg., "Hat"
Positive attribute, eg., Fashionable

Figure 1.10: $(n+2)$-node network for the example. The top row of nodes serve as the firm identifiers, Z is the node representing the product type, and Y is a desirable social attribute that the firms can advertise.

The first thing to realize is that whenever any firm advertises, they produce a positive externality. Specifically, they strengthen the weight w_{ZY} between the product type itself and the characteristic Y, leading to possible spillover effects for the entire market. This allows for an interesting phenomenon: firms may be able to free-ride off the advertising of other firms to raise the utility of their products. This phenomenon can actually be observed in fashion. For example, Von Dutch trucker hats gained a degree of popularity when stars like Britney Spears started wearing them, however, trucker hats with other logos became popular as well (see Figure 1.11). The trendiness was associated with the style of hat rather than the Von Dutch brand specifically.[24]

Figure 1.11: The rise of the trucker hat: Celebrities started wearing trucker hats made by Von Dutch in 2004. On the left, Madonna, Benicio DelToro, Leonardo DiCaprio and Lindsay Lohan are all shown wearing the brand. However, the trendiness of the product did not remain specific to the Von Dutch brand and other manufacturers began producing trucker hats with a variety of logos for the hipster population, one example is shown on the right. (Images were taken from www.newfaces.com and www.ilovejungle.co.uk respectively.)

Whether this free-riding can occur depends on the activation threshold for the social attribute Y, which is $\alpha_Y = \alpha$. The lower the threshold α, the easier it is for activation to "spread" to the attribute, i.e., the easier it is the create an association between the social attribute and any search attribute.

First case: $\alpha < 1$. As in the monopoly case, firms will never choose an advertisement with $Y < 1$, so the only advertisements that will ever occur will have $Z = Y = 1$. From this we can see that if any firm decides to advertise, the weight $w_{ZY} = 1$. This ensures that the input to Y, $I_Y \geq 1$ whenever $Z = 1$ and that $A(P, (\ldots, 1, 0)) = (\ldots, 1, 1)$ regardless of the states of the firm nodes F^i. In this case product from all the firms are perfect substitutes, differing only in price.

Now this turns into a simple public-good situation. Let $q = \min(\vec{q})$ be the minimum price charged for the product. If any firm decides to advertise, we denote the *total* demand for the good as $D_h(q) = D(q, (\ldots, 1, 1))$. Otherwise, we denote the total demand as $D_l(q) = D(q, (\ldots, 1, 0))$. This raises the question, when will firms advertise at all?

Claim 1.6.2 *If* $\alpha < 1$*, in any pure strategy equilibrium at most one firm will advertise. If advertising*

[24] Even though the "coolness" of the Von Dutch trucker hat spread to other brands, the Von Dutch brand did gain significant associations with young celebrities and all they entail. Von Dutch was able to use its association with these celebrities to begin marketing *other* products under the same brand: t-shirts, jeans, etc.

does occur only one advertisement will be purchased.

Proof:

Assume that this is not the case. That means that there exists some equilibrium where at least two firms advertise. As in the monopoly case, the only advertisements purchased will be of the form $a = (\ldots, 1, 1)$. Since P consists entirely of advertisements, this implies that the weight between the product type, Z, and the positive social characteristic, Y, is one: $w_{ZY} = 1$.

When $w_{ZY} = 1$, we know that the input to the social characteristic will be at least one, $I_Z \geq 1$, when the network is in any state where the Z node is active, $V_Z = 1$. Since $\alpha < 1$ by assumption and $I_Z \geq 1$, we have $I_Z > \alpha \Rightarrow T^Y_Y(V) = 1$ for all V such that $V_Z = 1$.

Consider one of the advertising firms: if it were to stop advertising the connection strength, w_{ZY} would still be 1. So, even without its advertisements the firm would face the exact same demand curve, but decrease its costs. This implies that advertising is not the firm's best response to the strategies of its competitors, therefore the market is out of equilibrium giving us a a contradiction. This proves that only one firm will advertise. The advertising level will be 1 for a similar reason: for any higher level of advertising the firm could do better by dropping to $|A^i| = 1$. *Q.E.D.*

We do not consider mixed strategy equilibria in this case since, although they may exist in this one-period model, they will not generally be reflective of real world advertising situations.[25]

There are two necessary conditions for advertising to occur:

1. At the equilibrium price, firms must not have any incentive to price under all other firms to monopolize the market.
2. The advertising firm must not have an incentive to stop advertising.

We can use these conditions to establish upper and lower bounds on the price in an advertising equilibrium. First we define four useful functions: the monopoly profit functions for the high and low-Y goods, and the oligopoly profit functions for these goods.

$$\begin{aligned}
\pi^h_m(q) &= qD_h(q) - c\left(D_h(q)\right) \\
\pi^l_m(q) &= qD_l(q) - c\left(D_l(q)\right) \\
\pi^h_o(q) &= q\frac{D_h(q)}{n} - c\left(\frac{D_h(q)}{n}\right) \\
\pi^l_o(q) &= q\frac{D_l(q)}{n} - c\left(\frac{D_l(q)}{n}\right)
\end{aligned}$$

[25] In a multi-period model even a mixed strategy in any given period will converge so that a consumer will have seen about $t \cdot \rho$ after t periods where ρ is the mean of the firm's one-period strategy. So for large t, mixed strategies don't look terribly different from pure strategies. When generalizing to this sort of model the important factor will be an advertising *rate* as opposed to an advertising level since the experience set, P, is always growing, and will never consist solely of advertisements.

We can rewrite the conditions for equilibrium with respect to these functions.

$$\begin{aligned}\pi_o^h(q) &\geq \pi_m^h(q^*),\ q^* = \min(q_m^h, q)\\ \pi_o^h(q) - c_A &\geq \pi_m^l(q^*),\ q^* = \min(q_m^l, q)\\ \pi_o^h(q) - c_A &\geq \pi_o^l(q)\end{aligned}$$

Claim 1.6.3 *If the monopoly profit functions for the high-Y and low-Y goods, $\pi_m^h(q)$ and $\pi_m^l(q)$ are quasi-concave, advertising will only be possible in a pure strategy equilibria if the monopoly price for the high-Y demand curve q_m^h is greater than q^* where $q^* = \inf\{q|q - c'\left(\frac{D_l(q)}{n}\right) > 0\}$.*[26] *(See Appendix A2 for Proof).*

These bounds are unfortunately extremely loose, so they tell us very little about how equilibrium prices should depend on the number of firms, or the cost of advertising. In fact, equilibrium prices will depend heavily on the exact form of the demand and cost functions above. However, we can derive much tighter bounds for a parametric example.

Claim 1.6.4 *For demand and cost functions of the form:*

$$\begin{aligned}\beta_h &< \beta_l\\ D_h(q) &= 1 - \beta_h q\\ D_l(q) &= 1 - \beta_l q\\ c(x) &= c \cdot x^2\end{aligned}$$

any equilibrium price q where advertising occurs is bounded:

$$\frac{c + \sqrt{c^2 + \frac{c_A n^2(n + c(\beta_h + \beta_l))}{\beta_l - \beta_h}}}{n + c(\beta_h + \beta_l)} \leq q \leq \frac{c(n+1)}{c\beta_h(n+1) + n}$$

(See Appendix A2 for Proof).

Consider the comparative statics of these bounds. The upper bound is decreasing in the number of firms n since the higher the price q, the more tempting it is to the advertising firm to monopolize the market. The lower bound, on the other hand, is increasing in the price of advertising, c_A, and the number of firms n. As the number of firms grow the range of possible equilibrium prices shrinks until, at high n and c_A, the interval is empty. This shows that as the cost of the public-good (the advertisement), and the number of firms increase, it becomes more and more difficult to find an

[26] *The lower bound ensures that even though this is a Bertrand oligopoly and firms are required to meet demand, marginal cost will never exceed marginal benefit in an advertising equilibrium. This is a result of the fact that if this is the case it will often be optimal for the advertising firm to purposefully decrease demand by refraining from advertising.*

equilibrium price.[27]

The profit functions, π^h_m, π^l_m, π^h_o, and π^l_o are plotted for $\beta_h = .05$, $\beta_l = .2$, $c = 1.5$, and $n = 2$ in Figure 1.12. Notice that as long as either of the monopoly profit functions is above the oligopoly function π^h_o, the system cannot be in equilibrium. Similarly, π^h_o must be above the low-Y oligopoly function, π^l_o. The dark region of the π^h_o curve shows the set of prices where they will be advertising in equilibrium when $c_A = 0$. Increasing c_A will simply transpose both high-Y curves down by c_A and lead to a smaller region of possible equilibrium prices, in particular the minimum equilibrium price with advertising is increasing in c_A and the maximum equilibrium price is decreasing in c_A.

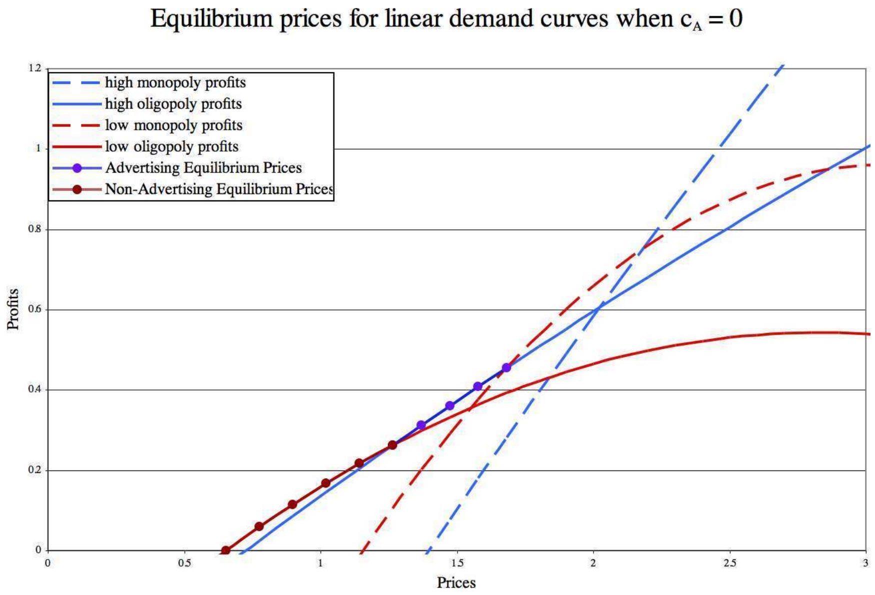


Figure 1.12: The four profit curves, π^h_m, π^l_m, π^h_o, and π^l_o for the demands and costs specified above. Notice that neither firm will ever have an incentive to *raise* prices above those of its competitor since then it would face a demand of 0. However, there may be incentives to *cut* prices or to stop advertising. Cutting prices is equivalent to switching to one of the monopoly curves and moving to the left, so in equilibrium the monopoly curves must be below the oligopoly curve in use (could be either high- or low-Y). Cutting advertising is equivalent to switching from the high-oligopoly curve to either of the low-Y curves (monopoly or oligopoly). So these must both fall below the high-oligopoly curve in equilibrium. Since all the curves in this example are quasi-concave and we only consider prices below their maxima, these conditions are not only necessary but sufficient for equilibrium. Advertising equilibrium prices are highlighted in blue, non-advertising equilibrium prices are highlighted in red.

Second Case: $1 \leq \alpha \leq 2$ In this case, the input w_{ZY} from Z to Y can never be strong enough to activate Y by itself, so a firm must advertise in order to form an association with Y. This changes the structure of the market radically. Any firm that wishes to associate its product with Y must

[27] These bounds only utilize the first and third equilibrium conditions listed above. The second, $\pi^h_o(q) - c_A \geq \pi^l_m(q)$, may bind from above or below depending on the specific values of β_h, β_l, and c. Notice that $\pi^h_o(q) - \pi^l_m(q) - c_A$ is a convex quadratic function, so it will generally have two roots. Equilibrium prices must be either below the lower root, or above the upper root of this quadratic. Increasing c_A simply transposes the quadratic profit function for the advertising firm down. So the lower root is decreasing in c_A and the upper root is increasing in c_A. This implies that the bound generated by this condition will get stricter as c_A whether it binds from above or from below.

advertise. As we will show, a limited number of firms will in fact be able to associate their product with Y. And the advertising cost of associating its product with Y is increasing in the number of other firms that choose to advertise.[28]

Once again, all advertisements will be of the form $(-, \ldots, -, 1, 1)$, so, likewise, $w_{ZY} = 1$ if any firm advertises and 0 otherwise. Let F^i advertise. We know that the firm's product will only be associated with Y if its advertisements make up a significant portion of total advertisements seen by the consumers. Formally,

$$A_Y(P, X^i) = 1 \Leftrightarrow w_{F^iY} = \frac{N_i}{\sum_j N_j} > \alpha_Y - 1$$

where N_j is the number of ads purchased by F^j.

Claim 1.6.5 *When the threshold for Y, $\alpha > 1$, there is an upper-bound of $\lfloor \frac{1}{\alpha_Y - 1} \rfloor$ on the number of firms that can be associated with Y.*

Proof:

If any firm advertises the sum of the weights

$$\sum_i w_{F^iY} = \sum_i \frac{N_i}{\sum_j N_j} = 1,$$

and as stated above, F^i will be associated with Y if and only if $w_{F^iY} > \alpha - 1$. At most $\lfloor \frac{1}{\alpha_Y - 1} \rfloor$ firms can satisfy this condition at once. We call this the *capacity* of the high-Y market. *Q.E.D.*

In order to understand this case we need to re-specify the demand functions to account for the fact that basic and high-Y products may coexist in the market. When $\alpha > 1$, there are cases where there will be firms with $A(P, X^i) = (-, \ldots, -, 1, 1)$ *and* firms where $A(P, X^i) = (-, \ldots, -, 1, 0)$. We need to know the preferences of individual consumers for the basic ($Y = 0$) vs. high-Y ($Y = 1$) version of the good at different prices in order to understand how the market splits between the two goods. We say that firms selling the high-Y good are part of the high-Y market. Similarly, firms selling the basic version of the good are part of the low-Y market.

Claim 1.6.6 *In a pure strategy equilibrium there will be one price for the high-Y market, q_h, and one for the low-Y market, q_l. When both types of products are sold, we will have $q_l < q_h$.*

Proof:

The firm itself will not directly affect the consumers' utility by the assumption that F^i is trivial. So, for any given consumer all products produced by firms F^i in the high-Y market are perfect

[28] The limited capacity feature of the high-Y market makes this game look somewhat like an *entry*-game from the economic literature.

substitutes. Consumers will maximize $u(x) - q_i x_1$ - $q_j x_2$, implying that consumers will all purchase high-Y goods from the firm(s) charging the lowest price, q_h^*. Any firm charging $q > q_h^*$ will sell no product and make negative profits $-N_i c_A$. Therefore, in a pure strategy equilibrium, all firms that enter the high-Y market will charge a single price q_h. The argument for the low-Y market is the same.

We said that Y is a universally positive social attribute, so for any consumer $u_1(x) > u_2(x)$ for all consumers and at all x. Therefore if $q_h \leq q_l$ the demand for the low-Y product will be 0, proving the claim. *Q.E.D.*

Without loss of generality, assume that firms $(F_1, \ldots, F_k)$ advertise and firms $(F_{k+1}, \ldots, F_n)$ do not.

Since we know that there will only be two prices q_h and q_l, we can define two functions:

$D_h(q_h, q_l)$	The demand for the high-Y good at prices q_h and q_l
$D_l(q_h, q_l)$	The demand for the low-Y good at prices q_h and q_l.

We assume that $\frac{\partial D_i}{\partial p_i} < 0$. In other words, demand is decreasing in own-price. We also define the following profit functions:

$$\begin{aligned} \pi_m^h(q, q_l) &= qD_h(q, q_l) - c(D_h(q, q_l)) \\ \pi_m^l(q, q_h) &= qD_l(q_h, q) - c(D_l(q_h, q)). \end{aligned}$$

$\pi_m^h(q, q_l)$ describes the profits that would be achieved if a firm monopolizes the high-Y market at price q when the price in the low-Y market is q_l, while $\pi_m^l(q, q_h)$ does the same for a firm monopolizing the low-Y market.

Assume that for all fixed q_l, $\pi_m^h(q, q_l)$ is quasi-concave in q and has some finite maximum $q_m^h(q_l)$. Similarly assume that for all fixed q_h, $\pi_m^l(q, q_h)$ is quasi-concave and achieves its maximum at $q_m^l(q_h)$.

To simplify the notation D_h and D_l refer to the demands for some equilibrium prices q_h and q_l unless noted otherwise.

Claim 1.6.7 *In any pure strategy equilibrium, all firms that advertise will advertise at the same level N.*

Proof:

Assume that this is not true. Then there must exist two firms F^i and F^j with advertising levels $N_i > N_j$. In equilibrium, F^j will only advertise if it successfully associates its product with Y.

Therefore we know that

$$\frac{N_j}{N_i + N_j + \sum_{k \neq i,j} N_k} > \alpha - 1.$$

A necessary and sufficient condition for F^i's product to be associated with Y is that

$$\frac{N_i}{N_i + N_j + \sum_{k \neq i,j} N_k} > \alpha - 1.$$

This means that F^i could decrease its advertising level to $N_i = N_j$ and still face D_h since

$$\frac{N_j}{2N_j + \sum_{k \neq i,j} N_k} > \frac{N_j}{N_i + N_j + \sum_{k \neq i,j} N_k} > \alpha - 1.$$

So for all $i \leq k$, $N_i = N$ as desired. *Q.E.D.*

Claims 1.6.5 and 1.6.6 narrow the relevant strategies considerably. An equilibrium will be completely described by four parameters:

1. k, the number of firms in the high-Y market
2. N, the number of advertisements purchased by firms in the high-Y market
3. q_h, the price of the high-Y good
4. q_l, the price of the low-Y good.

In a pure strategy equilibrium with non-empty high-Y and low-Y markets the following conditions must be met:

1. The number of firms who advertise must be less than the capacity of the high-Y market.

$$k < \frac{1}{\alpha - 1}$$

2. Firms should not have an incentive to undercut the price of their competitors to monopolize either the high or low-Y market.

$$q_h \frac{D_h}{k} - c\left(\frac{D_h}{k}\right) > \pi_m^h(q_h),\ q_h \leq q_m^h(q_l)$$

$$q_l \frac{D_l}{n-k} - c\left(\frac{D_l}{n-k}\right) > \pi_m^l(q_l),\ q_l \leq q_m^l(q_h)$$

3. Firms that advertise should not have an incentive to change the advertising level.
4. Firms that do not advertise should not want to enter the high-Y market. In other words, the cost of entry should outweigh the benefits of entry.

5. Finally, firms that advertise should not want to exit the high-Y market.

The first two conditions are fairly self-contained, so we will consider the implications of the third, fourth, and fifth conditions. The third condition places limits on the possible advertising levels given a fixed number of firms, k, in the market. Intuitively, advertising levels should be "minimal" in the sense that firms should not be able to lower the advertising level and still take part in the high-Y market. However, advertising levels must be high enough that it is too expensive for any firm to raise advertising levels and monopolize the high-Y market. Together, these conditions give us upper and lower bounds on the advertising level N.

Claim 1.6.8 *In any equilibrium with non-empty low and high-Y markets, the advertising level for firms in the high-Y market must fall within a specified range:*

$$\frac{\pi_m^h(q_m^h(q_l), q_l) - \frac{q_h}{k} D_h + c\left(\frac{D_h}{k}\right)}{c_A \left\lceil \frac{1-k(\alpha-1)}{\alpha-1} \right\rceil} \leq N \leq \frac{2-\alpha}{1-k(\alpha-1)}$$

(See Appendix A2 for Proof.)

Consider the comparative statics of this set of inequalities as k approaches the capacity of the high-Y market, $\frac{1}{\alpha-1}$. The right-hand side is increasing quickly toward ∞. The left-hand side is also increasing in k as long as $q_h \geq c'\left(\frac{D_h(q_h,q_l)}{k}\right)$, the marginal cost faced by an advertising firm.[29] When this is not true, the profit of the firm is actually decreasing in output, so no firm will ever want to monopolize the high-Y market. Assuming that this is not the case, i.e., q_h is high enough that firms would *want* a larger market share, advertising levels are increasing in k.

That brings us to the fourth and fifth conditions. How many firms should actually join the high-Y market in equilibrium? Once again we look at the problem intuitively. If there are too few firms, the benefits to entry will be high (assuming a high enough price q_h) and firms from the low-Y market will have an incentive to enter the high-Y market. However, as noted above, advertising costs are increasing in k. If too many firms enter the high-Y market those firms have an incentive to stop advertising and participate in the low-Y market.

First, we define functions specifying the benefits of both simple entry to the high-Y market, and entering and monopolizing the high-Y market:

$$\begin{aligned} b_e(k, q_h, q_l) &= \frac{q_h}{k+1} D_h - \frac{q_l}{n-k} D_l + c\left(\frac{D_l}{n-k}\right) - c\left(\frac{D_h}{k+1}\right) \\ b_m(k, q_l) &= p_m^h(q_l) D_h(q_m^h(q_l), q_l) - \frac{q_l}{n-k} D_l + c\left(\frac{D_l}{n-k}\right) \end{aligned}$$

[29] In this section we assume that firms must produce to meet demand. However, in a more general setting where we allow firms to choose their own output, the marginal cost of production $c'(x)$ must be less than the marginal revenue (the price of the good q).

We also define cost functions for entering the high-Y market, and entering and monopolizing the high-Y market

$$\begin{aligned} c_e(k,N) &= \left\lceil \frac{k(\alpha-1)}{2-\alpha} N \right\rceil c_A \\ c_m(k,N) &= \left\lceil \frac{1-(\alpha-1)k}{\alpha-1} N \right\rceil c_A \end{aligned}$$

Using these, along with the bounds on the advertising level N shown above, we can derive conditions on the number of firms in the high-Y market that don't depend on the specific value of N (for details of these derivations see Appendix A.2):

$$\begin{aligned} \left\lceil \frac{k(\alpha-1)}{1-k(\alpha-1)} \right\rceil c_A &> b_e(k,q_l,q_h) \\ \left\lceil \frac{2-\alpha}{\alpha-1} \right\rceil c_A &> b_m(k,q_l). \end{aligned}$$

Once again, we consider the comparative statics of these conditions. Consider the first inequality: The benefit of entry is decreasing in k while $q_h > c'\left(\frac{D_h}{k+1}\right)$ and $q_l > c'\left(\frac{D_l}{n-k}\right)$, and the cost of entry is always increasing in k. This means that this condition is more and more easily met as k approaches the capacity of the market, i.e., as the high-Y market approaches capacity the benefits to entry decrease while the costs to entry increase, removing the incentive for firms in the low-Y market to enter.

Finally, advertising firms should want to stay in the high-Y market. This requires that the benefits of staying outweigh the costs of staying. Using the functions defined above we can rewrite this as

$$\begin{aligned} b_e(k-1,q_l,q_m) &> c_e(k-1,N) \\ b_e(k,q_l,q_m) &\leq c_e(k,N). \end{aligned}$$

Since benefits are decreasing and costs are increasing in k, this boils down to finding the crossing point for these two functions.

We can now start to see the dynamics of this system. At low k, b_e is high and firms will enter. This, in turn, will increase the incentive for firms in the high-Y market to attempt to monopolize the high-Y market, which will drive the advertising levels, and cost of entry c_e up. This will continue until the costs outweigh the benefits. If too many firms enter, the cost of staying in the market will actually motivate firms in the high-Y market to exit and sell products in the low-Y market. This cycle is illustrated in Figure 1.13.

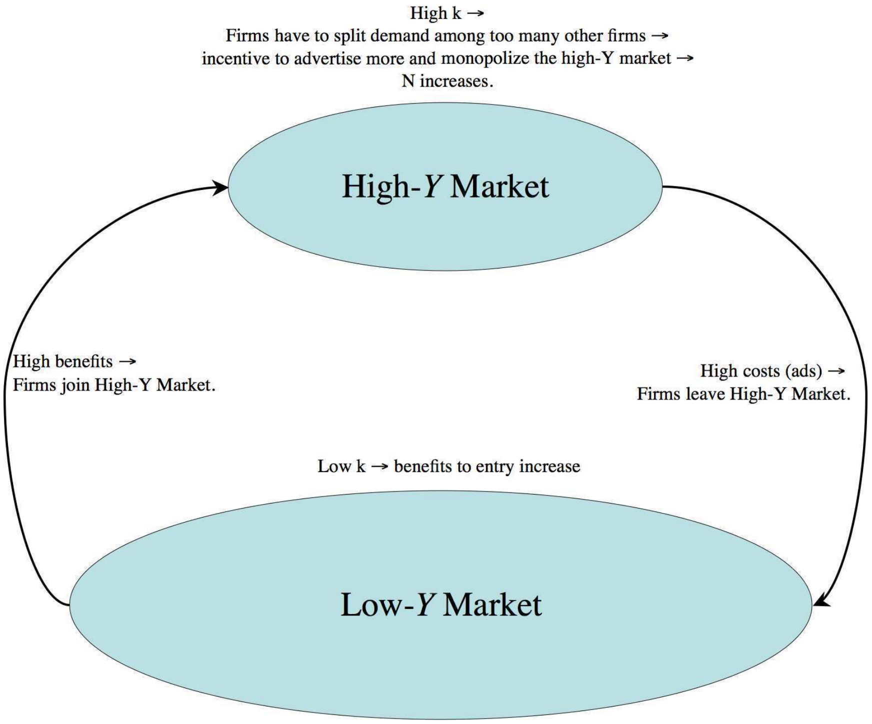

High k →
Firms have to split demand among too many other firms →
incentive to advertise more and monopolize the high-Y market →
N increases.
High-Y Market
High benefits →
Firms join High-Y Market.
High costs (ads) →
Firms leave High-Y Market.
Low k → benefits to entry increase
Low-Y Market

Figure 1.13: Flow diagram for oligopoly. If there are too many firms in the high-Y market, the benefits to staying decrease and firms have an incentive to leave. When there are too few firms, firms will enter, driving up the advertising costs of joining or staying in the high-Y market. When these costs have just exceeded the benefits, the system is in equilibrium

The main points to take from this example are that because of the existence of social attributes, it may be beneficial for even monopolies to advertise. However, without the presence of distinguishing attributes, monopolies are constrained to market either a high or low-Y good. In oligopolies when Y has a low threshold, free-riding is a significant problem that may prevent advertising even if it would increase overall welfare. This is because the cost of creating utility from the social attribute is fixed (the cost of one advertisement), but no one firm will be able to reap the benefits of associating the product with Y. And finally, for Y with a high threshold, oligopolies look like an entry game where the cost of entry is increasing in the number of firms in the high-Y market.

1.6.2 Product Differentiation

In this section we use the model developed in sections 1.3 and 1.4 to explore sub-branding as a means to create product differentiation. In these cases a firm uses a parent brand along with an extra sub-brand label to introduce a new product. This strategy is usually employed when firms are introducing a new version of the same type of product already sold under the brand name.

Prominent examples of this sort of extension are Diet Coke, or Miller Hi-Life.

For any given product type there is a distribution of preferences over possible characteristics. This gives firms an incentive to offer a diverse set of products. However, the associative structure of this model will constrain the degree to which they diversify under a given brand name. Diversifying over search attributes is generally easy because they are readily apparent (i.e., they are in the initial pattern X), and they are highly self connected ($w_{ii} \gg 0$), so they will remain in their initial state in pattern $A(P, X)$ regardless of the content of P. However, this is not true of experience or social attributes. We will once again focus on these attributes in the analysis in this section.

Diversification over social attributes will often be desirable even for universally positive characteristics like the one considered in section 1.6.1. This stems from that fact that even though all people might prefer the high-Y product to the basic product, the amount any given consumer would be willing to pay for the high-Y good varies. This may allow firms to engage in price discrimination, i.e., they can offer both a cheap, basic good and a more expensive, high-Y good. Consumers who place more value on the characteristic Y will purchase the more expensive good, but the firm will still be able to sell the basic product to consumers who value Y less. The benefits of diversifying over characteristics such as the quality of a good have been well studied both empirically and theoretically (3; 51; 69). But, as we saw in previous section, achieving diversification over social characteristics is non-trivial. We saw that a monopoly with only one brand had to choose between selling a good with social attribute Y or without Y. Even though monopolistic price discrimination might be desirable, the firm cannot practice discrimination without introducing a search attribute to differentiate between the two products.

The key purpose of advertising in this case is to create product differentiation: the two products offered by the firm must not be substitutes to the consumers. If this does not occur, the new product will, at best, cannibalize consumers from the original rather than actually expanding the market.

Sub-branding is a common method used to create this sort of product differentiation. However, this is not the only way to expand the firm's market. Greg Stine, a marketing expert who criticizes sub-branding as a strategy, said “The smarter, long-term plan however, is to apply the same branding techniques used to create your successful brand and create a new, completely distinct new entity. Many very successful companies do this very effectively and we don't notice (we just buy their stuff).” (70) The major Japanese car manufacturers used this strategy to enter the luxury car market: Lexus for Toyota, Acura for Honda, and Infinity for Nissan. In this section we compare these two methods with regard to the costs of differentiation.

We compare these methods under two different network structures: one where there are only four nodes, corresponding to the parent brand F_1, the new brand F_s, the product type Z, and the social characteristic Y; and one where there are five nodes, the aforementioned four and a fifth, shared social or experience attribute R.

For the remainder of the section we assume that the firms have a constant returns to scale production function. This means that the cost of producing some quantity x of the good is a constant multiple of x: $c(x) = c \cdot x$. Therefore, for any price $p > c' = c$, firms strictly prefer more demand and thus product diversification will always be desirable. This reduces the firm's problem to minimizing the costs of differentiation.

1.6.2.1 Product differentiation without a shared social or experience attribute

Consider the network in Figure 1.14 with nodes (F_1, F_s, Z, Y). The sub-branding strategy gives us advertisements

$$\begin{aligned} \mathcal{A}_1 &= (1,0,1,0) \\ \mathcal{A}_s &= (1,1,1,1). \end{aligned}$$

Let N_1 and N_s be the number of ads $\mathcal{A}_1$ and $\mathcal{A}_s$ purchased respectively. This gives us the weight matrix

$$W = \begin{pmatrix} w_{FF} & \frac{N_s}{N_1+N_s} & 1 & \frac{N_s}{N_1+N_s} \\ \frac{N_s}{N_1+N_s} & w_{F_sF_s} & \frac{N_s}{N_1+N_s} & \frac{N_s}{N_1+N_s} \\ 1 & \frac{N_s}{N_1+N_s} & w_{ZZ} & \frac{N_s}{N_1+N_s} \\ \frac{N_s}{N_1+N_s} & \frac{N_s}{N_1+N_s} & \frac{N_s}{N_1+N_s} & 0 \end{pmatrix}$$

The requirements for product diversification in this network are also shown in Figure 1.14. The input to the Y node must be below threshold when the network is in state $(1,0,1,0)$ and only the parent brand is salient, but above threshold when the network is in state $(1,1,1,0)$ and the sub-brand is salient and present. In addition, the sub-brand should not be too associated with the target brand. Specifically, it is important that the node representing the sub-brand is not activated by the parent brand. Mathematically this means

$$\begin{aligned} \frac{2N_s}{N_1+N_s} &< \alpha_Y < \frac{3N_s}{N_1+N_s} \\ \frac{2N_s}{N_1+N_s} &< \alpha_{F^s} \end{aligned}$$

We assume for this example that $\alpha_{F^s} > 2$ so the second condition is not binding.[30] We also assume that $\alpha_Y < 2$ so that the creating a new brand is a feasible strategy.[31] First, notice that the first condition requires $N_1 > 0$, so unlike the oligopoly example described in section 1.6.1, there is an incentive to actually advertise the low-Y product. Advertising the low-Y product purposefully

[30] However, activation of the sub-brand is a possible issue for companies. You do not want to create a sub-brand that is so popular that it overshadows, and possibly steals market share, from the parent brand.

[31] For $2 \leq \alpha_Y < 3$, $N_1 = 0$ and $N_s = 1$. Creating a new brand is not feasible since the input to Y from Z and F^s is at most 2.

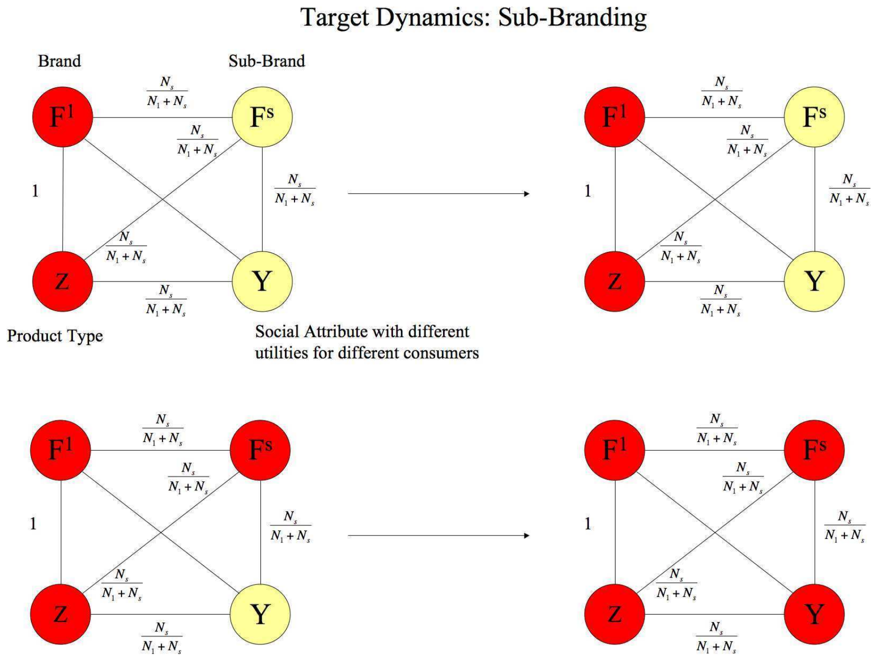

Target Dynamics: Sub-Branding
Brand
Sub-Brand
F1
Fs
Z
Y
1
$\frac{N_s}{N_1+N_s}$
Product Type
Social Attribute with different utilities for different consumers

Figure 1.14: Weights and target dynamics for the sub-branding strategy. F^1 is the node representing the parent brand while F^s represents the sub-brand. The firm wishes to diversify over attribute Y, which means that $A(P,(1,0,1,0)) = (1,0,1,0)$, so the parent brand alone does not activate Y as shown in the top row, and $A(P,(1,1,1,0)) = (1,1,1,1)$, so the presence of the sub-brand will activate Y as shown in the bottom row.

dilutes the parent brand's connection to Y. We can rewrite these inequalities as upper and lower bounds on the ratio for N_1 to N_s:

$$\frac{2-\alpha_Y}{\alpha_Y} < \frac{N_1}{N_s} < \frac{3-\alpha_Y}{\alpha_Y}$$

When $\alpha_Y > 1.5$, making the social attribute difficult to activate, the upper-bound is the limiting factor, so $N_1 = 1$ and $N_s = \left\lceil \frac{\alpha_Y}{3-\alpha_Y} \right\rceil$. This simply means that when Y is hard, the difficulty in advertising is actually establishing the high-Y version of the product and $N_s \geq N_1$. Since in this case we can only consider $\alpha_Y < 2$, this means that for all $\alpha_Y > 1.5$, $N_1 = 1$ and $N_s = 2$.

Using similar logic, when $\alpha_Y < 1$, the lower bound is the limiting factor and we generally get $N_s = 1$ and $N_1 = \left\lceil \frac{2-\alpha_Y}{\alpha_Y} \right\rceil$. The intuition behind this is that when Y is easy to activate, the difficulty for the firm is actually establishing the low-Y version of the product, so most advertising is aimed at diluting the connection between the parent brand and Y.

On the other hand, if we create a new brand we do not have to work as hard to differentiate the basic product from the new one. Specifically, the firm will not have to purposefully dilute any brand associations. The network for the new-brand strategy is shown in Figure 1.15 with the connection weights labeled.

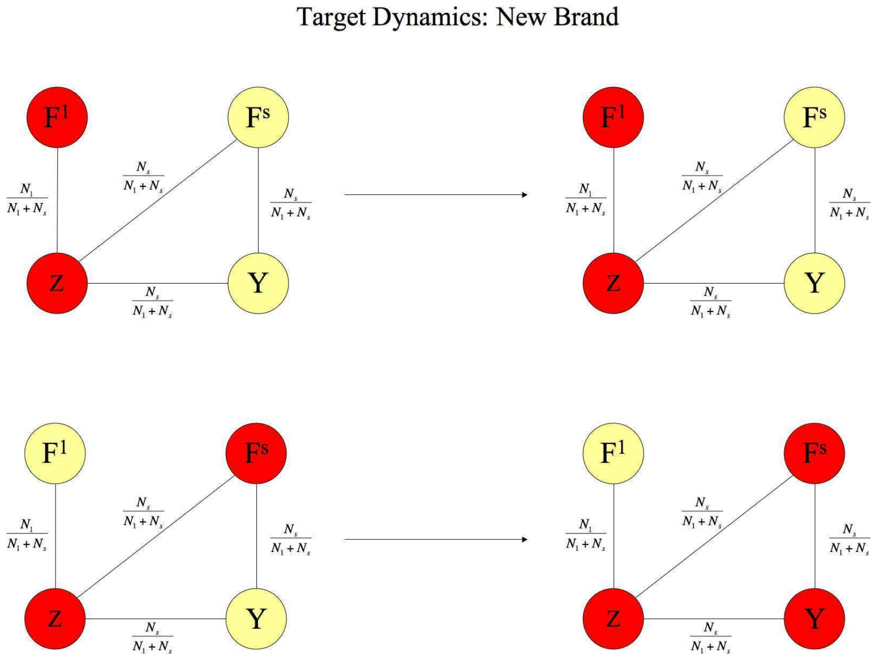


Figure 1.15: Weights and target dynamics for creating a new brand. Once again F^1 represents the parent brand, however F^s is now a completely new brand, meaning that the firm will not reference the parent brand F^1 in advertisements for F^s. The top row shows the desired network dynamics for the low-Y product and the bottom row shows the desired network dynamics for the high-Y product.

Our new condition for product differentiation is that the connection weights are such that the input to Y from Z alone will not activate Y (since there will be no connection between F and Y), but the input from F^s and Z to Y together will activate it. Mathematically, this comes to

$$\frac{N_s}{N_s + N_1} < \alpha_Y < \frac{2N_s}{N_s + N_1}.$$

Or, rewritten:

$$\frac{1-\alpha_Y}{\alpha_Y} < \frac{N_1}{N_s} < \frac{2-\alpha_Y}{\alpha_Y}$$

First, notice that for all $\alpha_Y > 1$, the advertising levels are $N_1 = 0$, $N_s = 1$ since the input from Z alone can never activate Y. However, when $\alpha_Y < 1$ the firm must still advertise for both products in order to dilute the connection between Z and Y, w_{ZY}. The advantage of this strategy is that the lower bound for $\frac{N_1}{N_s}$ is less strict than for the sub-branding strategy.

The total advertising levels required for product differentiation under both strategies are graphed in Figure 1.16. As you can see, inducing product differentiation by creating a new brand is *always* cheaper than using a sub-brand under these conditions.

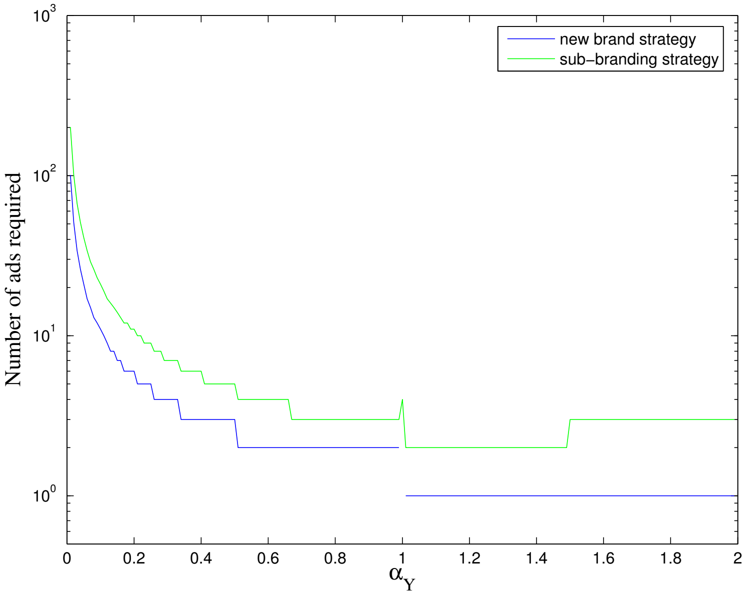


Figure 1.16: Required number of ads for the sub-branding strategy (green) and the new-brand strategy (blue) when there is no shared non-search attribute.

1.6.2.2 Product differentiation with one shared social or experience attribute

The justification frequently used by firms in order to sub-brand rather than create a new brand is that association with the parent brand somehow increases its probability of purchase. This is usually couched in terms of loyalty and reputation, but both of these terms really refer to the fact that the parent brand is associated with a specific characteristic that the firm wishes to transfer to the new product. At the same time, a sub-brand is created so that the firm can diversify over some other characteristic. To better understand this strategy, we will look at the dynamics of a 5-node network with characteristics (F^1, F^s, Z, R, Y). The firm wishes to diversify over Y while keeping an association with R. Using a sub-branding strategy they will broadcast two different advertisements

$$\begin{aligned} \mathcal{A}_1 &= (1,0,1,1,0) \\ \mathcal{A}_s &= (1,1,1,1,1) \end{aligned}$$

The search characteristics for the two products will be $X^1 = (1,0,1,0,0)$ and $X^s = (1,1,1,0,0)$. Notice that these two patterns are extremely similar, so for many different weight matrices they will be in the same basin of attraction for the network. This means that differentiating the products will be difficult.

Let N_1 and N_s be the number of the parent and sub-brand advertisements purchased by the firm. The resulting network is shown in Figure 1.17.

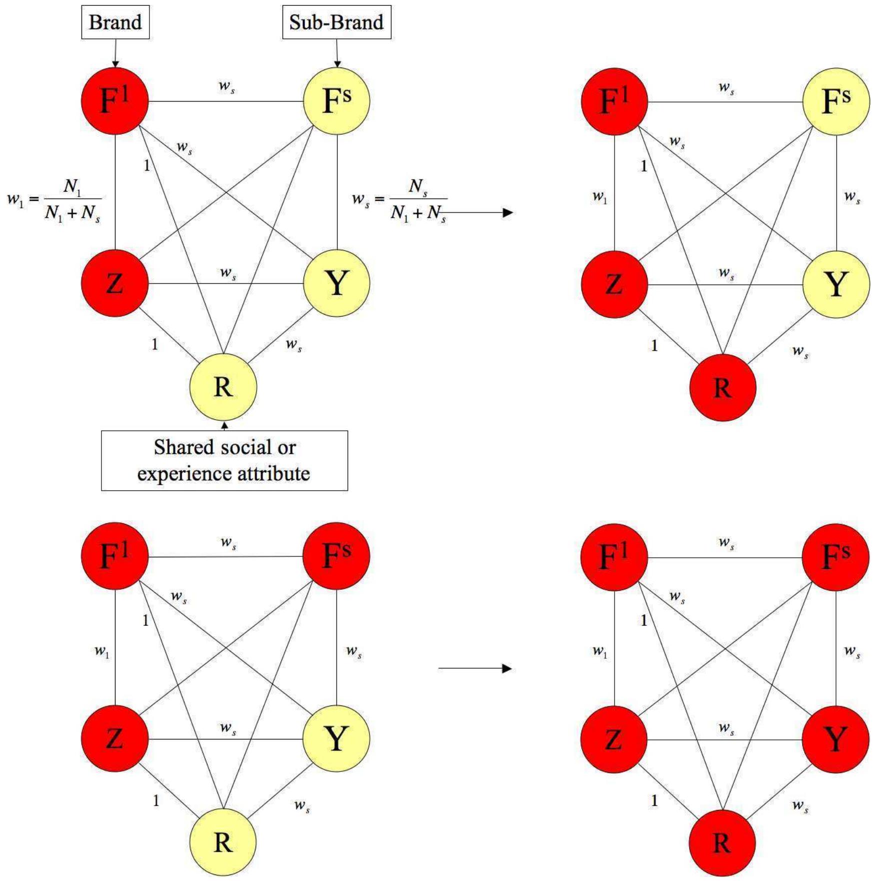

Brand
Sub-Brand
Shared social or experience attribute
F1
Fs
Z
Y
R
w_s
1
$w_1 = \frac{N_1}{N_1 + N_s}$
$w_s = \frac{N_s}{N_1 + N_s}$
w_1

Figure 1.17: Weights and target dynamics for the sub-branding strategy with a shared social attribute. The top row shows the desired network dynamics for the low-Y product and the bottom row shows the desired network dynamics for the high-Y product.

We assume that $\alpha_{F_s} > 4$ so we still do not have to worry about X^1 activating F_s. The necessary and sufficient condition for product differentiation over Y is that

$$\frac{3N_s}{N_1 + N_s} < \alpha_Y < \frac{4N_s}{N_1 + N_s},$$

which just means that the connection weight between F^s and Y, w_{F^sY} is just enough to take the input to Y over the activation threshold α_Y. The necessary and sufficient condition for both products to be associated with R is

$$2 > \alpha_R.$$

We assume that $\alpha_R < 2$ so that the second condition is met immediately. We also assume that $\alpha_Y < 3$ so that differentiation is possible using either a sub-branding strategy or by creating a new brand.

What is the minimal level of advertising necessary to achieve product differentiation? We can re-write the inequalities above as

$$\frac{3-\alpha_Y}{\alpha_Y} < \frac{N_1}{N_s} < \frac{4-\alpha_Y}{\alpha_Y}$$

Notice that since $\alpha_Y < 3$ the lower bound is greater than 0 so N_1, N_s must both be greater than 1. Using the conditions on $\frac{N_1}{N_s}$ we find that the advertising levels are generally as summarized in Table 1.4.[32]

Activation Threshold	Ads for low-Y product	Ads for high-Y product
$\alpha \in (1.5, 2)$	$N_1 = 1$	$N_s = 1$
$\alpha < 1.5$	$N_1 = \lceil \frac{3-\alpha}{\alpha} + \epsilon \rceil$	$N_s = 1$
$\alpha \in (2, 3)$	$N_1 = 1$	$N_s = \left\lceil \frac{\alpha}{4-\alpha} + \epsilon \right\rceil.$

Table 1.5: Required advertising levels to induce product differentiation using a sub-branding strategy

Now consider the alternative strategy, creating a new brand. The advertisements will now be

$$\begin{aligned} \mathcal{A}_1 &= (1,0,1,1,0) \\ \mathcal{A}_s &= (0,1,1,1,1). \end{aligned}$$

The network generated by these advertisements and the target network dynamics are shown in Figure 1.18. The new requirements for differentiation over Y while maintaining both brands' association with R are:

$$\begin{aligned} \frac{2N_s}{N_1+N_s} &< \alpha_Y < \frac{3N_s}{N_1+N_s} \\ \frac{2N_s+N_1}{N_1+N_s} &> \alpha_R \\ \frac{2N_1+N_s}{N_1+N_s} &> \alpha_R. \end{aligned}$$

The first thing to notice is that for $\alpha_R \geq 1.5$, this strategy is not feasible since the capacity for high-R brands will be 1. On the other hand, for $\alpha_R < 1$ the latter two conditions will never bind and we only need to worry about differentiation over Y, given by the first condition.

[32] There are a few exceptions where the prescribed advertising levels summarized here fall exactly on a boundary or both conditions bind N_1 and N_s away from 1.

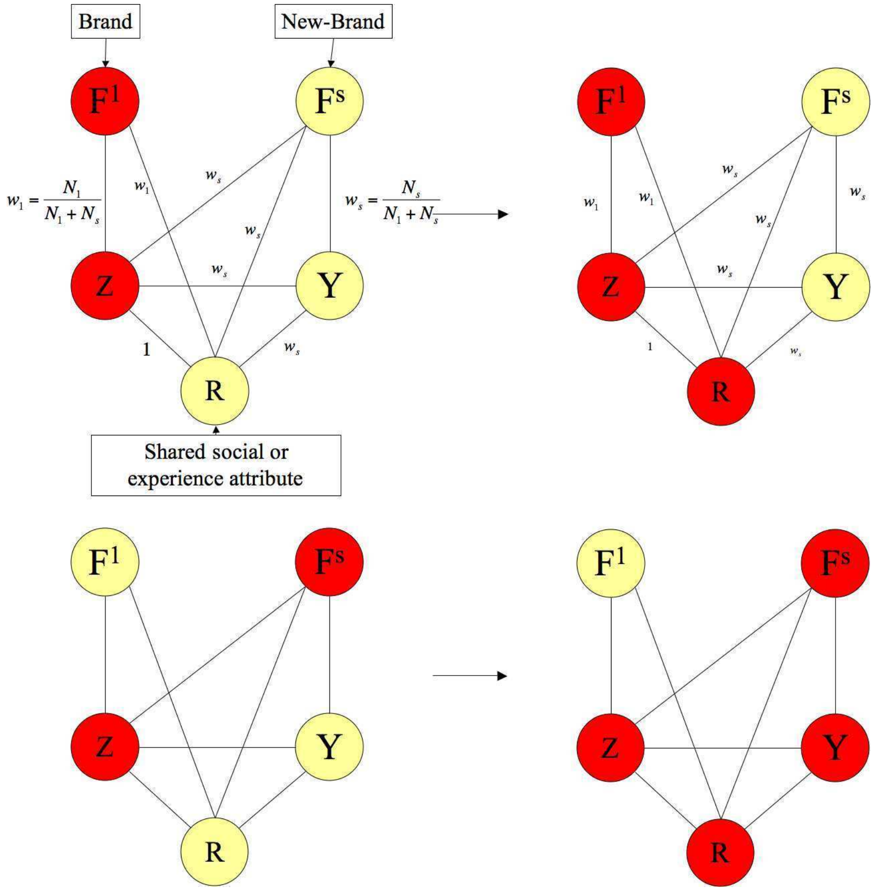


Figure 1.18: Target dynamics for the new brand strategy with one shared social or experience attribute. Once again, the top row shows desired dynamics for the parent brand while the bottom row shows desired dynamics for the new brand.

We can once again rewrite the condition for differentiation over Y as bounds on the ratio $\frac{N_1}{N_s}$

$$\frac{2-\alpha_Y}{\alpha_Y} < \frac{N_1}{N_s} < \frac{3-\alpha_Y}{\alpha_Y}$$

Figure 1.19 graphs the relative advertising levels required for both strategies with $\alpha_R < 1$ and $\alpha_Y \in (0,3)$. As you can see, developing a new brand is cheaper for all $\alpha_Y \notin (1.5, 2)$. This shows that when the shared attribute R is relatively easy to activate, starting a new brand is *still* generally cheaper than using a sub-brand.

For $\alpha_R \in [1, 1.5)$ the conditions for association with R actually bind, so the conditions for differentiation while retaining association with R are stricter:

$$\frac{N_1}{N_s} \in \left(\frac{2-\alpha_Y}{\alpha_Y}, \frac{3-\alpha_Y}{\alpha_Y}\right) \bigcap \left(\frac{\alpha_R - 1}{2-\alpha_r}, \frac{2-\alpha_R}{\alpha_R - 1}\right).$$

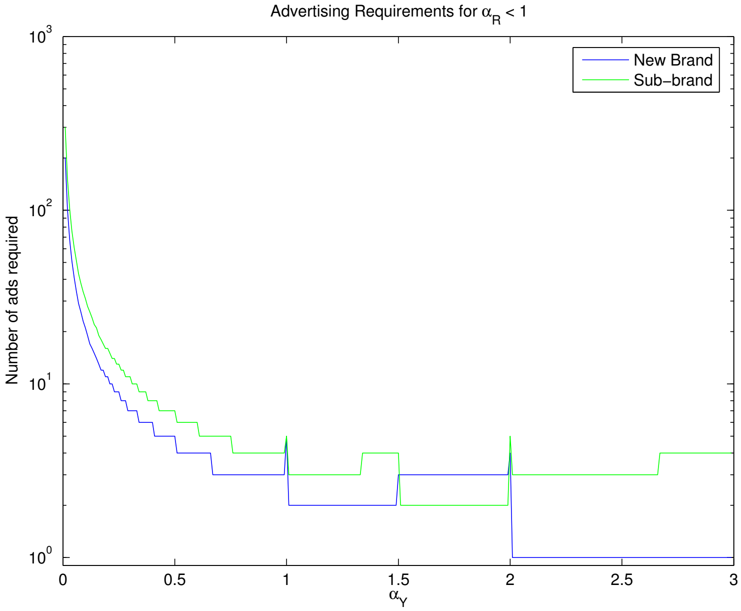


Figure 1.19: Advertising levels for the new and sub-brand strategy when there is a shared attribute R and $\alpha_R < 1$

In this case creating a new brand may not always even be a feasible strategy. For example, when $2\alpha_R - 2 > \alpha_Y$ or $6 - \alpha_R < \alpha_Y$ this intersection is empty. Even when the intersection exists, the sub-branding strategy is always cheaper over these values. By creating a new brand the firm is essentially competing with itself over the market for high-R products. In these cases using a sub-branding strategy is justified, but notice that this will only be true when the desirable shared characteristics is difficult to activate. The surface in Figure 1.20 shows the advertising costs for the new brand strategy as we vary both α_R and α_Y. Required advertising levels are increasing in both activation thresholds. The sections of the $\alpha_Y \times \alpha_R$ plane not covered by the surface are regions where creating a new brand is not a feasible strategy.

These calculations show that under these conditions it is almost always *cheaper* in terms of advertising costs to create a new brand than to use a sub-brand to diversify over a social attribute. The major exception is when the firm wishes to associate both products with some non-search attribute while diversifying over another. Even under these circumstances, it is generally only cheaper to sub-brand if the shared attribute is “hard” in the sense of having a high threshold $\alpha_R > 1$.

It is important to note that if the shared attribute is a search attribute, there is no reason for the firm to advertise this attribute with *either* brand, so the situation would be the same as in the first case we examined.

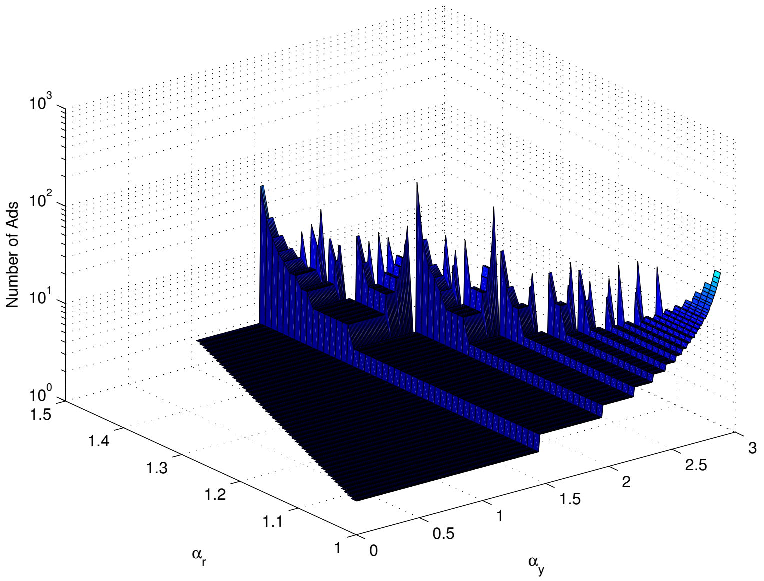


Figure 1.20: Advertising levels for the new brand strategy over $\alpha_R \in (1, 1.5)$ and $\alpha_Y \in (0, 3)$. The z-axis shows the number of ads required while the portion of the $x-y$-plane not covered by the surface are regions where product differentiation using a new brand is not possible.

1.7 Discussion

1.7.1 Limits

There are several open questions that we haven't discussed yet in this paper. The most important of these is attention. We restrict the initial vector X^i for a product to be the *salient* search attributes of the product, but the salience of any characteristic will be influenced by a host of environmental factors, as well as endogenously guided by the consumer. In our example markets we avoid these question by keeping the number of attributes low, but in practice any given attribute will have a host of search attributes.

In addition to determining which attributes are salient, there are cases where the *degree* of salience is relevant. For example, we showed that creating a new brand is often a better strategy than sub-branding to create product differentiation since it doesn't require the parent brand to dilute its connection with the differentiated attribute. However, Milberg, Park and McCarthy show experimentally that sub-branding is actually more beneficial than a direct extension where the firm does not create a new brand label at all. (54) This might be because creating a new label actually distracts from the parent brand name, attenuating the dilution effects of the extension. While the parent brand name is still salient, it is somehow less salient.

The study by Brown and Carpenter is another example of how directors of attention can effect choice. Trivial attributes in their study provided a meaningless differentiation of a product (either the products without the trivial attribute when most products had it, or the product with the trivial attribute when most did not) that none the less, directed attention at the differentiated product and increased the probability that it would be chosen. (13)

Another major issue is encoding. A consumer might encode an advertisement in any number of ways depending on how his attention is directed and what his existing associations are. Just as advertisements allow brands to compress information via associations, advertisements use symbols that compress information. For example, the ever present beer commercial with women in bikinis is not meant to associate beer with bikinis, but with the implied social characteristics of sexual attractiveness. Similarly, celebrity endorsements are meant to associate a brand with the positive social characteristics of the celebrity. The celebrity serves to represent a large number of attributes efficiently. While as noted above, the model in this paper is ill-suited to deal with the issues of attention, it can actually be applied to information compression within advertisements.

In general, the model described in this paper is best suited to deal with social as opposed to experience attributes. Since experience attributes are actually objective traits, we are able to assess them using various methods of inference as well as via association. Though the Gilovich study shows that experience attributes are still susceptible to the effects of association, associations are almost certainly not the only determining factor for our beliefs about these attributes. This will be especially true when subjects are specifically motivated to report their beliefs about product benefits as opposed to simply choosing. Problems framed in this manner are more likely to engage higher cognitive functions ("System II" in the language of Kahneman (34)). A study by Meyvis and Janiszewki suggests that when subjects are asked to report beliefs about specific product benefits they actually use a form of hypothesis testing such that irrelevant information in the form of trivial attributes actually *lowers* their beliefs. (53)

1.7.2 Parameter Estimation

The major contribution of this model is it proposes a concrete mechanism whereby the experiences of the consumer, specifically advertisements, affect his perceptions of products. However, the structure of the model raises several important questions, especially when it comes to the *application* of the model. How can we determine what the nodes of these networks should really be? Even once we do, the activation thresholds and self-connections constitute a large number of free parameters. I propose that while it is difficult to approach these problems from a purely theoretical standpoint, existing data sets and data analysis techniques will be able to constrain both the size of these networks and their free parameters.

In section 1.4 we address at least part of the problem by constraining the self-connection strengths based on the characteristic type. By assuming that social attributes have no self-connection we immediately remove a large number of free parameters, and we significantly limit the self-connections for search and experience attributes as well. However, in order to make this framework applicable to specific problems we need a concrete way to estimate the activation thresholds for characteristics.

The most direct way to measure activation thresholds would be to measure how much priming is necessary to induce a subject's free recall of the attribute. Subjects could be primed with varying numbers of moderately related concepts and asked to engage in a free recall task. The activation threshold for the attribute will be proportional to the number of concepts necessary to invoke recall. In order to avoid input "building" up, this sort of experiment would either have to be done between subjects using a very homogeneous population, or with significant breaks between free recall tasks. Either way, the estimated activation threshold would be specific to consumers from similar cultural and personal backgrounds.

Another, more indirect but perhaps easier, way to estimate the activation thresholds of different characteristics is to assume that they are inversely proportional to how common they are. How often do we actually see anything with a given characteristics? Characteristics like "nice" that we see frequently would have relatively low activation thresholds, while characteristics like "luxury" that we see relatively rarely would have high activation thresholds.[33]

In fact, studies have shown that the lexical frequency of a word, or a particular meaning of a word, modulates the ease with which it is accessed (73). While these studies examine lexical rather than semantic frequency,[34] they suggest that we may be able to use semantic frequency as a proxy for how easily the concept is accessed, which will be inversely proportional to its activation threshold.

It is also unclear how large these networks should be. How do we determine what the nodes of the network should be? Ideally the network should be large enough to reasonably describe the space of possible goods while remaining as small as possible to increase tractability. One way to approach this problem is to consider existing techniques for language processing. Latent semantic analysis is a specific technique using large bodies of text to index key concepts and analyze the relationships among them. A major contribution of this approach is that it specifies a means to effectively reduce the dimension of the semantic space considered. This dimension-reduction reduces the problems introduced both by synonyms (multiple words describing the same semantic concept), and polysemy (one word may refer to several different concepts). These methods might be used in the future to isolate the core characteristics relevant to either a particular type of good or even a large set of goods, limiting the size of the semantic network used to model the perceptions of these goods (47).

The final set of parameters that must be estimated are the connection strengths themselves. One psychological tool for estimating connection strengths is the implicit association test (26). In

[33] The cultural aspect of some of these characteristics, like luxury, will actually restrict the frequency with which we see them. To some degree, rarity is a necessary component of luxury.

[34] Lexical frequency refers to the frequency with which we see a specific word, whereas semantic frequency refers to the frequency with which we deal with a particular concept

this paradigm, subjects are asked to classify one of two groups of stimuli, for example pictures of light and dark skinned people or positive or negative words, and to respond with either a right- or left-hand button press. When associated stimuli share a response, for example light skinned people and positive emotional words both require a right hand button press, response times are faster than when dissociated stimuli share a response. The difference in response times is used to measure the degree of association, or dissociation between concepts. When this difference is large, the concepts — in the example, light skin and positive emotions — are strongly connected, corresponding to a high value for the connection weight. When there is little or no difference the connection strength will be close to zero. While these tests are best known for their use as tools to measure subjects' implicit racial, ethnic, or gender bias, the basic paradigm could easily be used to measure the degree of association between any two concepts for a given subject, or population of subjects.

1.7.3 Testable Implications

At the consumer level the model implies that advertising will increase the strength of association between a brand and some social characteristics of a product. This in turn should make the consumer attribute the social characteristics to the brand's products.

The IAT can be used to measure the effect of advertising on connection strengths. After a burst of advertising and a delay, response times for and IAT session where a brand is paired with an advertised social attribute should decrease. The magnitude of the decrease should be inversely related to both the familiarity of the brand and the age of the subject since the marginal effect of advertising depends on the change in the relative frequency of brand advertisements.

In addition, we can test the prediction that higher association strengths as measured by the IAT will increase the likelihood that consumers will actually attribute a characteristics to a branded product by measuring association strengths between a brand and a social characteristic, and then after a substantial delay (on the order of days or weeks), either ask them to engage in a free recall task listing the characteristics of a product from the target brand or directly elicit their beliefs about the advertised characteristic. In the free recall condition, subjects with higher association strengths should be more likely to list the target attribute. And in the belief elicitation condition, there should be a positive correlation between association strength as measured by the IAT and beliefs about the target attribute.

There are also testable implications at the firm level. First, we should observe that advertising levels for "social products," i.e., products valued largely for their social implications such as clothing, should be significantly higher than those that Nelson would classify as "search goods."

Recall that in an oligopoly, when the activation threshold for a social attribute is low, the connection between the product category and the social attribute may be sufficient to activate the social attribute. On the other hand, for high threshold characteristics, input from brand identifiers are necessary to activate the characteristic. This means that for “hard” social attributes, e.g. high social status, advertising will be purely predatory, i.e., advertising levels should have little to no effect on aggregate demand for the product category. This means that advertisements for luxury cars denoting status should not affect the aggregate demand for sedans (i.e., luxury will not rub off onto competitors). On the other hand, for low threshold attributes, e.g. “environmentally friendly,” will spillover to all similar products. So advertisements claiming that say Honda sedans are better for the environment should have global effects on the demand for sedans. This latter claim is supported by Seldon et. al’s (67) result that advertisements for specific cigarette brands do have positive effects on total demand for cigarettes since these advertisements essentially state that “(brand name)’s cigarettes are cool,” and cool appears to be a relatively low-threshold social characteristic.

Since advertising for high-threshold attributes is purely predatory, advertising levels for goods with these attributes will be increasing in the number of competing firms. We should only observe “advertising wars” for goods with high-threshold social attributes.

The model implications relating to product differentiation within a firm are somewhat harder to test empirically since so many firms fail to effectively differentiate their products. Failed product differentiation within the firm should lead to constant overall market share for the firm. For example, if Coke fails to differentiate its extension Diet Coke Plus from the original Diet Coke, demand for the two products should be the same as previous demand for Diet Coke alone. The model predicts that product differentiation over higher threshold attributes actually requires fewer advertisements, so we should observe that successful differentiation occurs more often over high-threshold attributes.

1.7.4 Extensions

1.7.4.1 Experiences and The Multi-period Model

Extending the current model to multiple periods importantly involves the expansion of the set P to include a variety of other types of experiences. I take the view that, with respect to associative memories, advertisements and experiences only differ in the relative strength of their impact. In this respect, the model is similar to Shapiro’s since advertisements and experiences are essentially pooled. However, in my model this pooling is not the result of a cognitive mistake, but a simple function of how associations are generated in the mind.

This raises the question, how strongly should non-advertising experiences be weighted in comparison to advertisements? It is clear that a consumption experience should heavily outweigh any single advertisement. But is it reasonable to think that advertising is able to attenuate the effects of a bad consumption experience? In addition to direct consumption experiences, consumers may also observe others consuming a product or hear about another person’s consumption experience. How should these indirect consumption experiences be weighted in relation to both direct experiences and advertisements?

The simplest way to reflect the relative strength of one experience as opposed to another is to literally count it multiple times. For example, we can assert that the pattern c, of attributes experienced by the consumer during consumption counts as if the consumer had seen m advertisements with pattern c. So if a consumer with an initial experience set P consumes a product which he encodes as having a pattern of attributes p, the new experience set is

$$P' = P \uplus \underbrace{\{c, \ldots, c\}}_{m}$$

We can similarly vary the weight of other experiences by manipulating the multiplicity, m, with which they enter the experience set.

Given these new weights we can specify the model in the following way:

1. At the beginning of every period, consumers start with some initial experience set $P^t - 1$.

2. Consumers then go through a consumption and exposure stage where they consume the goods purchased in period $t-1$ and view a set of advertisements. These consumption experiences and advertisements are added to the experience set yielding P^t.

3. Consumers are given a set of products to choose from with salient search characteristics X^i. They assign perceived characteristics to each product $V^i = A(P^t, X^i)$ independently as in the one period model.

4. Given the perceived characteristics of each good, consumers choose the consumption bundle x for the next period that maximizes expected future utility.[35]

One interesting thing to notice is that the experience set P^t is increasing in t. This means that the marginal effect of advertisements seen in period t on the network connections in period t is decreasing in t. Since advertisements only affect consumption decisions through these network connection, the decreasing marginal effect of advertising on connection strengths imply that advertisements seen at time t have a greater effect on period-t consumption than ads seen at time $t+1$ do on period $t+1$-consumption. Heuristically this means that advertisements will affect younger consumers' consumption decisions more than older consumers' and agrees with the fact that advertisers tend to target younger audiences.

Once we include consumption experiences into P, we must ask: how are social characteristics experienced during consumption? Consumers will get objective feedback on the experience attributes

[35] It will be easier to initially assume that all goods are perishable and consumers are completely myopic and so only maximize utility for the current period. This would allow us to consider the implications of evolving perceptions without also dealing with issues of dynamic programming.

of a product during consumption which will either reinforce or dilute a brand's connection to these attributes. However, there will be no such objective feedback for social characteristics. One way to deal with this question is to assume that if a social characteristic is activated during the evaluation of the product it will also be activated during its consumption.

Finally, the multi-period model raises another interesting question. Are consumers aware of the effects of advertisements? It has been suggested that if consumers are aware that certain stimuli will change their effective utility function, it may be in their best interest to avoid these stimuli. (45; 9) Given the opportunity, should consumers avoid or seek out certain advertisements, and if they should, will they? Considering the problem from this angle potentially has implications for the regulation of advertisements that are difficult to avoid, e.g., beer ads during the Super Bowl, or billboards on a highway.

1.7.4.2 Identity and Social Networks

In this paper, I have focused on attributes that are generally seen as positive by everyone. However, many social attributes are valued differently by different consumers. Some of this has to do with a consumer's ideal perception of himself, or his perception of himself as a member of some social grouping, as in Akerlof and Kranton's model of identity (2). Sub-group identifications can themselves be viewed as social attributes associated with and defined by sets of other attributes. Consumers identifying with these groups will have a high utility coefficient for the attribute representing the social grouping. So, for example, someone identifying themselves as "conservative" will tend to consume products that are associated with conservatism. Some of these may have a rational basis, like an American made car. But others simply gain from their association with conservatives and conservative causes. Country music is a good example of a product that often has no political content, but is still overwhelmingly consumed by conservatives. The political preferences of country listeners is illustrated by the fan reactions to the Dixie Chicks in the example from section 1.2. When the group made prominent liberal statements that were then covered repeatedly by the press, consumers developed a negative association between their music and conservatism, making them far less popular with the country music fan base.

Social networks, an emerging field in economics that borrows from sociology, (25; 19; 32; 31; 43), are a possible framework for understanding *how* certain product attributes become associated with particular social groups. Consumers who are closely connected in a social network are likely to have similar experience sets, and therefore, similar internal associative networks. Since these networks are formed endogenously, we could hypothesize that consumers with similar preferences (basically similar attribute utilities) will seek each other out — leading to the social groupings that individuals eventually identify with. Since experiences such as observing a friend consuming a product will affect a consumer's implicit associations, these network might explain trends where the conspicuous consumption decisions of a few members of some closely connected group of individual "spread" to the rest.

1.7.4.3 Complements

While the network structure outlined in this paper allows for complementarities between characteristics, our chosen utility function does not lead to complementarities between products. However, one could imagine cases where the attributes of different products might interact with a consumer's associations in the same way they would for a single product. For instance, separate pieces of clothing can interact to generate an image consistent with any number of different social groups.

The major hurdle for understanding complementarities between products in this framework is to understand which products are actually likely to be consumed simultaneously. This becomes particularly problematic when dealing with durable goods like clothing since they will be valued differently depending on the consumption context. When making a purchasing decision, consumers must consider the benefits from the good in every likely consumption context to assess its actual utility.

1.7.4.4 Determining the Consideration Set

This paper demonstrates some of the applications of the model to how advertising can effect the evaluation of a product, but advertising also serves the simple purpose of making a product a salient choice to the consumer. In the above example we assume that both firms' products are always salient and that assumption actually leads to very low overall levels of advertising because of the incentive to free-ride. Adding brand-recall as a requirement for purchase places an upward pressure on the advertising level.

In marketing the set of products actually consciously evaluated by the consumer is called the "consideration set". The overall choice set for a consumer is almost always larger than the consideration set. This set is generally going to be a function of the choice situation. In the simplest example if a consumer is in a particular store, the salient products will be those that they can actually see. For precisely this reason, shelf-space that is at eye-level is at a premium in stores.

We can use the same network used to evaluate products to determine the salient choice set for a consumer. We model the context effects as a "stimulus" to the consumer, which is simply a state of the network $V \in \mathcal{S}$. We say that a brand F^i is *recalled* if it is active in $A(P, V)$.

Definition 1.7.1 *A brand F^i is* recalled *in response to a stimulus V if $A_{F^i}(P, V) = 1$.*

For a product to be successful it needs to actually be in the consumer's choice set as often as possible. One way to do that would be to ensure decent shelf placement in all the relevant stores; this is precisely the strategy chosen by store-brand generics. However, there are many product types, such as clothing, prepared food, or cars, where brand recall must occur before the consumer even reaches the store. For these sorts of products it is important that a stimulus of the product class itself cue the recall of the brand.

In the network model advertisements not only strengthen the connection between a brand name and positive characteristics, but the connection between the brand and the type of product itself. This in turn increases the probability that the brand will be recalled when they consider a product type. For some brands, like Kleenex, Xerox, and Coca-Cola, this has happened to such a degree that the brand-names are practically synonymous with the product type.

1.7.4.5 Order effects of choice

In this model the initial information X network is exogenous and, and the evaluation of a choice is independent of the order of evaluation. However in many situations products are evaluated in rapid succession so that the evaluation of one actually influences the evaluation of the next.

In order to consider these sorts of effects we need to introduce the concept of *external* input to a node. Instead of setting the state of the network to X, we can think of a vector of external inputs I^E. This way we can consider the effects of these inputs I^E to some non-trivial state of the network, such as $A(X^1, P)$, the perceived characteristics of a previously evaluated good.

One possible effect of this would be the “persistence” of certain characteristics. This would be a particularly important effect for the experience attributes of products. Recall that unlike social attributes, we allow experience attributes to have some degree of self-connection ($w_{ii} > 0$). This means that in cases where the search characteristics of a product would not be sufficient to activate some experience attribute Y themselves, they may be sufficient to *sustain* this activation if Y had already been activated by the previous option. This effect is similar to the free-riding observed in section 1.6.1, except instead of free-riding off of another firm’s advertising, the product can free-ride directly off of another product’s *evaluation.*

This inertial effect would be another mechanism that generic products could take advantage of. As mentioned in section 1.5.3, store-brand generics often use similar packaging to their brand-name counterparts, but they are also almost always placed in close proximity to these counterparts to emphasize their similarities to each other.

1.7.4.6 Controlling the updating path

As noted in section 1.3, $A(P, X)$ is not necessarily a deterministic function. For initial vectors that are on the edge between attractor basins, the order in which characteristics are updated will affect the finally perceived characteristics. In general it is convenient to think of this updating sequence as random, however, there are cases where this sequence may be affected by external stimuli. This is most apparent in situations where the information X is given to the decision-maker in some fixed sequence. For example, when an employer is reading the resumes of potential employees, they scan information in a specific exogenously determined order.

To model this, it is once again useful to think of the information as an external input to a particular node. When a consumer receives information about a particular characteristic, the network updates the node corresponding to that characteristic — adding the external input I_i^E to the endogenously generated input I_i[36] to generate a total input I_i^t. The node is then updated according to this aggregate input.

One way to think about this is that each new piece of information is "checked" against what is already known, as represented by the state of the network V. This means that information received earlier will have precedence over new information. After all the information has been received, the network will be in some state V that will converge independently as described in section 1.3. However, since each new bit of information is checked against the previous pieces, V should not have too many pieces of contradictory information, i.e., the network energy should be low and V is more likely to be inside a basin of attraction.

This extension will be particularly relevant to discussing issues of stereotyping: for example, the order in which the characteristics of a job applicant are presented may change the degree to which implicit racial, ethnic, or gender biases will affect evaluations. For example, consider the job applicant scenario. Bertrand and Mullainathan showed in a field experiment that resumes with stereotypically black names were less likely to get interviews than identical resumes with stereotypically white names (10). This disparity occurs *despite* any institutional pressures that may exist to increase racial diversity at firms, suggesting that this is the result of implicit racial bias rather the explicit discriminatory intent. This extension of the neural network model implies that placing racial information at the end of a resume as opposed to the beginning might attenuate this bias.

1.7.5 Welfare Implications

In this model advertisements can be thought of as *creating* utility in some real way. We see this reflected in higher consumer demand for goods associated with a positive social attribute in section 1.6.1. This could be seen as a positive externality to advertising. However, there has been a great deal of criticism of advertising for creating *need*, so much so that it is taken as given that advertising has the capacity to create desire for what some would call trivial or unimportant goods.

Economists and philosophers have debated this question for some time. The most common criticisms of capitalist societies stem from the assumption that the market, and society itself, creates desire for what critics think of as meaningless commodities. Many of these criticisms have come from various Marxist philosophers, but the value of increasing levels of consumption have been questioned

[36] Recall from section 1.3, the input vector $I = W(P) \cdot V$, where $W(P)$ is the weight matrix of the network, and V is the current state of the network.

by economists as well.

> Few economists in recent years can have escaped some uneasiness over the kinds of goods which their value system is insisting they must maximize. They have wondered about the urgency of numerous products of great frivolity. They have been uneasy about the lengths to which it has been necessary to go with advertising and salesmanship to synthesize the desire for such goods. [37] That uneasiness has reflected the crucial weakness of the literature on this point. (John Kenneth Galbraith *The Affluent Society*, p. 116) (23)

Thus far, our exploration of the neural-network model has displayed how advertising can increase demand for a product by associating it with a positive social attribute. But it is not obvious that ads might have negative repercussions. An implicit assumption of those who criticize the advertising industry is that while ads create need, they also somehow have the intertemporal effect of decreasing the utility of a consumer's pre-advertising consumption bundle, i.e., advertising creates the constant need to "keep up with the Jones's" on a massive scale.

An interesting property of these networks is that if a pattern V is stable, so is its opposite, $-V$, so that these advertisements may create negative externalities as well. If the salient presence of some set of search attributes creates utility from a social attribute Y, such as status, their salient absence can create negative utility. So if, for example, a particular brand, or a type of product, is positively associated with some characteristic Y, its *salient absence* will create negative input to Y; therefore advertising has the potential for creating negative as well as positive utility.

These sorts of negative externalities are desirable to firms since they increase the marginal utility of their product: not only do you gain utility from a social attribute, but you get rid of the negative utility generated by its salient absence. However, these negative externalities are distinctly undesirable to the firm's competitors and potentially to consumers. Products associated with Y will generally be more expensive than those without, making them outside the price range of some set of the population. However, if advertisement makes the lack of Y salient, the utility of the low-Y products will actually decrease relative to their pre-advertising evaluations.

1.7.6 Conclusion

I have laid out an expansion of the hedonic pricing model where advertisements create complementarities among the search attributes of a good — in essence changing the attributes of the product as seen by the consumer. This change allows us to model advertising's effect on the marginal utility of a product while maintaining a stable underlying utility function like the complementary model proposed by Becker and Murphy.

[37] Once again, we see a reference to advertisers manipulating the demand and perceptions from decades ago. However, despite this early concern, there have been very few models that attempt to address the mechanisms of advertising.

Advertisements operate by modulating the associations among attributes, allowing the tangible features of a product to imply abstract cultural attributes such as a piece of clothing being fashionable. Attributing such social characteristics to a product will often change a consumer's utility for a product.

I use a network model, based on standard models from cognitive psychology and neuroscience, to model how advertisements actually change the associations among attributes and how these associations interact to affect the consumer's perception of the product.

Unlike Becker and Murphy's model, the complementarities we consider are between attributes, such as a brand name and a social attribute, instead of between the products themselves. This allows us to consider not only the effects of an advertisement on a consumer's evaluation of the advertised product, but also how the changes in attribute complementarities may affect other products (the externalities generated by an advertisement). We roughly divide these externalities into two classes, spillovers and dilution. The interconnected nature of the network model means that the externalities are almost impossible to avoid. If I firm wants to diversify its product over a social or experience attribute, it is often forced to dilute the useful associations it built up in the past. On the other hand, when a firm advertises it creates positive associations that may benefit its competitors.

This framework is about fast, automated processing (System I in the language of Kahneman (34)). The implicit associations among attributes that develop over a host of experiences will affect the way we process the information we receive about a good. Since this is a System I model, it is most appropriately applied to low-involvement situations or those where inference is either difficult or impossible. This is why I have focused mostly on what I call social attributes. Since these are in essence culturally agreed upon and impossible to assess objectively.

Even though this is a System I model, the associations among nodes contain information. They keep a record of the correlations between attributes, which could be particularly useful when considering experience characteristics in a multi-period version of the model. So while I have talked about these networks in relation to the complementary view of advertising, it has some relation to informative models as well.

As with any model, there is an inherent tension between descriptive power and parsimony. This model lacks the parsimony of Nelson's informative model of advertising where the only important feature of advertising is its cost, or Shapiro's model where consumers misremember advertisements as consumption experiences. But, it allows us to use a single framework to describe a host of phenomena relating to advertising including product differentiation, product blurring, the importance of frequency and the role of trademarks. While there are a large number of free parameters in this model, I have proposed a number of methods to estimate these parameters using data and existing psychological methods.

Since I have used a set of tools new to the economic literature I have only looked at networks with very few product characteristics. However, even with this limited set of attributes, we see a number of interesting phenomena related to the general categories of dilution and spillover. I hope that these examples serve as a proof of principle that formal associative networks can be usefully applied to economic problems.

Bibliography

[1] Aaker, D. A., and Keller, K. L. Consumer evaluations of brand extensions. *Journal of Marketing 54*, 1 (1990), 27–41.

[2] Akerlof, G. A., and Kranton, R. E. Economics and identity. *The Quarterly Journal of Economics 115*, 3 (2000), 715–733.

[3] Armstrong, M. Multiproduct nonlinear pricing. *Econometrica 64*, 1 (1996), 51–75.

[4] Bagwell, K. The economic analysis of advertising. www.columbia.edu/%7Ekwb8/-papers.html, 2005.

[5] Bain, J. S. *Barriers to New Competition: Their Character and Consequenes in Manufacturing Industries*. Harvard University Press, Cambridge, 1956.

[6] Bargh, J., Chaiken, S., Raymond, P., and Hymes, C. The automatic evaluation effect: Unconditional automatic attitude activation with a pronunciation task. *Journal of Experimental Social Psychology 32* (1996), 104–128.

[7] Becker, C. A. Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory and Cognition 8* (1980), 493–512.

[8] Becker, G. S., and Murphy, K. M. A simple theory of advertising as a good or bad. *The Quarterly Journal of Economics 108* (1993), 941–964.

[9] Bernheim, D., and Rangel, A. Addiction and cue-triggered decision processes. *American Economic Review 94*, 5 (2004), 1558–1590.

[10] Bertrand, M., and Mullainathan, S. Are emily and greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *American Economic Review 94*, 4 (2004), 991–1013.

[11] Boush, D. M., and Loken, B. A process-tracing study of brand-extension evaluation. *Journal of Marketing Research 28* (1991), 16–28.

[12] Broniarczyk, S. M., and Gershoff, A. D. The reciprocal effects of brand equity and trivial attributes. *Journal of Marketing Research 11* (2003), 161–175.

[13] Brown, C. L., and Carpenter, G. S. Why is the trivial important? a reasons-based account for the effects of trivial attributes on choice. *Journal of Consumer Research 26* (2000), 372–385.

[14] Burger wars: Jack in the box sued over ad — competitor has a beef over angus commercial: 'they're not being funny'. Associated Press, www.msnbc.com, May 29 2007. www.msnbc.msn.com/id/18894390/.

[15] Carpenter, G. S., Glazer, R., and Nakamoto, K. Meaningful brands from meaningless differentiation. *Journal of Marketing Research 31* (1994), 339–350.

[16] Collins, A. M., and Loftus, E. A spreading activation theory of semantic processing. *Psychological Review 82*, 6 (1975), 407–428.

[17] Comanor, W. S., and Wilson, T. A. *Advertising and Market Power.* Harvard University Press, Cambridge, 1974.

[18] Dacin, P. A., and Smith, D. C. The effect of brand portfolio characteristics on consumer evaluations of brand extensions. *Journal of Marketing Research 31* (1994), 229–242.

[19] Demange, G., and Warwick, M., Eds. *Group Formation in Economics.* Cambridge University Press, 2005.

[20] Dixie chicks pulled from air after bashing bush. Reuters, www.cnn.com, March 14 2003. www.cnn.com/2003/SHOWBIZ/Music/03/14/dixie.chicks.reut/.

[21] Dixie chicks nix concert tour dates. CB Music message board, 2006. www.cinemablend.com/-music/Dixie-Chicks-Nix-Concert-Tour-Dates-731.html.

[22] Fehr, E., and Schmidt, K. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics 114* (1999), 817–68.

[23] Galbraith, J. K. *The Affluent Society: Fortieth Anniversary Edition.* Houghton Mifflin, New York, 1998.

[24] Gilovich, T. Seeing the past in the future: The effect of associations to familiar events on judgments and decisions. *Journal of Personality and Social Psychology 40* (1981), 797–808.

[25] Goeree, J. K., Riedl, A., and Ule, A. In search of stars: Network formation among heterogeneous agents. *Iza Discussion Paper Series 1754* (2005).

[26] Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology 75*, 6 (1998), 1464–1480.

[27] Gürhan-Canli, Z., and Maheswaran, D. The effects of extensions on brand name dilution and enhancement. *Journal of Marketing Research 35* (1998), 464–473.

[28] Hamann, S., and Squire, L. Intact perceptual memory in the absence of conscious memory. *Behavioral Neuroscience 111* (1997), 850–854.

[29] Hebb, D. O. *The organization of behavior: A neuropsychological theory.* Wiley, New York, 1949.

[30] Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Procedures of the National Academy of Sciences 79* (1982), 2554–2558.

[31] Jackson, M. O., and Rogers, B. W. The economics of small worlds. *Journal of the European Economic Association 3* (2005), 617–627.

[32] Jackson, M. O., and Yariv, L. Diffusion of behavior and equilibrium properties in network games. *American Economic Review 97*, 2 (2007), 92–98.

[33] Jana, R. Bad for the brand? here's a look at seven brand extensions that are a real stretch. *BusinessWeek Online* (2006). images.businessweek.com/ss/06/07/-brand_extensions/source/1.htm.

[34] Kahneman, D. Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review 93*, 5 (2003), 1449–1475.

[35] Kahneman, D., and Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica 47*, 2 (1979), 263–292.

[36] Kaldor, N. V. The economic aspects of advertising. *Review of Economic Studies 18* (1950), 1–27.

[37] Kardes, F., and Allen, C. Perceived variability and inferences about brand extensions. In *Advances in Consumer Research*, R. H. Holman and M. R. Solomon, Eds. Association for Consumer Research, Provo, Utah, 1991, pp. 392–398.

[38] Keller, K. L. Conceptualizing, measuring, and managing customer-based brand equity. *Journal of Marketing 57* (1993), 1–22.

[39] Keller, K. L., and Aaker, D. A. The effects of sequential introduction of brand extensions. *Journal of Marketing Research 29* (1992), 25–50.

[40] Keller, K. L., and Sood, S. Brand equity dilution. *MIT Sloan Management Review 45*, 1 (2003), 12–15.

[41] Klauer, K. C., and Musch, J. Goal-dependent and goal-independent effects of irrelevant evaluations. *Personality and Social Psychology Bulletin 28*, 6 (2002), 802–814.

[42] Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., and D'Friederici, A. Music, language and meaning: brain signatures of semantic processing. *Nature Neuroscience 7*, 3 (2004), 302–307.

[43] Kosfeld, M. Economic networks in the laboratory: A survey. *Review of Network Economics 3*, 1 (2004), 20–42.

[44] Laibson, D. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics 112*, 2 (1997), 443–477.

[45] Laibson, D. A cue-theory of consumption. *Quarterly Journal of Economics 116*, 1 (2001), 81–119.

[46] Lancaster, K. A new approach to consumer theory. *Journal of Political Economy 74* (1966), 132–157.

[47] Landauer, T., Folts, P. W., and Laham, D. Introduction to latent semantic analysis. *Discourse Processes 25* (1998), 259–284.

[48] Levy, D. A., Stark, C. E. I., and Squire, I. R. Intact conceptual priming in the absence of declarative memory. *Psychological Science 15*, 10 (2004), 680–686.

[49] Loken, B., and John, D. R. Diluting brand beliefs: When do brand extensions have a negative impact? *Journal of Marketing 57* (1993), 71–84.

[50] Maclennan, D. *Housing Economics*. Longman, 1982.

[51] McAfee, R. P., McMillan, J., and Whinston, M. D. Multiproduct monopoly, commodity bundling and correlation of values. *Quarterly Journal of Economics 104*, 2 (1989), 371–383.

[52] McGurk, H., and MacDonald, J. Hearing lips and seeing voices. *Nature 264* (1976), 746–478.

[53] Meyvis, T., and Janiszewski, C. Consumers' beliefs about product benefits: The effect of obviously irrelevant product information. *Journal of Consumer Research 28* (2002), 618–635.

[54] Milberg, S. J., Park, C. W., and McCarthy, M. S. Managing negative feedback effects associated with brand extensions: The impact of alternative branding strategies. *Journal of Consumer Psychology 6*, 2 (1997), 119–140.

[55] Milgrom, P., and Roberts, J. Price and advertising signals of product quality. *The Journal of Political Economy 94*, 4 (1986), 796–821.

[56] Morrin, M. The impact of brand extensions on parent brand memory structures and retrieval processes. *Journal of Marketing Research 36* (1999), 517–525.

[57] Morrin, M., and Jacoby, J. Trademark dilution: Empirical measures for an elusive concept. *Journal of Public Policy and Marketing 19*, 2 (2000), 265–276.

[58] Mullainathan, S., Schwartzstein, J., and Shleifer, A. Coarse thinking and persuasion. Working Paper, 2006.

[59] Nelson, P. Advertising as information. *The Journal of Political Economy 82*, 4 (1974), 729–754.

[60] Nelson, S. B. Hebb and anti-hebb meet in the brainstem. *Nature Neuroscience 7*, 7 (2004), 687–688.

[61] Quillian, M. A revised design for an understanding machine. *Mechanical Translation 7* (1962), 17–29.

[62] Quillian, M. R. Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science 12* (1967), 410–430.

[63] Rabin, M. Incorporating fairness into game theory. In *Advances in Behavioral Economics*, C. F. Camerer, G. Loewenstein, and M. Rabin, Eds. Princeton University Press, Princeton, 2004, pp. 297–325.

[64] Ries, A., and Ries, L. *The 22 Immutable Laws of Branding*. Collins Business, New York, 2002.

[65] Robinson, J. *The Economics of Imperfect Competition*. MacMillan and Co., London, 1933.

[66] Rosen, S. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy 82*, 1 (1974), 34–55.

[67] Seldon, B. J., Banerjee, S., and Boyd, R. G. Advertising conjectures and the nature of advertising competition in an oligopoly. *Manegerial and Decision Economics 14*, 6 (1993), 489–498.

[68] Shapiro, J. M. Fooling some of the people some of the time: Advertising and consumer memory. Harvard Mimeo, 2005.

[69] Shepard, A. Price discrimination and retail configuration. *Journal of Political Economy 99*, 1 (1991), 30–53.

[70] Stine, G. Branding truth #6: Avoid sub-brands at all costs. *Venturer Newsletter June* (2004).

[71] Stine, G. Branding truth #7: Perception vs. quality. *Venturer Newsletter September* (2004).

[72] Storbeck, J., and Robinson, M. D. Preferences and inferences in encoding visual objects: A systematic comparison of semantic and affective priming. *Personality and Social Psychology Bulletin 30*, 1 (2004), 81–93.

[73] Trueswell, J. C. The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language 35* (1996), 566–585.

[74] Weibull, J. Price competition and convex costs. Stockholm School of Economics, Working Paper No. 622, 2006.

Appendix A

A.1 Convergence of networks to stable states

The classical Hopfield network is a binary network where each node has a single threshold α_i, rather than two thresholds, as in this model. For these types of networks there is a simple quadratic energy function

$$E(V) = -\frac{1}{2}\sum_{i,j} w_{ij}V_iV_j + \sum_i x_i\alpha_i$$

that is strictly decreasing under asynchronous updating.

These models assume that each node is salient to each pattern and that the network always has "complete information," i.e., each node is at either 1 or -1. To apply these associative networks to economics it was important to introduce the neutral 0 state for unknown or non-salient nodes. This does several things, including introducing the zero state as a permanent and strong attractor of the network. For the purposes of showing convergence, the new model introduces an extra threshold which the energy function must now deal with. We define a new energy function where the threshold itself depends on the state of the network:

$$E(V) = -\sum_i w_{ii}V_i(t) - \frac{1}{2}\sum_{i\neq j} w_{ij}V_i(t)V_j(t) + \sum_i V_i(t)V_i(t)\alpha_i$$

This essentially states that the *relevant* threshold for the energy function depends on the state of the node. If $V_i = -1$, the relevant threshold is $-\alpha_i$ since passing this threshold will change the state of the node. Conversely, if $V_i = 1$ the relevant threshold is α_i. When the state of the node is at 0 both thresholds are relevant, but if the state of the node were to *change* the only important fact is the *sign* of the input, so calculating the energy with a 0 threshold is sufficient.

Now we show that this energy is in fact decreasing under the transition function. If the node remains unchanged, the energy is stable, so we will only consider cases where the updated node changes.

Since we are updating nodes one at a time we can isolate the portion of the energy function pertaining to the ith node. We use the fact that W is symmetric to simplify the expression:

$$\Delta E = \Delta - V_i\left(w_{ii}V_i + \sum_{j\neq i} w_{ij}V_j - V_i\alpha_i\right) = \Delta - V_i(I_i - V_i\alpha_i)$$

where I_i is the input to the ith node.

- Case 1: $V_i(t) \neq 1$, $V_i(t+1) = -V_i(t)$. In this case the bit actually flips. Without loss of generality, let $V_i(t) = 1$. Since the bit flips we know that $I_i(t) < -\alpha_i$. So $-V_i(t)(I_i(t) - \alpha_i V_i(t)) > 0$. After the bit flips $I_i(t+1) - I_i(t) = -w_{ii} - w_{ii} < 0$ so $I_i(t+1) < -\alpha_i$ as well. That means that $-V_i(t+1)(I_i(t+1) + \alpha_i) < 0$ and $\Delta E < 0$ as desired.

- Case 2: $V_i(t) \neq 0$, $V_i(t+1) = 0$. Once again assume without loss of generality that $V_i(t) = 1$. $V_i(t+1) = 0$ so $I_i(t) < \alpha_i$ and $-V_i(t)(I_i(t) - \alpha_i V_i(t)) > 0$. After the transition $-V_i(t+1)(I_i(t+1) - V_i(t+1)\alpha_i) = 0$ so $\Delta E < 0$ again.

- Case 3: $V_i(t) = 0$, $V_i(t+1) \neq 0$.

  $$V_i(t)(I_i(t) - V_i(t)\alpha_i) = 0.$$

  $V_i(t+1) \neq 0$ so $|I_i(t)| > \alpha_i$ and $sign(V_i(t+1)) = sign(I_i(t))$,

  $$I_i(t+1) = I_i(t) + w_{ii}sign(I_i(t)) \Rightarrow |I_i(t+1)| > \alpha_i.$$

  Taken together we get

  $$-V_i(t+1)(I_i(t+1) - V_i(t+1)\alpha_i) < 0 \Rightarrow \Delta E < 0$$

  .

This shows that if $V_i \neq T_i^i(V_i)$ $\Delta E < 0$. Since the state space is finite this ensures that the energy of the network must reach a minimum, and that this minimum must be a stable state.

A.2 Conditions on advertising and entry in oligopoly — from section 1.6.1

A.2.1 Conditions on price

Claim A.2.1 *If the monopoly profit functions for the high-Y and low-Y goods, $\pi_m^h(q)$ and $\pi_m^l(q)$ are quasi-concave, advertising will only be possible in a pure strategy equilibria if the monopoly price* *for the high-Y demand curve, q_m^h is greater than q^* where* $q^* = \inf\{q|q - c'\left(\frac{D_l(q)}{n}\right) > 0\}$.

Proof:

Mathematically condition (1) implies that

$$\pi_o^h(q) \geq \pi_m^h(q)$$

Basically, the increased cost of producing for the entire market must exceed the extra revenue from sales. Otherwise, a firm would profit by slightly undercutting its competitors.

Weibull establishes that such prices will generally exist if the monopolistic profit function π_m^h, is quasi-concave (74),[1] and achieves its maximum at some finite $q_m^h > 0$. He further shows that this can only occur at $q \leq q_m^h$, establishing an upper bound on the equilibrium price, $q < q_m^h$.

We can derive the lower bound from condition (2), which is equivalent to:

$$\pi_o^h(q) \geq \left\{ \begin{array}{c} \pi_m^l(q^*) + c_A \\ \pi_o^l(q) + c_A \end{array} \right. ,$$

where $q^* = \min(q, q_m^l)$, and q_m^l is the monopoly price for the low-Y good. The first part of this condition states that the advertising firm has no incentive to deviate by dropping its advertisement and monopolizing the market, while the second states that it has no incentive to deviate by dropping its advertisement and sharing the market at price q.

Since costs are convex this second part of the second condition gives us a lower bound on q. To see this, first rearrange and note that $c_A > 0$ to get:

$$q\left(\frac{D_h(q)}{n} - \frac{D_l(q)}{n}\right) > c\left(\frac{D_h(q)}{n}\right) - c\left(\frac{D_l(q)}{n}\right).$$

Since c is convex and $D_l(q) \leq D_h(q)$ for all q we can put a lower bound on the right side of this inequality:

$$\begin{array}{rcl} c\left(\frac{D_h(q)}{n}\right) - c\left(\frac{D_l(q)}{n}\right) & > & c'\left(\frac{D_l(q)}{n}\right)\left(\frac{D_h(q)}{n} - \frac{D_l(q)}{n}\right) \Rightarrow \\ q & > & c'\left(\frac{D_l(q)}{n}\right) \Rightarrow \\ q - c'\left(\frac{D_l(q)}{n}\right) & > & 0 \end{array}$$

Taking the derivative of the function $q - c'\left(\frac{D_l(q)}{n}\right)$ we see that it is strictly increasing, so we get a lower bound on possible equilibrium prices $q > \inf\{x|x - c'\left(\frac{D_l(x)}{n}\right) > 0\} = q^*.Q.E.D.$

[1] A function is quasi-concave if all its upper contour sets (sets of the form $\{x : f(x) > \lambda\}$) are convex. This includes any single-peaked or monotonic function.

Claim A.2.2 *For demand and cost functions of the form:*

$$\begin{aligned} \beta_h &< \beta_l \\ D_h(q) &= 1-\beta_h q \\ D_l(q) &= 1-\beta_l q \\ c(x) &= c\cdot x^2 \end{aligned}$$

For any equilibrium where advertising occurs the range of prices q bounded:

$$\frac{c+\sqrt{c^2+\frac{c_A n^2(n+c(\beta_h+\beta_l))}{\beta_l-\beta_h}}}{n+c(\beta_h+\beta_l)} \le q \le \frac{c(n+1)}{c\beta_h(n+1)+n}$$

Proof:

The first condition is $\pi_o^h(q) \ge \pi_m^h(q')$ This means that

$$\begin{aligned} \pi_o^h(q)-\pi_m^h(q) &\ge 0 \;\Rightarrow \\ (1-\beta_h q)\frac{n-1}{n}\left(c\cdot(1-\beta_h q)\frac{n+1}{n}-p\right) &\ge 0 \end{aligned}$$

$\pi_o^h(q)-\pi_m^h(q)=0$ at $q^*=\frac{1}{\beta_h}$ and $q_*=\frac{c(n+1)}{c\beta_h(n+1)+n}$. Notice that $\pi_o^h(q)-\pi_m^h(q)$ is convex in q, so this function will be weakly positive everywhere outside the interval (q_*, q^*). Demand is 0 at prices $q>q^*$, so we throw out prices above this interval and we know that any equilibrium price q, is less than q_*:

$$q \le \frac{c(n+1)}{c\beta_h(n+1)+n}.$$

Second, we need $\pi_o^h(q) \ge \pi_o^l(q)+c_A$. So,

$$\begin{aligned} \pi_o^h(q)-\pi_o^l(q) &\ge c_A \;\Rightarrow \\ \frac{q^2}{n}(\beta_l-\beta_h)-\frac{c}{n^2}[(1-\beta_h q)^2-(1-\beta_l q)^2] &\ge c_A \;\Rightarrow \\ \frac{q(\beta_l-\beta_h)}{n^2}(n+c(\beta_h+\beta_l)q-2c) &\ge c_A \end{aligned}$$

Notice that $\pi_o^h(q)-\pi_o^l(q)$ is convex in q, and decreasing at $q=0$. This means that the upper solution to $\pi_o^h(q)-\pi_o^l(q)-c_A=0$ forms the lower bound for q. Using the quadratic formula we get

$$q \ge \frac{c+\sqrt{c^2+\frac{c_A n^2(n+c(\beta_h+\beta_l))}{\beta_l-\beta_h}}}{n+c(\beta_h+\beta_l)}.$$

Q.E.D.

Claim A.2.3 *In any equilibrium with non-empty low- and high-Y markets, the advertising level for firms in the high-Y market must fall within a specified range:*

$$\frac{\pi_m^h(q_m^h(q_l),q_l)-\frac{q_h}{k}D_h+c\left(\frac{D_h}{k}\right)}{c_A\left\lceil\frac{1-k(\alpha-1)}{\alpha-1}\right\rceil}\leq N\leq\frac{2-\alpha}{1-k(\alpha-1)}.$$

First, recall that the advertising level must be minimal. This just comes down to asking if a firm can drop an advertisement and still be associated with Y. For this to be true we need $\frac{N-1}{kN-1}<\alpha-1$ and $k<\frac{1}{\alpha-1}$. These inequalities reduce to

$$N(1-k(\alpha-1))<2-\alpha\Rightarrow N<\frac{2-\alpha}{1-k(\alpha-1)},$$

giving us an upper bound on the number of advertisements that is increasing in k.

On the other hand, if a firm unilaterally raises the advertising level they will either push out *all* of the other advertising firms or none. The later case is clearly not optimal for the firm since the extra advertising has no effect on revenue, but will increase costs. So we only need consider the first case where the firm raises the advertising level enough to monopolize the high-Y market. For N' to be large enough ensure a monopoly over high-Y goods $N'-N>\frac{1-(\alpha-1)k}{\alpha-1}N$. This inequality implies that the closer m is to the capacity $\frac{1}{\alpha-1}$ of the high-Y market, the easier it is to push out the other firms. Set N' at the smallest such integer. To ensure that monopolizing the high-Y market won't be profitable we must have:

$$\pi_m^h(q_m^h(p_l),p_l)-\left\lceil\frac{1-(\alpha-1)k}{\alpha-1}N\right\rceil c_A\leq D_h\frac{p_h}{k}-c\left(\frac{D_h}{k}\right),$$

giving us a lower bound on the advertising level N.

Taken together with the condition above we can constrain the range of possible advertising levels N:

$$\frac{\pi_m^h(q_m^h(p_l),p_l)-\frac{p_h}{k}D_h+c\left(\frac{D_h}{k}\right)}{c_A\left\lceil\frac{1-k(\alpha-1)}{\alpha-1}\right\rceil}\leq N\leq\frac{2-\alpha}{1-k(\alpha-1)}$$

as desired. *Q.E.D.*

A.2.2 Conditions on k

First we calculate the costs of entry into the high-Y market. The firm can either simply purchase enough ads to enter the high-Y market and split the high-Y market with the other firms, or it can purchase enough ads to actually monopolize the high-Y market, pushing all competing firms out. The lowest N' to simply enter the high-Y market is $\left\lceil\frac{k(\alpha-1)}{2-\alpha}N\right\rceil$. The lowest N' needed to gain a monopoly over the high-Y market is $\left\lceil\frac{1-(\alpha-1)k}{\alpha-1}N\right\rceil$.

So the cost $c_e(k,N)$ of entry to the high-Y market is

$$c_e(k,N) = \left\lceil \frac{k(\alpha-1)}{2-\alpha} N \right\rceil c_A$$

and the cost $c_m(k,N)$ of gaining a monopoly over the high-Y market is

$$c_m(k,N) = \left\lceil \frac{1-(\alpha-1)k}{\alpha-1} N \right\rceil c_A$$

The benefits of entry and monopoly are

$$\begin{aligned} b_e(k,p_h,p_l) &= \frac{p_h}{k+1} D_h - \frac{p_l}{n-k} D_l + c\left(\frac{D_l}{n-k}\right) - c\left(\frac{D_h}{k+1}\right) \\ b_m(k,p_l) &= p_m^h(p_l) D_h(q_m^h(p_l), p_l) - \frac{p_l}{n-k} D_l + c\left(\frac{D_l}{n-k}\right) \end{aligned}$$

So the no-entry condition reduces to

$$\begin{aligned} b_e(k,p_h,p_l) &\leq c_e(k,N) \\ b_m(k,p_l) &\leq c_m(k,N) \end{aligned}$$

Notice that when $k > \frac{1}{\alpha-1} - 1$, i.e., when it is already at capacity, the entry option is irrelevant. A firm must either choose to monopolize the high-Y market or stay in the low-Y market. Recall that we know from the condition above that $N \leq \frac{2-\alpha}{1-k(\alpha-1)}$, so we can get necessary conditions on k that don't involve the advertising level N, specifically

$$\begin{aligned} \left\lceil \frac{k(\alpha-1)}{1-k(\alpha-1)} \right\rceil c_A &> b_e(k,p_l,p_h) \\ \left\lceil \frac{2-\alpha}{\alpha-1} \right\rceil c_A &> b_m(k,p_l) \end{aligned}$$

Chapter 2

Self-referential thinking and equilibrium as states of mind in games: fMRI evidence

with Colin F. Camerer

2.1 Introduction

Game theory has become a basic paradigm in economics and is spreading rapidly in political science, biology, and anthropology. Because games occur at many levels of detail (from genes to nations), game theory has some promise for unifying biological and social sciences [27].

The essence of game theory is the possibility of strategic thinking: Players in a game can form beliefs about what other players are likely to do, based on the information players have about the prospective moves and payoffs of others (which constitute the structure of the game). Strategic thinking is central to game theory, but is also important in market-level phenomena like signaling, commodity and asset market information aggregation, and macroeconomic models of policy setting.

Despite the rapid spread of game theory as an analytical tool at many social levels, very little is known about how the human brain operates when thinking strategically in games. This paper investigates some neural aspects of strategic thinking using fMRI imaging. Our eventual goal is to build up a behavioral game theory that predicts how players choose and the neural processes that occur as they play. The data can also aid neuro-scientific investigations of how people reason about other people and in complex strategic tasks.

In our experiments, subjects' brain activity is imaged while they play eight 2-player matrix games which are "dominance-solvable"[1] – that is, iterated deletion of dominated strategies (explained further below) leads to a unique equilibrium in which players' beliefs about what other players will

[1] In a dominance-solvable games, if players do not play dominated strategies, and guess that others will not, iteratively, then the result is an equilibrium configuration of strategy choices by players, and beliefs about what others will do, which are mutually consistent.

do are accurate and players best respond to their beliefs. (In equilibrium, nobody is surprised about what others actually do, or what others believe, because strategies and beliefs are synchronized, presumably due to introspection, communication or learning.)

The subjects perform three tasks in random orders: They make choices of strategies (task C); they guess what another player will choose ("beliefs", task B); and they guess what other players think they will choose ("2^{nd}-order beliefs", task $2B$). Every player being scanned plays for money with another subject who is outside of the scanner.

In a game-theoretic equilibrium, beliefs are correct, and choices are optimal given beliefs. One way for the brain to reach equilibrium is for neural activity in the C, B, and $2B$ tasks to be similar, since at equilibrium all three tasks "contain" the others, i.e. choice is a best response to belief, so the choice task invokes a belief formation. Any difference in activation across the three conditions is suggestive that different processes are being used to form choices and beliefs. In fact, as we show below, in experimental trials in which choices and beliefs are in equilibrium, there is little difference in activity in making a choice and expressing a belief; so this provides a purely neural definition of equilibrium (as a "state of mind"). Differences in activity across the three tasks might help us understand why players are out of equilibrium, so these differences are the foci of most of our analyses.

The first focus is the difference between making a choice and expressing a belief (i.e., the comparison between behavior and fMRI activation in the C and B conditions). If choices are best-responses to beliefs, then the thinking processes underlying choice and belief formation should highly overlap; choice and belief are like opposite sides of the same coin. (Put differently, if you were going to build brain circuitry to make choices and form beliefs, and wanted to economize on parts, then the two circuits would use many shared components.)

In contrast, disequilibrium behavioral theories that assume limited strategic thinking allow players to choose without forming a belief, per se, so that C and B activity can differ more significantly. For example, Camerer, Ho and Chong [9, 8] present a theory of limited strategic thinking in a cognitive hierarchy (building on earlier approaches [45, 54, 17, 35, 4]). In their theory some "0-step" players just choose randomly, or use some algorithm which is thoughtful but generates random choice – in any case, they will spend more energy on choice than belief. "One-step" thinkers act as if they are playing 0-step players, so they compute a choice but do not think deeply while forming a belief (e.g., they do not need to look at the other player's payoffs at all since they do not use these to refine their guess about what others will do). Two-step players think they are playing a mixture of 0- and 1-step players; they work harder at forming a belief, look at other players' payoffs, and use their belief to pick an optimal choice. Models of this sort are precise (more statistically precise than equilibrium theories) and fit most experimental data sets from the first period of a game (before learning occurs) better than Nash equilibrium does [9]. These limited-thinking theories allow larger differences in cognitive activity between the acts of *choosing* a strategy and *expressing a belief* about another players' strategy than equilibrium theories do. A 1-step player, for example, will look at all of her own payoffs and calculate the highest average payoff when making a choice, but when guessing what strategy another player with choose she can just guess randomly. Such a player will do more thinking when choosing than when stating a belief. This possible difference in processing motivates our analysis of differential brain activity during the C and B tasks.[2]

The second focus of the analysis is on the difference in activity while forming beliefs in the B task and 2^{nd}-order beliefs in the $2B$ task. One way agents might form 2^{nd}-order beliefs is to use general circuitry for forming beliefs, but apply that circuitry as if they were the other player (put themselves in the "other player's brain"). Another method is self-referential: Think about what they would like to choose, and ask themselves if the other player will guess their choice or not. These two possibilities suggest, respectively, that the B and $2B$ conditions will activate similar regions, or that the C and $2B$ regions will activate similar regions.

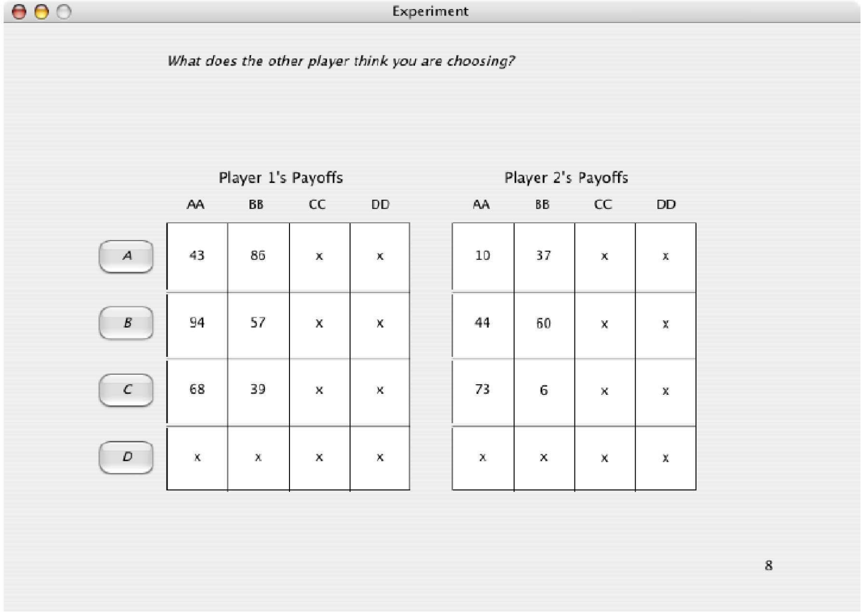
Besides contributing to behavioral game theory [6], imaging the brain while subjects are playing games can also contribute to basic social neuroscience [1]. Cognitive social neuroscientists are interested in spectrum disorders[3] like autism, in which people lack a normal understanding of what other people want and think. The phrase "theory of mind" (ToM) describes neural circuitry that enables people to make guesses about what other people think and desire (sometimes called "mind-reading" or "mentalizing" [51, 25, 52]).

Using game theory to inform designs and generate sharp predictions can also provide neuroscientists interested in ToM and related topics with some new tools which make clear behavioral predictions and link tasks to a long history of careful theory about how rational thinking relates to behavior.

In this spirit, our study extends ToM tasks to include simple matrix games. While there has been extensive research into first order beliefs: the simple consideration of another person's beliefs, there has been very little investigation of 2^{nd}-order beliefs, especially when they are self-referential i.e., what goes on in a person's brain when they are trying to guess what another person thinks *they* will do?

[2] An ideal test would compare activity of subjects who are capable of performing different thinking steps across games of different complexity. For example, a low-step thinker should show similar activity in simple and complex games (because they lack the skill to think deeply about complex games). A high-step thinker would stop at a low-level choice in a simple game (where k and higher steps of thinking prescribe the same choice) but would do more thinking in complex games. Unfortunately, we have not found a solid psychometric basis to "type-cast" players reliably into steps of thinking; when we can do so, the comparison above will provide a useful test.

[3] A "spectrum" disorder is one which spans a wide range of deficits (inabilities) and symptoms – it has relatively continuous gradation. This suggests a wide range of neural circuits or developmental slowdowns contribute to the disorder, rather than a single cognitive function.



Experiment

What does the other player think you are choosing?

	Player 1's Payoffs				Player 2's Payoffs			
	AA	BB	CC	DD	AA	BB	CC	DD
A	43	86	x	x	10	37	x	x
B	94	57	x	x	44	60	x	x
C	68	39	x	x	73	6	x	x
D	x	x	x	x	x	x	x	x

8

Figure 2.1: A three-step game used in the experiment, as presented in the scanner (game 3). C is dominated. Deleting C makes AA dominated. Deleting AA and C makes A dominant. The unique Nash equilibrium is therefore (A,BB). Only 31% and 61% (respectively) chose these strategies (see Appendix). The Camerer-Ho CH model (see text) with $\tau = 1.5$ predicts 7% and 55%.

2.1.1 Why study choices, beliefs and 2^{nd} order beliefs?

Figure 2.1 shows the exact display of a matrix game (our game 3) that row players saw in the scanner, in the $2B$ task where they are asked what the column player thinks they will do.[4] The row and column players' payoffs are separated onto the left and right halves of the screen (in contrast to the usual presentation).[5] Row payoffs are in a submatrix on the left; column player payoffs are in a submatrix on the right (which was, of course, explained to subjects).

The Figure 2.1 game can be solved (that is, a Nash equilibrium can be computed) by three steps of iterated deletion of dominated strategies.[6] The row players strategy C is dominated by strategy B (i.e., regardless of what the column player does, B gives a higher payoff than C); if the row player prefers earning more she will never choose C. If the column player guesses that row will never play

[4] The placeholder letter x is placed in cells and rows which are inactive in an effort to create similar amounts of visual activity across trials, since matrices had different numbers of entries.

[5] The split-matrix format was innovated by Costa-Gomes et al (2001), who used it to separate eye movements when players look at their own payoffs or the payoffs of others, in order to judge what decision rules players were using (see also Camerer et al, 1994 [10]). The matrices are more complex than many fMRI stimuli but we chose to use affine transformations of the CGCB matrices to permit precise comparability of our choice data to theirs. Our current study did not track eye movements but it would be simple to use this paradigm to link eye movement to fMRI activity, or to other temporally-fine measures of neural activity.

[6] A *strictly* dominated strategy is one that has a lower payoff than another strategy, for every possible move by ones opponent; A weakly dominated strategy has weakly lower payoffs than another strategy against all strategies and strictly lower payoffs against at least one of the opponents strategies. A dominant strategy is one that gives the highest possible payoff against all of the opponents strategies.

C (the dominated strategy is "deleted", in game theory language – i.e., the column player thinks C will never be played by an earnings-maximizing row player), then strategy BB becomes a dominant strategy for the column player. If the row player guesses that the column player guesses she (the row player) will never play C, and the row player infers that the column player will respond with BB, then strategy A becomes dominant for the row player. Of course, this is a long chain of reasoning which presumes many steps of mutual rationality.

Putting aside the fMRI evidence in our study, simply comparing choices, beliefs and iterated beliefs as we do could be interesting in game theory for a couple of reasons. A common intuition is that higher-order beliefs do not matter. But Weinstein and Yildiz show that in games which are not dominance-solvable, outcomes depend sensitively on higher-order beliefs (if they are not restricted through a common knowledge assumption a la Harsanyi) [55]. Empirically, their theorems imply that knowing more about higher-order beliefs is necessary to guess what will happen in a game.

Goeree and Holt's theory of noisy introspection assumes that higher-order beliefs are characterized by higher levels of randomness or uncertainty. Increased uncertainty might appear as lower levels of overall brain activity (or higher, if they are thinking harder) for 2^nd-order beliefs compared to beliefs and choices. Furthermore, increased uncertainty should be manifested by poorer behavioral accuracy for higher-order beliefs [30].

Second-order beliefs also play a central role in games involving deception. By definition, a successful deception requires a would-be deceiver to know she will make one choice A, but also believe the other player thinks she will make a *different* choice, B. The capacity for deception therefore requires a player to hold "false 2^nd-order beliefs" in mind – that is, to plan choices which are different from what (you think) others think you will do.[7]

Finally, second-order beliefs also play an important role in models of social preferences, when a player's utility depends directly on whether they have lived up to the expectations of others [47]. Dufwenberg and Gneezy studied trust games in which players could pass up a sure amount x and hope that a second player gave them a larger amount y from a larger sum available to divide [19]. They found that the amount the second player actually gave was modestly correlated (.44) with the amount the second player thought the first player expected (i.e., the second player's 2^nd-order belief). The second player apparently felt some obligation to give enough to match player 1's expectations.[8] These kinds of emotions require 2^nd-order beliefs as input.

Trying to discern what another person believes about *you* is also important in games with

[7] Whether or not a person can understand false beliefs is a key component of theory of mind and is also a test used to diagnose autism. In a classic "Sally-Anne" task, a subject is told that Sally places a marble in her basket and leaves the room. Anne then moves the marble from the basket to a box and also leaves the room. Sally re-enters the room. The subject is then asked where Sally will look for her marble. Since the child believes that the marble is in the box, she must be able to properly represent Sally's different belief – a *false* belief – to answer correctly, that Sally will look in the basket. Most children switch from guessing that Sally will look for the marble in the box (a self-referentially-grounded mistake) to guessing that she will be looking in the basket at around 4 years old. Autistic children make this switch later or not at all. See Gallagher and Frith for more detail [25].

[8] However, about a third of the player 2's gave less than they thought other expected.

asymmetric information, when players have private information that they know others know they have, and in games where a "social image" might be important, when people care what others think about them (in dictator and public goods games, among others).

2.1.2 Neuroeconomics, and what it is good for

This study is a contribution to "neuroeconomics", a rapidly-emerging synthesis which ground details of basic economic processes in facts about neural circuitry [11, 12, 56, 28].

Neuroeconomics is an extension of *behavioral economics*, which uses evidence of limits on rationality, willpower and self-interest to reform economic theory; neural imaging is just a new type of evidence. Neuroeconomics is also a new part of *experimental economics*, because it extends experimental methods which emphasize paying subjects according to performance, and tying predictions to theory, to include studies with animals, lesion patients (and "temporary lesions" created by TMS), single-neuron recording, EEG and MEG, psychophysiological recording of heart rate, skin conductance, pupil dilation, tracking eye movements, and PET and fMRI imaging [43]. Neuroeconomics is also part of *cognitive neuroscience*, since these studies extend the scope of what neuroscientists understand to include "higher-order cognition" and complex tasks involving social cognition, exchange, strategic thinking, and market trading that have been the focus of microeconomics for a long time.

One reaction to the idea of neuroeconomics is that economic models do not need to include neural detail to make good predictions, because they are agnostically silent about whether their basic assumptions are actually satisfied, or simply lead to outcomes "as if" the assumptions were true.[9] As a result, one can take a conservative or radical view of how empirical studies like ours should interact with conventional game theory.

The conservative view is that neural data are just a new type of evidence. Theories should get extra credit if they are consistent with these data, but should not be penalized if they are silent about neural underpinnings.

The radical view is that all theories, eventually, will commit to precisely how the brain (or some institutional aggregation, as in a firm or nation-state's actions) carries out the computations that are necessary to make the theory work. Theories that make accurate behavioral predictions and also account for neural detail should be privileged over others which are neurally implausible.

Our view leans toward the radical. It cannot be bad to have theories which predict choices from observable structural parameters and which *also* specify precise details of how the brain creates those choices. (If we could snap our fingers and have such theories for free, we would.) So the

[9] The "as if" mantra in economics is familiar to cognitive scientists in the form of David Marr's influential idea that theories can work at three levels – "computational" (what an economist might call functional or as-if); "algorithmic" or "representational" (what steps perform the computation); and "implementation" or hardware (see Glimcher, 2003 for a particularly clear discussion [29]). Ironically, Marr's three-level idea licensed cognitive scientists to model behavior at the highest level. We invoke it to encourage economists who operate exclusively at the highest level, to commit game theory to an algorithmic view, to use evidence of brain activity to make guesses about algorithms and to therefore discipline ideas about highest-level computation.

only debatable question is whether the cognitive and neural data available *now* are good enough to enable us to begin to use neural feasibility as a central way to judge the plausibility of as-if theories of choice.

We think this is a reasonable time to begin using neural activation to judge plausibility of theories because there are many theories of choice in decision theory and game theory, and relatively few data to sharply separate those theories. Virtually all theories appeal vaguely to plausibility, intuition, or anecdotal evidence, but these are not scientific standards. Without more empirical constraint, it is hard to see how progress can be made when there are many theories. Neural data certainly provide more empirical constraint.

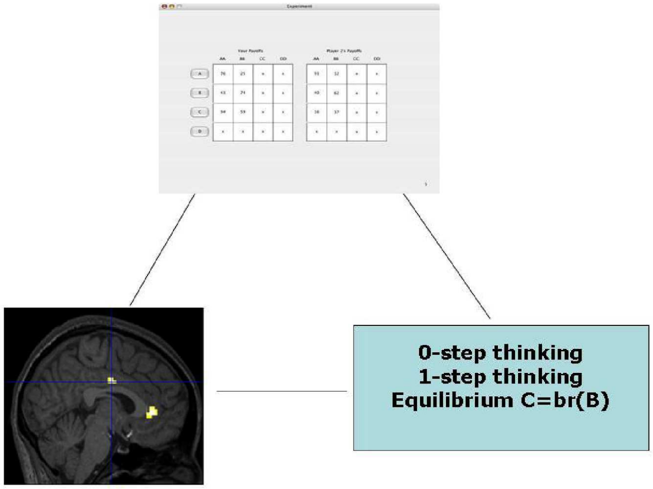
Furthermore, in many domains current theories *do not* make good behavioral predictions. For example, equilibrium game theories clearly explain many kinds of experimental data poorly [6]. Studying cognitive detail, including brain imaging, will inevitably be useful for developing new concepts to make *better* predictions.[10]

An argument for the imminent value of neural data comes by historical analogy to recent studies which track eye movements when subjects play games [10, 17, 15, 40, 39]. When payoffs are placed on a computer screen, different algorithms for making choices can be tested as joint restrictions on the choices implied by those algorithms, *and* whether players look at the payoff numbers they need to execute an algorithm.

Eye tracking has been used in three published studies to separate theories which make similar behavioral predictions. Camerer et al [10] and Johnson et al. studied three-period bargaining games in which empirical offers are somewhere between an equal split and the sub-game perfect self-interest equilibrium (which requires subjects to "look ahead" to future payoffs if bargaining breaks down in early periods; see Camerer, 2003, chapter 4 [6]) [40]. They found that in 10-20% of the games subjects literally did not glance at the possible payoff in a future period, so their offers could not be generated by sub-game perfect equilibrium. Johnson and Camerer found that the failure to look backward, at the possible payoffs of other players in previous nodes of a game, helped explain deviations from "forward induction" [39], CGCB found that two different decision rules, with very similar behavioral predictions about chosen strategies, appeared to be used about equally often, when only choices were used to infer what rules were used. But when lookup information was used, one rule was inferred to be much more likely. If CGCB had only used choices to infer rules, *they would have drawn the wrong conclusion about what rules people were using.*

Those are three examples of how inferences from choices alone do not separate theories nearly as well as inferences from both choices *and* cognitive data. Perhaps neural activity can have similar

[10] Furthermore, neuroeconomics will get done whether economists endorse it or not, by smart neuroscientists who ambitiously explore higher-order cognition carefully but without the benefit of decades of training about how delicate theoretical nuances might matter and which can guide design. Engaging with the energetic neuroscientists is therefore worthwhile for both sides.



	AA	BB	CC	DD	AA	BB	CC	DD
A	76	25	x	x	55	12	x	x
B	43	74	x	x	40	62	x	x
C	94	59	x	x	16	37	x	x
D	x	x	x	x	x	x	x	x

Figure 2.2: Neuroeconomics design: Designs relate stimuli (top) to latent variables or algorithms (right) which generate interpretable activation (left). Experimental economics studies link stimuli (top) and variables (right). Many neuroscience studies just report links between stimuli (top) and activation (left). The neuroeconomics challenge is to make all 3 fit.

power as attentional measures, as evidence accumulates and begins to make sense.

The hard part is creating designs that link neural measures to underlying latent variables. Our work is guided by the “design triangle” illustrated in Figure 2.2. The triangle shows experimental stimuli (on the top of the triangle) which produce measured output – brain activation, skin conductance, eye movements, and so on (lower left) – which can, ideally, be interpreted as expressions of underlying variables or algorithms which are not directly observable (lower right). For the experiments reported in this paper, the underlying constructs which are illuminated by brain activity are hypotheses about the decision processes players are using to generate choices and beliefs.

Keep in mind that while brain pictures like those shown below highlight regions of activation, we are generally interested not just in regions but in neural *circuitry* – that is, how various regions collaborate in making decisions. Understanding circuitry requires a variety of methods. fMRI methods are visually impressive but place subjects in an unnatural (loud, claustrophobic) environment and the signals are weak so many trials are needed to average across. Neuroscience benefits from many tools. For example, looking at tissue in primate brains helps establish links between different regions (“connectivity”). Other methods include psychophysiological measurement (skin conductance, pupil dilation, etc.), studies of patients with specialized brain damage, animal studies, and so forth. Neuroscience is like detective work on difficult cases: There is rarely a single piece of evidence that is definitive. Instead, the simplest theory that is consistent with the most different types of evidence is the one that gets provisionally accepted, and subject to further scrutiny. This paper should be read in this spirit, as extremely tentative evidence which will eventually be combined with many new studies to provide a clear picture.

2.2 Neural correlates of strategic thinking

2.2.1 Methods

Sixteen subjects were scanned,[11] one at a time, in a 3T Siemens Trio scanner at Caltech (Broad Imaging Center) as they performed C, B and $2B$ tasks across each of eight games. The games and order of the three tasks were fixed across subjects. Appendices show the games (which are transformations of games in CGCB), the instructions, and give some methodological details.

In keeping with healthy experimental economics convention, both players were financially rewarded for one task and game that was chosen at random after they came out of the scanner. If a choice task was chosen, then the choices of both players determined their payoffs ($.30 times experimental points). If a belief or 2^{nd}-order belief task was chosen for payment, a player earned $15 if her belief B matched the other players choice, or $15 if her 2nd-order belief $2B$ matched the other players belief.

Pairs of subjects were recruited on campus at Caltech through SSEL lab recruiting software.[12] One subject performed the tasks in the scanner, as the row player, while the other performed them in an adjacent room, as the column player.

We give only a quick sketch of fMRI technique here. Methods of measurement and analysis are extremely complex and still evolving. The Appendix has more detail (or see, e.g., Huettel, Song and McCarthy [38]).

Each subject first has their brain "structurally scanned" (as in medical applications) to establish a sharper picture of the details of brain anatomy for six minutes. Then each subject proceeds through a series of screens (like Figure 2.1) one at a time, at their own pace (response times averaged 8-25 seconds; see the Appendix). They make choices and express beliefs by pressing buttons on a box they hold in their hand. After each response is recorded, there is a random lag from 6-10 seconds with a "fixation cross" on a blank screen to hold their visual attention in the center of the screen and allow blood flow to die down. The entire set of tasks took from 7 to 15 minutes.

The scanner records 32-34 "slices" of brain activity every 2 seconds (one TR). Each slice shows blood flow in thousands of three-dimensional "voxels" which are $3 \times 3 \times 3$ millimeters in size. Our

[11] To experimental social scientists, 16 seems like a small sample. But for most fMRI studies this is usually an adequate sample to establish a result because adding more subjects does not alter the conclusions much.

[12] Since Caltech students are selected for remarkable analytical skill, they are hardly a random sample. Instead, their behavior is likely to overstate the average amount of strategic thinking in a random population. This is useful, however, in establishing differential activation of regions for higher-order strategic thinking since the subjects are likely to be capable of higher-order thinking in games that demand it.

analysis is "event-related", which means we ask which voxels are unusually active when a particular stimulus is on the screen. The analysis is a simple linear regression where dummy variables are "on" when a stimulus is on the screen and "off" otherwise. This "boxcar" regression is convolved with a particular function that is well-known to track the hemodynamic response of blood flow. The regression coefficients of activity in the BOLD (blood oxygen-dependent level) signal in each voxel tell us which voxels are unusually active. Data from all subjects are then combined in a random effects analysis. We report activity which is significantly different from chance at a p-value¡.001 (a typical threshold for these studies), and for clusters of at least 5 adjacent voxels where activity is significant (with exceptions noted below).

2.2.2 Behavioral data

Before turning to brain activity, we first describe some properties of the choices and expressed beliefs. The Appendix shows the relative frequencies of subject choices, expressed beliefs, and expressed 2^{nd}-order beliefs, in each game.

Table 2.1 shows the percentages of trials, for games solvable in different numbers of steps of deletion of dominated strategies, in which players made equilibrium choices. The table includes the choice data from CGCB's original study using these games. First note that the percentages of subjects making the equilibrium strategy choice in our study is similar for row and column players, who are respectively, in and out of the scanner. (None of the row-column percentages are significantly different). However, equilibrium play in our games is less frequent than in CGCB's experiment, significantly so in the simplest games.[13] Since the frequencies of equilibrium play by the in-scanner row player and the out-of-the-scanner column player are similar, the lower percentage of equilibrium play in our experiments is probably due to some factor other than scanning.[14]

Table 2.2 reports the frequency of trials in which $C = br(B)$ (where $br(B)$ denotes the best response to belief B), $B = br(2B)$, $C = 2B$, and in which all three of those conditions are met simultaneously (our stringent working definition of an equilibrium trial hereafter). Equilibrium trials are generally rare (23%). Comparing the match of beliefs and choices across categories, a natural

[13] Of course, eliciting choices, beliefs, and 2^{nd}-order beliefs in consecutive trials might affect the process of choice, perhaps promoting equilibration. But the close match of our observed $C = br(B)$ rate to the Costa-Gomes and Weizsäcker rate, and the lower rate of equilibrium choices compared to CGCB's subjects (who only made choices) suggests the opposite [17, 16]. Also keep in mind that our subjects report a single strategy as a belief, and are rewarded if their guess is exactly right, which induces them to report the mode of their distribution. (For example, if they think AA has a p chance and BB has a $1-p$ chance they should say AA if $p > .5$). Costa-Gomes and Weizsacker elicited a probability distribution of probability across all possible choices. Their method is more informative but we did not implement it in the scanner because it requires a more complex response which is difficult and time-consuming using button presses.

[14] The difference between our rate of conformity to equilibrium choice and CGCB's may be due to the fact that beliefs are elicited, although one would think that procedure would increase depth of reasoning and hence conformity to equilibrium. We think it is more likely to result from a small number players who appeared to act altruistically, trying to make choices which maximize the total payoff for both players (which often leads to dominance violatione.g., cooperation in prisoners dilemma games). Since this kind of altruism is surprisingly difficult to pin down carefully, we continue to use all the data rather than to try to separate out the altruistically-minded trials.

Type of Game	row player (in scanner)	column player (no scanner)	row+column mean	CGCB mean	new data - CGCB z-statistic
2 × 2, row has a dominant decision	0.75	0.61	0.68	0.93	-3.21^*
2 × 4, row has a dominant decision	0.56	0.72	0.65	0.96	-3.24^*
2×2, column has a dominant decision	0.50	0.61	0.56	0.80	-2.46^*
2×4, column has a dominant decision	0.63	0.56	0.59	0.70	-0.94
2 × 3, 2 rounds of iterated dominance	0.47	0.58	0.53	0.69	-1.49
3 × 2, 3 rounds of iterated dominance	0.22	0.22	0.22	0.22	-0.02

Table 2.1: Proportion of equilibrium play across games and player type

intuition is that as players reason further up the hierarchy from choices, to beliefs, to iterated beliefs, their beliefs become less certain. Therefore, 2^{nd}-order beliefs should be less consistent with beliefs than beliefs are with choices, and 2^{nd}-order beliefs and choices should be least consistent [30]. (In terms of the Table 2.2 statistics, the three rightmost column figures should decline from left to right.) That intuition is wrong for these data. The fractions of trials in which $C = br(B)$, and $B = br(2B)$ are about the same. The number of subjects who make optimal choices given their belief ($C = br(B)$) is only 66%. This number may seem low, but it is similar to statistics reported by Costa-Gomes and Weizsacker (2004) (who also measured beliefs more precisely than we did).

More interestingly – and foreshadowing brain activity we will see later – the frequency with which choices match 2^{nd}-order beliefs ($C = 2B$) is actually *higher*, for all classes of games, than the frequency with which $B = br(2B)$ (75% versus 63% overall). This is a hint that the process of generating a self-referential iterated belief might be similar to the process of generating a choice, rather than simply iterating a process of forming beliefs to guess what another player believes about oneself.

Given these results, and the success of parametric models of iterated strategic thinking [9], an obvious analysis is to sort subjects or trials into 0, 1, 2 or more steps of thinking and compare activity. But the current study was not optimally designed for this analysis, so analyses of this type are not insightful.[15]

[15] Comparing trials sorted into low-steps of thinking (0 or 1) and high steps shows very little differential activation

Type of game	equilibrium (all 3 conditions hold)	$C = br(B)$	$B = br(2B)$	$C = 2B$
row has dominant decision	0.31	0.66	0.59	0.69
column has dominant decision	0.44	0.75	0.75	0.88
2×3, 2 rounds of iterated dominance	0.13	0.63	0.66	0.69
3×2, 3 rounds of iterated dominance	0.06	0.59	0.53	0.75
Overall	0.23	0.66	0.63	0.75

Table 2.2: Frequencies of choice and belief matching for the row players

2.2.3 Differential neural activity in choice (C) and belief (B) tasks

In cognitive and neural terms, 0- and 1-step players do not *need* to use the same neural circuitry to make choices and to express beliefs. Thus, any difference in neural activation in the two conditions (C and B) is a clue that some players, on some trials, are making choices without forming beliefs of the sort that require any deep processing about what other players will do, so that belief elicitation is actually a completely different sort of neural activity than choice.[16] Therefore, the first comparison we focus on is between row players *choosing* strategies and *expressing* beliefs about what column players will do.

Figure 2.3 shows brain "sections" which reveal four significantly higher activations in the choice (C) condition compared to the belief (B) condition (i.e., the "$C > B$ subtraction") which have 10 or more adjacent voxels ($k > 10$).[17] The differentially active regions are the posterior cingulate cortex

of high relative to low in either choice or belief tasks, and substantial activation of low relative to high in cingulate and some other regions. The a priori guess is that higher thinking steps produce more cingulate (conflict) activation, so we do not think the sorting into apparent 0- and 1-step trials is accurate enough to permit good inferences at this stage. A design tailored for this sort of typecasting analysis could be used in future research. There are many handicaps from the current design for linking inferred thinking steps to brain activity. One problem is that in many games, choices of higher-step thinkers coincide. Another problem is that it is difficult to weed out altruistic choices, so they are typically misclassified in terms of steps of thinking which adds noise. A cross-subject analysis (trying to identify the typical number of thinking steps for each subject) did not work because individual subject classification is noisy with only eight games (see also [14]). It is also likely that these highly skilled subjects did not vary enough in their thinking steps to create enough variation in behavior to pick up weak behavior-activation links.

[16] An important caveat is that different tasks, and game complexities, will produce different patterns of eye movement. Since we do not have a complete map of brain areas that participate in eye movements for the purpose of decision (though see [29]), some of what we might see might be part of general circuitry for eye movement, information acquisition, etc., rather than for strategic thinking per se. The best way to tackle this is to record eye tracking simultaneously with fMRI and try to use both types of data to help construct a complete picture.

[17] A very large fifth region not shown in Figure 2.3 is in R occipital cortex $(9, -78, 9, k = 202, t = 6.77)$. When we use a smaller k-voxel filter, $k = 5$ (used in Figure 2.3) there are four additional active regions besides the R occipital and those shown in Figure 2.3 (see Appendix Table B.4) which are not especially interpretable in terms of strategic thinking.

(PCC),[18] the anterior cingulate cortex (ACC), the transitional cortex between the orbitofrontal cortex (OFC) and the agranular insula (which we call frontal insula, FI),[19] and the dorsolateral prefrontal cortex (DLPFC). The sections each show differential activity using a color scale to show statistical significance. A 3-dimensional coordinate system is used which locates the middle of the brain at $x = y = z = 0$. The upper left section (1) is "sagittal", it fixes a value of $X = -3$ (that is 3 mm to the left of the zero point on the left-right dimension) The upper right section (2) is "coronal" at $Y = +48$ (48 mm frontal or "anterior" of the Y=0 point). The lower left section (3) is "transverse" (or "axial") at $Z = -18$, 18 mm below the zero line.

Figure 2.4 shows the time courses of raw BOLD signals on the y-axis (in normalized percentage increases in activity) in the PCC region identified above (left, or superior, in the upper left section Figure 2.3), for the C (thick line), B (thin line) and $2B$ (dotted line) tasks. These pictures show how relative brain activity increases or decreases in a particular area over time, for different tasks. The time courses also show standard error bars from pooling across trials; when the standard bars from two lines do not overlap, that indicates statistically significant patterns of activation. The 0 time on the x-axis is when the task stimulus is first presented (i.e., the game matrix appears). The x-axis is the number of scanning cycles (TR's). Each TR is 2 seconds, so a number 4 on the x-axis is 8 seconds of clock time. Perhaps surprisingly, when the stimulus is presented the ACC actually *deactivates* during these tasks (the signal falls). Since blood flow takes one or two TR cycles to show up in imaging (about 3-5 seconds), the important part of the time sequence is in the middle of the graph, between 3 TR's and 8 TR's (when most of the responses are made, since they typically take 8-10 seconds; see Appendix for details).

The important point is that during the belief task (thick line), PCC deactivation is lower than in the $2B$ and B task – hence the differential activation in C minus B shown in the previous Figure 2.3. Most importantly, note that the $2B$ task activity lies between the C and B activity. This is a clue that guessing what someone thinks you will do ($2B$) is a mixture of a guessing process (B), and choosing what you will do (C). This basic pattern – $2B$ is between C and B – also shows up in time courses of activity for all the other areas highlighted in the brain sections in Figure 2.3.

Figure 2.5 shows the location of anterior cingulate cortex (ACC, in yellow) and orbitofrontal

[18] We use the following conventions to report locations and activity: The vector (-3, -9, 33, $k = 5$, positive in 14 of 16 subjects) means that the voxel with peak activation in the cluster has coordinates $x = -3$,$y = -9$, $z = 33$. The coordinates x, y, and z respectively measure distance from the left to the right of the brain, from front ("anterior") to back ("posterior"), and bottom ("inferior") to top ("superior"). The figure $k = 5$ means the cluster has 5 voxels of 3 cubic millimeters each. The number of subjects with positive regression coefficients is an indication of the uniformity of the activation across subjects. Appendix Table B.4 shows coordinates for all regions mentioned in this paper, and some regions that are not discussed in the text.

[19] FI and ACC are the two regions of the brain known to contain spindle cells. Spindle cells are large elongated neurons which are highly "arborized" (like a tree with many branches, they project very widely, and draw in information and project information to many parts of the brain) that are particular to humans and higher primate kin, especially bonobos and chimpanzees [2]. It is unlikely that any of these brain areas are solely responsible for our ability to reason about others. In fact it seems that the pathologies where individual do not have these abilities, namely Autism and Asperger syndrome, do not involve lesions of any specific areas of the brain, but rather more generalized developmental problems including a decreased population of spindle cells (Allman, Caltech seminar), decreased connectivity to the superior temporal sulcus [13], and defects in the circuitry of the amygdala [51].

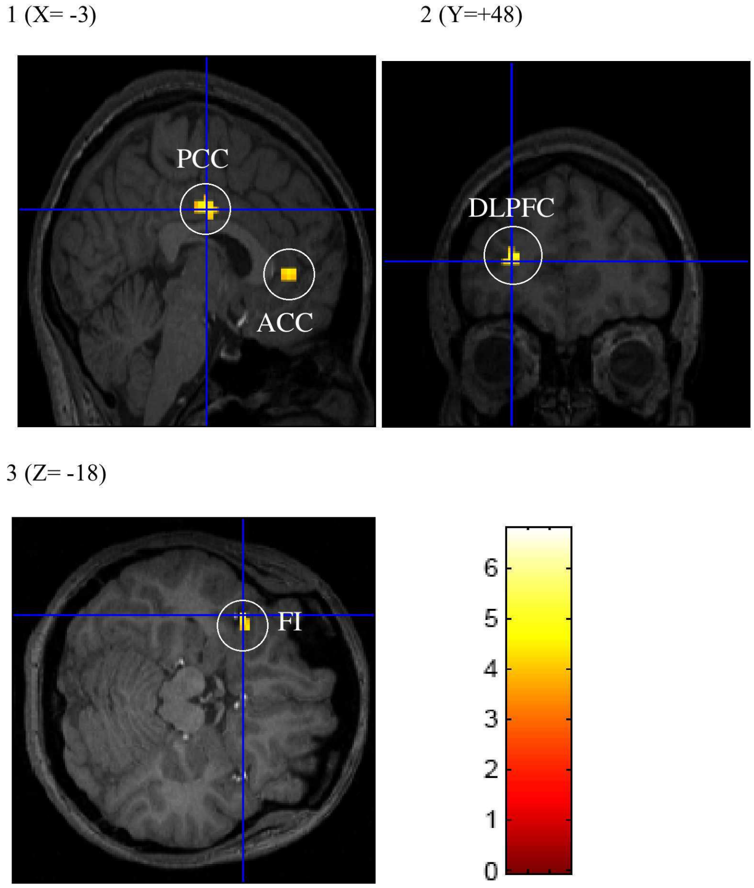


Figure 2.3: Areas of significantly differential activity in choice minus belief conditions, all trials, at $p < .001$ (uncorrected). (1) Top area is posterior cingulate cortex, PCC $(-3, -12, 33, k = 24, t = 5.12$; 14 of 16 subjects positive); right area is anterior cingulate cortex/genu ACC $(6, 42, 0; k = 33, t = 4.62$; 15 of 16 subjects positive). (2) dorsolateral prefrontal cortex DLPFC $(-27, 48, 9; k = 14, t = 4.74$; 15 of 16 subjects positive). (3) transition cortex/FI $(-42, 12, -18; k = 31, t = 4.60$; 14 of 16 subjects positive).

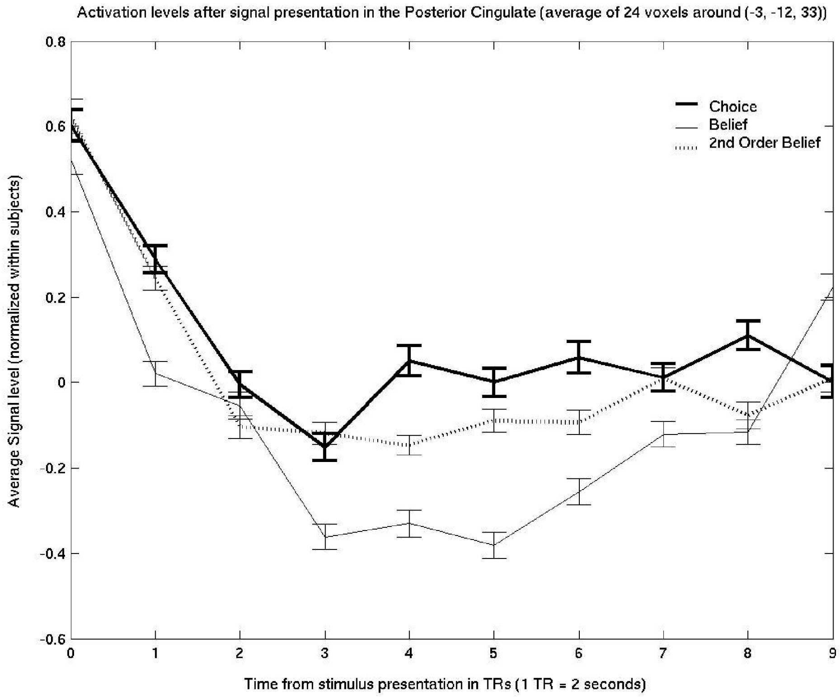


Figure 2.4: Time course of activity in posterior cingulate (-3,-12, 33) in choice (C, thick line), belief (B, thin line) and 2^{nd}-order belief ($2B$, dotted line) tasks.

cortex (pink). The cingulate cortex is thought to be important in conflict resolution and "executive function" [44]. The ACC and PCC regions that are differentially active in choosing rather than forming beliefs have both been implicated in ToM and in other social reasoning processes. The PCC is differentially active in moral judgments that involve personal versus impersonal involvement and many other kinds of processing that involve emotional and cognitive conflict [32]. D. Tomlin (personal communication) has found relative activation in the very most anterior (front) and posterior (back) cingulate regions that are shown in Figure 2.3 in repeated trust games with a very large sample (almost 100 pairs of players), after another player's decision is revealed.[20] Since their subjects are playing repeatedly, presentation of what another player actually does provides information on how he may behave in the next trial, it is possible that this evidence is immediately used to start making the players next decision.

The fact that all these regions are more active when people are making choices, compared to expressing beliefs, suggests that a very simple neural equation of forming a belief and choosing is

[20] Tomlin et al reported a "self-other" map of the cingulate which includes the most anterior and posterior regions we see in Figure 2.3. They studied brain activation during repeated partner trust games. When the other players behavior was shown on a screen, the most anterior (front of the brain) region was active, independent of the player role. When ones own behavior was shown, more middle cingulate regions were activated. The most posterior (back) regions were activated when either screen was shown. The brain often "maps" external parts of the world (retinotopic visual mapping) or body (somatosensory cortex). The cingulate map suggests a similar kind of "sociotopic" mapping in the cingulate.

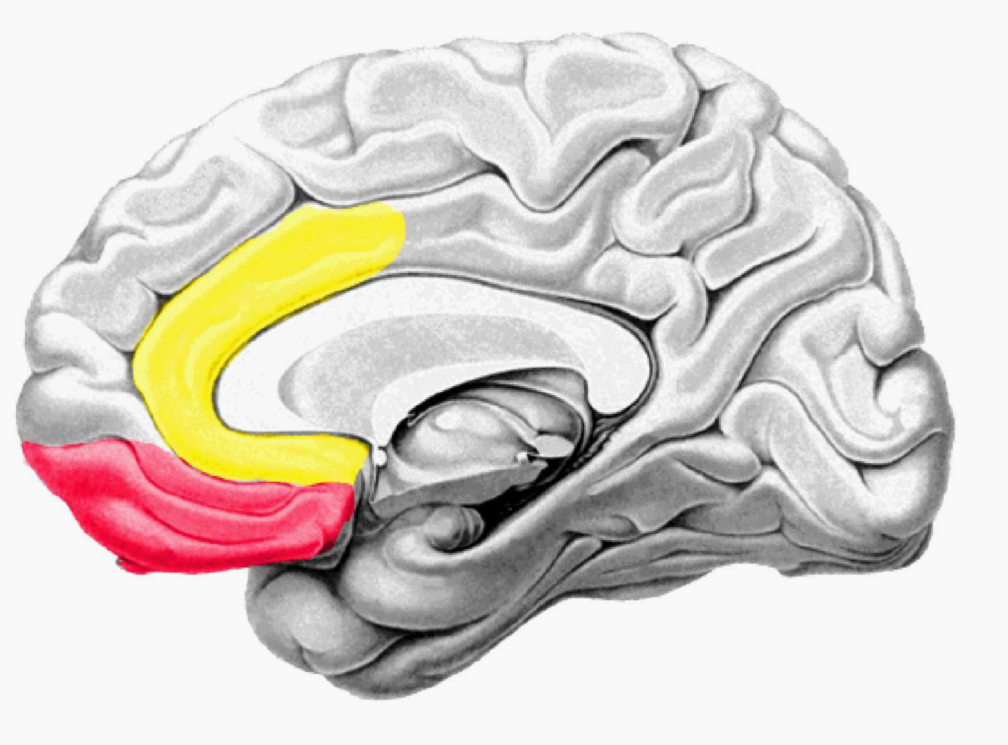


Figure 2.5: A brain drawing showing anterior cingulate cortex (ACC, yellow) and orbitofrontal cortex (OFC, pink). The front of the brain (anterior) is to the left. Reprinted with permission of Ralph Adolphs.

leaving out some differences in neural activity that are clues to how the processes may differ.

The FI region we identify is close to an area noted by Gallagher et al. (38, 24,-20) in inferior frontal cortex [24]. Their study compared people playing a mixed-equilibrium (rock, paper, scissors) game against human opponents versus computerized opponents. The identification of a region differentially activated by playing people, which is nearby to our region is a clue that this inferior frontal/FI region might be part of some circuitry for making choices in games against other players.

Differential activation in frontal insula (FI) is notable because this area is activated when people are deciding how to bet in ambiguous situations relative to risky ones, in the sense of Ellsberg or Knight [37]. This suggests choice in a game is treated like an ambiguous gamble while expressing a belief is a risky (all-or-none) gamble. This interpretation is consistent with 0- and 1-step thinking, in which evaluating strategies and likely payoffs occurs with a shallow consideration of what other players will do, which seems more ambiguous than forming a belief.

2.2.4 Equilibrium as a state of mind: Choice and belief in- and out-of-equilibrium

The evidence and discussion above suggests that the processes of making a strategic choice and forming a belief are *not* opposite sides of a neural coin. Interesting evidence about this neural-equivalence hypothesis emerges when the trials are separated into those in which all choices and beliefs are in equilibrium (i.e., $C = br(B)$, $B = br(2B)$ and $C = 2B$) and those which are out of equilibrium (one or more of the previous three parenthetical conditions does not hold).

Figure 2.6 shows sections of differential activity in the C and B tasks during equilibrium trials. This is "your brain in equilibrium": There is only one area actively different (at p¡.001) in the entire brain. This suggests that equilibrium can be interpreted not only as a behavioral condition in which choices are optimal and beliefs rational, but also can be interpreted neurally as a *state of mind*: When choices, beliefs and 2^{nd}-order beliefs all match up accurately, and are mutual best responses, there is only a minimal difference in activation between choice and belief, which means the mechanisms performing those tasks are highly overlapping.[21]

Figure 2.6 does show one important differential activation, however, in the ventral striatum, This region is involved in encoding reward value of stimuli and predicting reward [50]. This area is also differentially activated when we compare choice to the 2^{nd}-order belief task, t-statistic ¿ 4 in several overlapping voxels). This difference could be due to the difference in rewards in the choice and belief tasks. Note that activation in FI is not significantly different between the C and B tasks in equilibrium (cf. Figure 2.3), which is a clue that perceived ambiguity from choosing is lower when choices and beliefs are in equilibrium.

Figure 2.7 shows the C minus B differential activation in trials when choices and beliefs are out of equilibrium. Here we see some areas of activation similar to those in the overall C minus B subtraction.[22] The novel activity here is in the paracingulate frontal cortex region (Brodmann area BA 8/9; Figure 2.7, upper left section). This region has appeared in mentalizing tasks in two studies. One is the Gallagher et al. (2002) study of "rock, paper, scissors"; a paracingulate area just anterior to the one in Figure 2.7 is differentially active when subjects played human opponents compared to computerized algorithms [24].[23] McCabe et al. also found significant differential activations in the

[21] The difference between in- and out-of-equilibrium $C > B$ activity does not simply reflect the complexity of the games which enter the two samples, because separating the trials into easy (solvable by dominance for row or column) and hard (solvable in 2-3 steps) does not yield a picture parallel to Figures 2.6-2.7. The difference is also not due to lower test power (there are fewer in-equilibrium than out-of-equilibrium trials) because the strategic areas active in Figure 2.7 are not significantly activated in the in-equilibrium C¿B subtraction (paracingulate t=.36; dorsolateral prefrontal, t=1.34).

[22] Note that the Figure 2.3 activations, which pool all trials, do not look like a mixture of the Figure 2.6 (in-equilibrium trials) and Figure 2.7 (out-of-equilibrium trials) activities. However, the areas which are differentially active below the p¡.001 threshold when all trials are pooled do tend to have activation in the in- and out-of-equilibrium subsamples, but activation is more weakly significant in the subsamples and vice versa. In the $C > B$ subtraction for out-of-equilibrium trials, the PCC is active at p¡.01 and the ACC at p¡.005. The dorsolateral prefrontal region (see Figure 2.7) at (-30,30,6,k=14) which is active (p¡.001) in the out-of-equilibrium trials is just inferior to the region active in all trials (-27,48,9,k=14).

[23] In both conditions the subjects were actually playing against randomly chosen strategies (which is the Nash

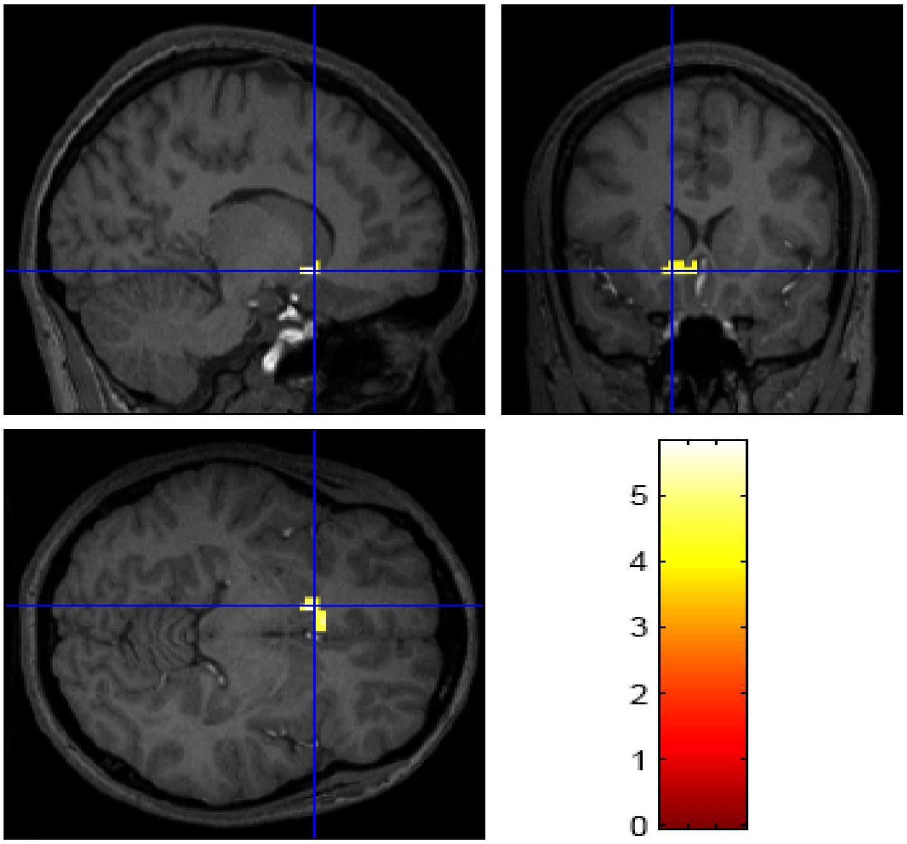


Figure 2.6: This is your brain in equilibrium: Area of significant differential activation in $C > B$ for in-equilibrium trials. The only significant area at p¡.001 (-3,21,-3; k=20, t=5.80) is ventral striatum.

same area among subjects who were above the median in cooperativeness in a series of trust-like games, when they played humans versus computers [42].

In our tasks, of course, choosing and expressing belief are both done with another opponent in mind (in theory). Activation of the paracingulate region in our non-equilibrium $C > B$ subtraction and in Gallagher et al's and McCabe et al's human-computer difference suggests that people are reasoning more thoughtfully about their human opponent when choosing rather than believing. This pattern is consistent with low-level strategic thinking in which players do not spend much time thinking about what others will do in forming beliefs, when they are out of equilibrium. In those cases, assuming your opponent will behave randomly is like treating your human opponent like a computer randomizer.

The difference we observe in brain activity in- and out-of-equilibrium is similar to Grether et al's fMRI study of bidding in the incentive-compatible Vickrey second-price auction. After players were taught they should bid their values (a dominant strategy), activity in the ACC was diminished [33].

2.2.5 Self-referential iterated strategic thinking: 2^{nd}-order beliefs versus beliefs

The second comparison we focus on is differential activity in the brain when row players are asked what they think the column players think *they* (the row players) will do– their 2^{nd}-oder beliefs – compared to brain activity when they are just asked to state beliefs about what column players will do.

Figure 2.8 shows differential activity in the $2B$ condition, compared to B, in those trials where players were out of equilibrium.[24] The large (k=35 at p=.005) voxel area is the anterior insula (a smaller subset of these voxels, k=3, are still significant at p=.001).

The insula is the region in the brain responsible for monitoring body state and is an important area for emotional processing (see Figure 2.9 for a picture of where the insula is). Parts of the insula project to frontal cortex, amygdala, cingulate, and ventral striatum. The insula is hyperactive among epileptics who feel emotional symptoms from seizures (fear, crying, uneasiness [20]), and in normal subjects when they feel pain, disgust and social anxiety. Sanfey et al. found that the insula was activated when subjects received low offers during the ultimatum game [49]. Eisenberger et. al. found the area was activated when subjects were made to feel socially excluded from a computerized game of catch [21]. Importantly for us, the insula is also active when players have a sense of self-

equilibrium for this game). The occasional practice of deception in economics experiments conducted by neuroscientists raises a scientific question of whether it might be useful to agree on a no-deception standard in this emerging field, as has been the stubborn and useful norm in experimental economics to protect the public good of experimenter credibility.

[24] This $2B > B$ subtraction for the in-equilibrium trials yields no significant regions at p¡ .001. As noted earlier, this shows that being in equilibrium can be interpreted as a state of mind in which forming beliefs and 2^{nd}-order beliefs are neurally similar activities

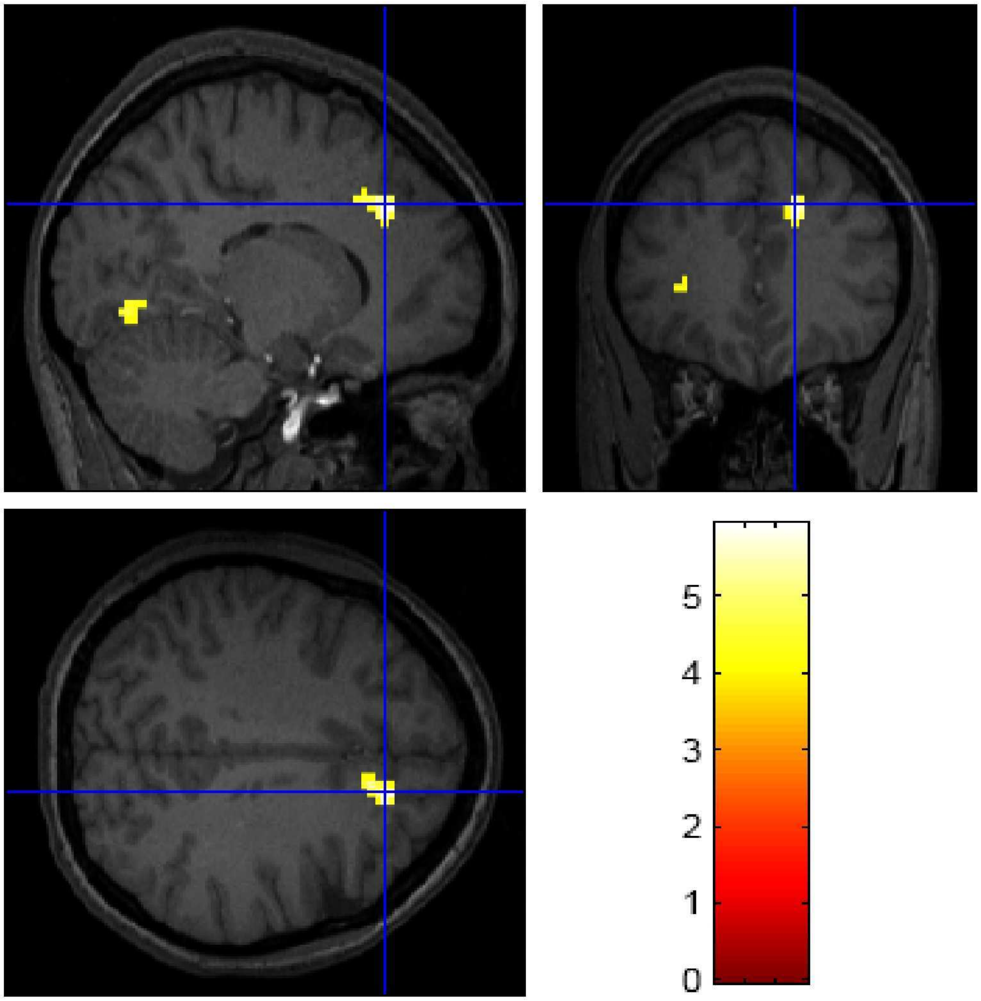


Figure 2.7: This is your brain out-of-equilibrium: Areas of significant differential activation in $C > B$ for out-of-equilibrium trials. Largest area (15,36,33; k=39; t=5.93, 12 of 13 positive) is paracingulate cortex (BA 9), visible in all three sections. Posterior area in the sagittal section (left in upper left section) is occipital cortex (12, -75, -6; k=19, t=4.84). Ventral area in the coronal section (leftmost activity in the upper right section) is dorsolateral prefrontal cortex (-30,30,6,k=14,t=4.85).

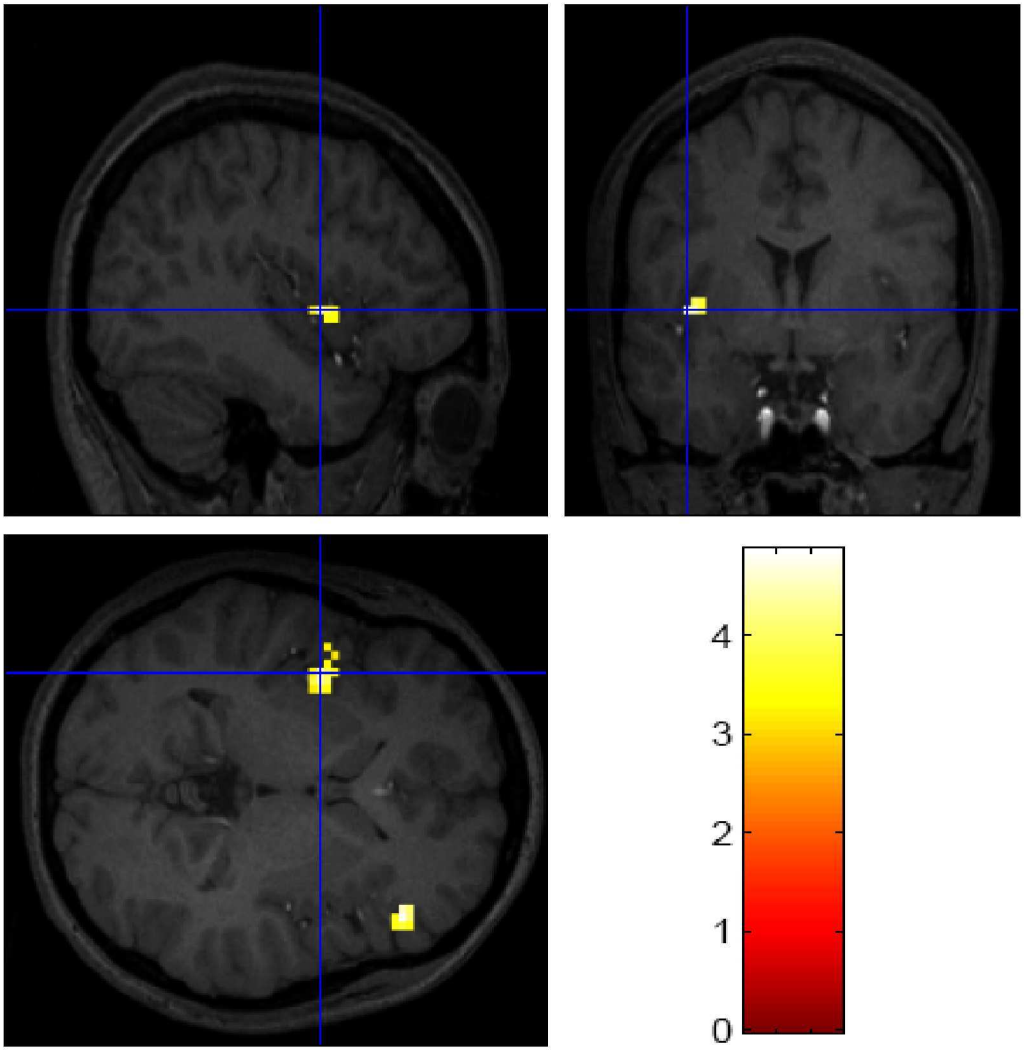


Figure 2.8: Differential activity in iterated belief $(2B)$ minus belief (B) conditions, out-of-equilibrium trials only. Significance level p¡.005 (uncorrected). N=13 because some subjects did not have enough non-Nash trials to include. Area visible in all three sections is left insula (-42,0,0, k=35, t=4.44, 12 of 13 positive). This area is still active but smaller in cluster size at lower p-values (k=9 at p=.002, k=3 at p=.001). The other active region in the transverse slice (lower left) is inferior frontal gyrus (45,33,0; k=13, t=4.85).

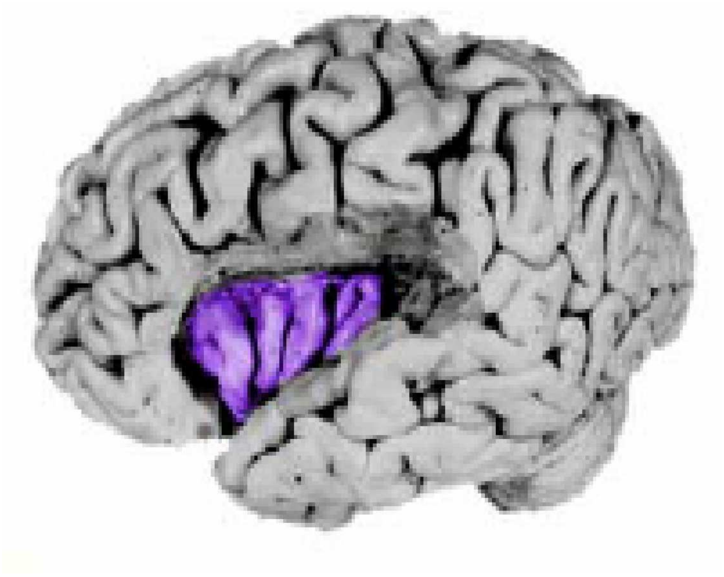


Figure 2.9: A brain drawing showing insula cortex (in purple), as it would appear with the temporal lobe peeled back at the Sylvian fissure. The front of the brain (anterior) points to the left. Drawing reprinted with permission of Ralph Adolphs.

causality from driving a cursor around a screen (compared to watch equivalent cursor movement created by others [22]), or recall autobiographical memories [23]. These studies suggest that insula activation is part of a sense of "agency" or self-causation, a feeling to which bodily states surely contribute. Our region overlaps with the area found by Farrer and Frith [22].

The insula activation in creating 2^{nd}-order beliefs supports the hypothesis that 2^{nd}-order belief formation is not simply an iteration of belief formation applied to imagine how what other players belief about you. Rather, it is a combination of belief-formation and choice-like processes. We call this the self-referential strategic thinking hypothesis. The basic facts that C and $2B$ activations tend to be very similar, C and $2B$ choices often match up (Table 2.2), and that activations in the C and $2B$ tasks both tend to be different from B in similar ways,[25] supports this hypothesis too.

2.2.6 Individual differences: Brain areas that are correlated with strategic IQ

All the analyses above pool across trials and subjects (assuming random effects). Another way to approach the data is to treat each subject as a unit of analysis, and ask how activation is correlated with behavioral differences in skill, across subjects.

To do this we first calculate a measure of "strategic IQ" for each subject. Remember that subjects actually had a human opponent in these games. Since subjects did not receive any feedback until

[25] Differential C¿B activation in the same insula region observed in the $2B > B$ subtraction is marginally significant (t=2.78), and is positive for 10 out of the 13 subjects in the sample.

they came out of the scanner (and one of each of the C, B, and $2B$ trials was chosen randomly for actual payment), it makes sense to judge the *expected* payoffs from their choices, and the accuracy of their beliefs, by comparing each row subject with the population average of *all* the column players.[26] We use this method to calculate the expected earnings for each subject from their choices, and from accuracy of their beliefs (i.e., how closely did their beliefs about column players' choices match what the column players *actually did*?) and similarly for 2^{nd}-order beliefs. Their earnings in each of the three tasks are then standardized (subtracting the task-specific mean and dividing by the standard deviation). Adding these three standardized earnings numbers across the C, B, and $2B$ tasks gives each subject's strategic IQ relative to other subjects. (The three numbers are only weakly correlated, about .20, across the three tasks, as is typical in psychometric studies.)

We then regressed activation during the choice task on these strategic IQ's. The idea is to see which regions have activity that is correlated with strategic IQ.

We expected to find that players with higher strategic IQ might have, for example, stronger activation in ToM areas like cingulate cortex or the frontal pole BA 10. However, we found no correlations with strategic IQ in areas most often linked to ToM. Positive and negative effects of skill on activation in these areas might be canceling out. That is, players who are skilled at strategic thinking might be more likely to think carefully about others, which activates mentalizing regions. However, they may also do so more effortlessly or automatically, which means activity in those regions could be lower (or their responses more rapid).[27]

However, choice-task activity in a k=13 voxel cluster in the precuneus and a k=11 voxel cluster in the caudate (dorsal striatum), are positively correlated with SIQ (p¡.001 and p¡.05 respectively), as shown in Figure 2.10. The precuneus neighbors the posterior cingulate (PCC) and is implicated in "integration of emotion, imagery and memory" [32]. Perhaps high-SIQ players are better at imagining what others will do, and this imaginative process in our simple matrix games uses all-purpose circuitry that is generally used in creating empathy or doing emotional forecasting involving others. The SIQ-caudate correlation shown in Figure 2.10 is naturally interpreted as reflecting the greater certainty of rewards for the high SIQ subjects. This shows a sensible link between actual success at choosing and guessing in the games (experimental earnings) and the brain's *internal* sense of reward in the striatum.

We also find interesting *negative* correlations between strategic IQ and brain activity during the

[26] This is sometimes called a "mean matching" protocol. It smooths out the high variance which results from matching each in-scanner subject with just one other subject outside the scanner.

[27] The identification problem here is familiar in labor economics, where there is unobserved skill. If you run a regression on output (y) against time worked (t) across many workers, for example, it might be negative because the most skilled workers are so much more productive per unit time that they can produce more total output in a shorter time than slow workers, who take longer to produce less. Similarly, Chong, Ho, and Camerer [14] recorded response times of subjects and then inferred the number of steps of thinking the subjects were doing from their choices. Surprisingly, they found that the number of thinking steps was negatively correlated with response time. This puzzle can be explained if the higher-step thinkers are much faster at doing each step of thinking. It might also mean, as noted in footnote 14, that subjects classified as 0-step thinkers are actually doing something cognitively sophisticated which the model cannot classify as higher-level thinking. (In some games, this even includes Nash equilibrium choices.)

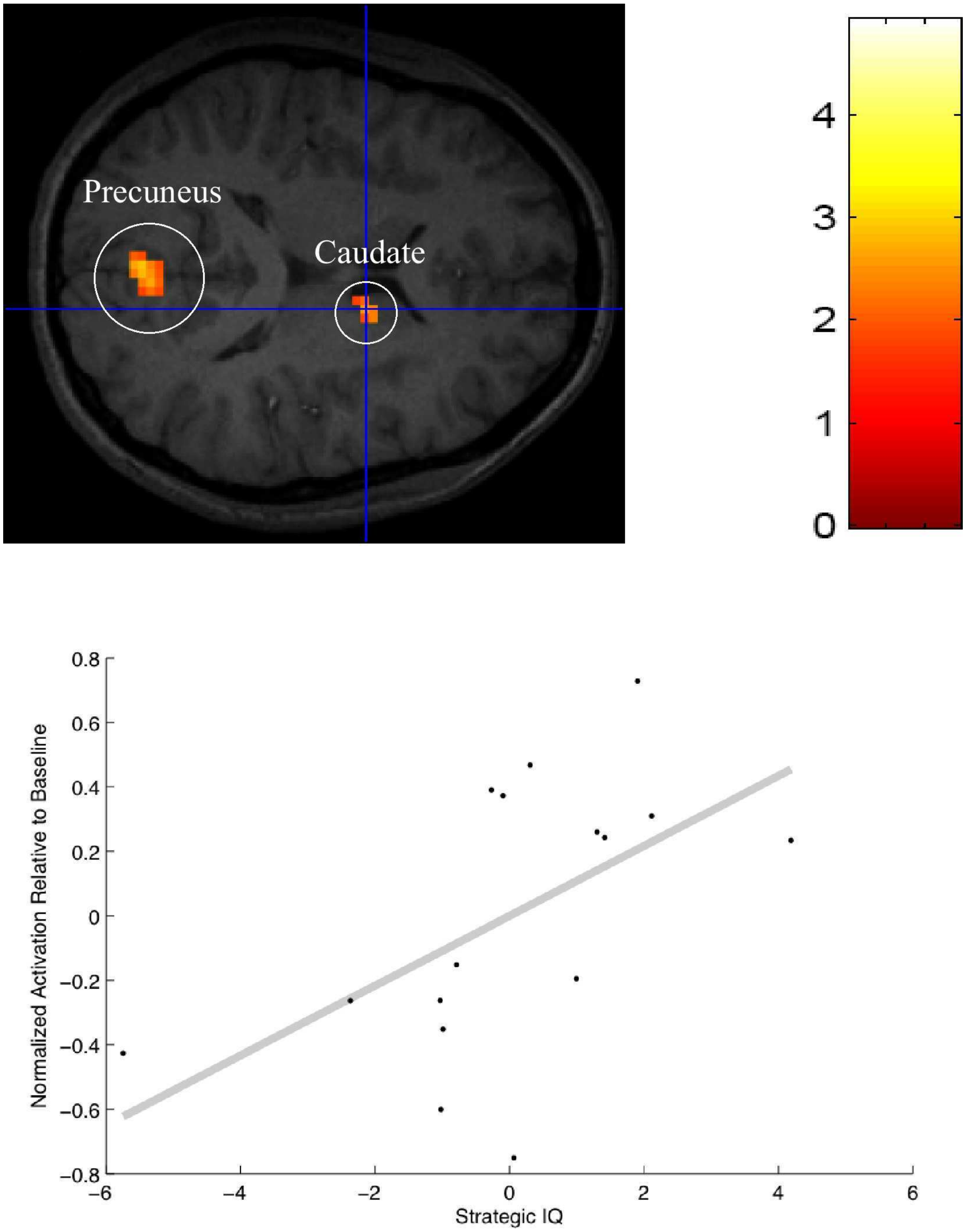


Figure 2.10: (top) Areas positively correlated with SIQ (p ¡ .05): Precuneus (on left, 3,-66,24, k=312, t=4.90), ; caudate (dorsal striatum) (12, 0, 15, k=11, t=2.52). (bottom) Cross-subject correlation between relative caudate activity (y-axis) and relative SIQ (x-axis) (r=.56, p¡.025; rank-order correlation=.60)

choice task. Figure 2.11 shows the strong negative correlation between SIQ and activity in the left anterior insula (-39,6,-3, k=25) in the choice task, relative to a baseline of all other tasks, and also shows the insula region of interest in a sagittal slice.[28] Note that the low-SIQ players have an *increase* in activation relative to baseline (i.e., the y-axis values for those with negative standardized SIQ are positive), while the high-SIQ players have a decrease (negative y-axis values). This makes it a little hard to interpret why SIQ and insula activity are correlated.

As noted above, the region of anterior insula in Figure 2.11 which is correlated with SIQ is also differentially active in the $2B$ task relative to the B task. We interpret this as evidence that subjects are self-focused when forming self-referential iterated beliefs. The increase in insula activity might be an indication that too much self-focus in making a choice is a mistake – subjects who are more self-focused do not think enough about the other player and make poorer choices and less accurate guesses. An alternative explanation is that subjects who are struggling with the tasks, and earn less, feel a sense of unease, or even fatigue from thinking hard while lying in the scanner (remember that the insula is activated by bodily discomfort). The higher insula activation for lower strategic IQ players may be the body's way of expressing strategic uncertainty to the brain. The fact that there is *de*activation in the choice task for higher SIQ players suggests a different explanation for them – e.g., by concentrating harder on the games they "lose themselves" or forget about body discomfort.

The fact that insula activity is negatively correlated with strategic IQ suggests that self-focus may be harmful to playing games profitably. A natural followup study to explore this phenomenon is to compare self-referential iterated beliefs of the form "what does subject A think that B think I (i.e., A) will do" with "what does *someone else* (C) think B thinks A will do" (a non-self-referential 2^{nd}-order belief task). If self-focus harms the ability to guess accurately what B thinks you (A) will do, a third party (C) may be more accurate about guessing B's beliefs about As move than A is. This possibility is related to psychology experiments on "transparency illu[26] and "curse of knowledge" [5, 41]. In these experiments, subjects find it hard to imagine that other people do not know what they the subjects themselves know.

At this point, we don't know empirically if non-self-referential 2^{nd}-order beliefs are more accurate than self-referential 2^{nd}-order beliefs. The key point is that we would never have thought to ask this question until the neuroeconomic method suggested a link between insula activity, self-reference, and low strategic IQ. This is one illustration of the capacity of neural evidence to inspire new hypotheses.

[28] The y-axis is the regression coefficient in normalized signal strength (%) for each subject from a boxcar regression which has an independent dummy variable of +1 when the choice task stimulus is on the screen–from onset to the time that the subject made a decision with a button press–and 0 otherwise. The activation is scaled for each subject separately in percentage terms, so the results do not merely reflect differences in overall activation between subjects. The rank-order correlation corresponding to the correlation in Figure 2.11 is -.81 (t=5.08) so it is not simply driven by outliers.

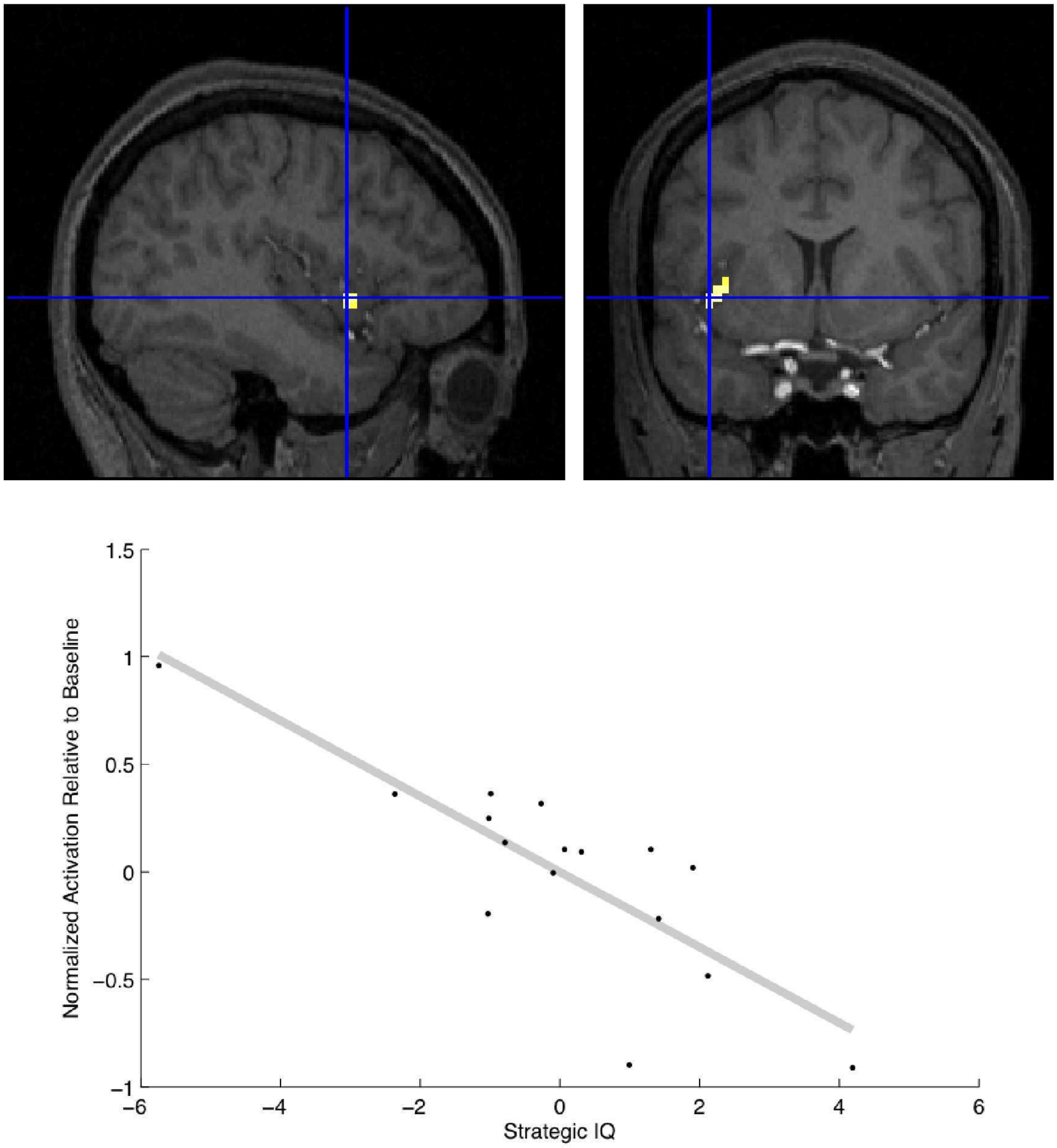


Figure 2.11: (top) Sagittal slice showing L insula (-42, 6,-3, k=12, t=5.34), p¡.0005. (bottom) Cross-subject correlation between L insula relative activity (y-axis) and relative SIQ (x-axis) (r= -.82, p¡.0001; rank-order correlation= -.81).

2.3 Discussion and conclusion

Our discussion has two parts. We first mention some earlier findings on neuroscientific correlates of strategic thinking. Then we will summarize our central findings, and briefly conclude about how to proceed.

2.4 Other neuroscientific evidence on strategic thinking

An irony of neuroeconomics is that neuroscientists often find the most basic principles of rationality *useful* useful in explaining human choice, while neuroeconomists like ourselves hope to use neuroscience to help us understand *limits* of rationality in complex decision making (usually by suggesting how to weaken rationality axioms in biologically-realistic ways).[29] As a result the simplest studies of strategic thinking by neuroscientists focus on finding brain regions that are specially adapted to do the simplest kind of strategic thinking–reacting differently to humans compared to nonhuman computerized algorithms. As noted earlier, when subjects played mixed-equilibrium and trust games, respectively, against humans rather than computerized opponents, Gallagher et al found activation in inferior frontal areas and paracingulate areas [24], and McCabe et al. found activity in the frontal pole (BA10), parietal, middle frontal gyrus and thalamic areas [42].

A few other studies have focused on reward and emotional regions in games. Rilling et al. found striatal activation in response to mutual cooperation in a PD, which they interpret as a rewarding "warm glow" that sustains cooperation [48]. deQuervain et al find nucleus accumbens activation when third-party players sanction players who betrayed the trust of another player, showing a "sweet taste of revenge" (which is also price-sensitive, revealed by prefrontal cortical activity) [18]. The Sanfey et al. study on ultimatum games showed differential insula, ACC, and dorsolateral prefrontal activation for low offers [49]. Singer et al. found that merely seeing the faces of players who had cooperated activated reward areas (striatum), as well as the insula [53]. The latter finding suggests where game-theoretic concepts of a persons "reputation" are encoded in the brain and are linked to expected reward. Tomlin et al (personal communication) find that the most anterior and posterior cingulate regions are active when players are processing what other players have done in a repeated trust games.

[29] The same irony occurs in models of risky choice where strategic thinking plays no role. Glimcher (2003) shows beautifully how simple expected value models clarified whether parietal neurons encode attention, intention or–the winner– something else (expected reward) [29]. At the same time, decision theorists imagine that neural circuitry might provide a foundation in human decision making for theories showing how choices violate simple rationality axioms–viz., that evaluations are reference-dependent, probabilities are weighted nonlinearly, and emotional factors like attention and optimism play a central role in risky decision making. A way to reconcile these views is to accept that simple rationality principles guide highly-evolved pan-species systems necessary for survival (reward, food, sex, violence) but that complex modern choices are made by a pastiche of previously-evolved systems and are unlikely to have evolved to satisfy rationality axioms only discovered in recent decades. Understanding such modern decisions forces us to become amateur neuroscientists and learn about the brain, and talk to those who know the most about it.

Many of these regions are also active in our study. The insula, active in evaluating low ultimatum offers and upon presentation of cooperating partners, is also active in creating 2^{nd}-order beliefs in our study. The cingulate regions in Tomlin et al. are also prominent when players are choosing strategies, compared to guessing what other players will do.

Special subject pools are particularly informative in game theory, where stylized models assume players are both self-interested (almost sociopathic) and capable of great foresight and calculation. Hill and Sally compared autistic children and adults playing ultimatum games. About a quarter of their autistic adults offered nothing in the ultimatum game, which is consistent with an inability to imagine why others would regard an offer of zero as unfair and reject it. Offers of those adult autistics who offer more than zero cluster more strongly around 50% than the autistic childrens' offers, which are sprinkled throughout the range of offers [36]. The child-adult difference suggests that socialization has given the adults a rule or "work around" which tells them how much to offer, even if they cannot derive an offer from the more natural method of emotionally forecasting what others are likely to accept and reject. Gunnthorsdottir, McCabe and Smith found that subjects high on psychometric "Machiavellianism" ("sociopathy lite") were twice as likely to defect in one-shot PD games than low-Mach subjects [34].

A sharp implication in games with mixed equilibria is that all strategies that are played with positive probability should have equal expected reward. Platt and Glimcher found neurons in monkey parietal cortex that have this property [46]. Their parietal neurons, and dorsolateral prefrontal neurons in monkeys measured by Barraclough, Conroy and Lee [3], appear to track reinforcement histories of choices, and have parametric properties that are consistent with Camerer and Ho's dual-process EWA theory [7], which tracks learning in many different games with human subjects.[30]

Still other studies have focused on coarse biological variables rather than detailed brain processes. In sequential trust games, Zak et al. find a link between levels of oxytocin–a hormone which rises during social bonding (such as intimate contact and breast-feeding) and trust [57]. Gonzales and Loewenstein found that circadian rhythms (whether you're a night or morning person) affected behavior in repeated trust (centipede) games–players who are "off peak" tended to cooperate less [31].

[30] In the Camerer-Ho theory, learning depends on two processes: (1) A process of reinforcement of actual choices, probably driven by activity in the limbic system (striatum), and (2) a potentially separate process of reinforcing unchosen strategies according to what they would have paid (which probably involves a frontal process of counterfactual simulation similar to that involved in regret). A parameter δ represents the relative weight on the counterfactual reinforcement relative to direct reinforcement. Estimates by Barraclough, Conroy and Lee [3] from activity in monkey prefrontal cortex support the two-process theory. They estimate two reinforcements: When the monkeys choose and win (reinforcement by ?1), and when they choose and lose (?2). In their two-strategy games, the model is mathematically equivalent to one in which monkeys are not reinforced for losing, but the unchosen strategy is reinforced by ?2. The fact that ?2 is usually less than ?1 in magnitude (see also Lee et al., in press) is equivalent to ?¡1 in the Camerer-Ho theory (less reinforcement in the second process from unchosen strategies), which corresponds to parametric measures from many experimental games with humans [8].

2.4.1 What we have learned

In this paper, we scanned subjects' brain activity using fMRI as they made choices, expressed beliefs, and expressed iterated "2^{nd}-order" beliefs. There are three central empirical findings from our study:

- A natural starting point for translating game theory into hypotheses about neural circuitry is that most of the processes in making choices and forming beliefs should overlap when players are in equilibrium. Indeed, in trials where choices and beliefs are in equilibrium, this hypothesis is true– the only region of differential activation between choice and belief tasks is the striatum, perhaps reflecting the higher "internal expected reward" from making a choice compared to guessing. In general, however, making a choice (rather than making a guess) differentially activates posterior and anterior cingulate regions, frontal insula, and dorsolateral prefrontal cortex. Some of these regions are part of "theory of mind" circuitry, used to guess what others believe and intend to do. The cingulate activity suggests that brains are working harder to resolve cognitive-emotional conflicts in order to choose strategies.

- Forming self-referential 2^{nd}-order beliefs–guessing what others think you will do–compared to forming beliefs, activates the anterior insula. This area is also activated by a sense of agency or self-causation (as well as by bodily sensations like disgust and pain). Combined with behavioral data and study of the time courses of activation, this suggests that guessing what others think you will do is a mixture of forming beliefs and making choices. For example, this pattern of activity is consistent with people anchoring on their own likely choice and then guessing whether other players will figure out what they will do, when forming a self-referential 2^{nd}-order belief.

- Since subjects actually play other subjects, we can calculate how much they earn from their choices and beliefs–their "strategic IQ". When they make choices, subjects with higher strategic IQ have stronger activation in the caudate region (an internal signal of predicted reward which correlates with actual earnings) and precuneus (an area thought to integrate emotion, imagery and memory, suggesting that good strategic thinking may use circuitry adapted for guessing how other people feel and what they might do). Strategic IQ is negatively correlated with activity in insula, which suggests that too much self-focus harms good strategic thinking, or that poor choices are neurally expressed by bodily discomfort.

It is too early to know how these data knit together into a picture of brain activity during strategic thinking. However, activity in cingulate cortex (posterior, neighboring precuneus, anterior, and paracingulate) all appear to be important in strategic thinking, as does activity in dorsolateral prefrontal cortex, the insula region and in reward areas in the striatum. The most novel finding is that activity in creating self-referential 2^{nd}-order beliefs activates insula regions implicated in a sense of self-causation. That interpretation, along with the fact that 2^{nd}-order beliefs are highly correlated with choices, is a clue that higher-order belief formation is not a simple iteration of belief formation. Furthermore, the link between self-focus suggested by insula activity and its negative correlation with low strategic IQ suggests that third-party 2^{nd}-order beliefs (C guessing what B thinks A will do) might be more accurate than self-referential 2^{nd}-order beliefs (A guessing what B thinks A will do). This novel prediction shows how neural evidence can inspire a fresh idea that would not have emerged from standard theory.

Note that the study of brain activation is not really intended to confirm or refute the basic predictions in game theory; that kind of evaluation can be done just by using choices [6]. Instead, our results provide some suggestions about a *neural* basis for game theory which goes beyond standard theories that are silent about neural mechanisms. Neural game theories will consist of specifications of decision rules and predictions about *both* the neural circuitry that produces those choices and its biological correlates (e.g., pupil dilation, eye movements, etc.). These theories should also say something about how behavior varies across players who differ in strategic IQ, expertise, autism, Machiavellianism, and so forth. Linking brain activity to more careful measurements of steps of strategic thinking is the next obvious step in the creation of neural game theory.

Bibliography

[1] R. Adolphs. Cognitive neuroscience of human social behavior. *Nature Reviews Neuroscience*, 1:165–178, 2003.

[2] J. Allman, A. Hakeem, and K. Watson. Two phylogenetic specializations in the human brain. *Neuroscientist*, 8:335–346, 2002.

[3] D. Barraclough, M. Conroy, and D. Lee. Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, 7:404–410, 2004.

[4] H. Cai and J. Wang. Overcommunication and bounded rationality in strategic information transmission games: An experimental investigation. *Games and Economic Behavior*, 56(1):7–36, 2006.

[5] C. Camerer, G. Loewenstein, and M. Weber. The curse of knowledge in economic settings. *Journal of Political Economy*, 97:1232–1254, 1989.

[6] C. F. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton University Press, Princeton, 2003.

[7] C. F. Camerer and T. Ho. Experience-weighted attraction (ewa) learning in normal-form games. *Econometrica*, 67:827–874, 1999.

[8] C. F. Camerer, T. Ho, and J. K. Chong. Behavioral game theory. thinking, learning, teaching. In S. Hück, editor, *Experiments, and Bounded Rationality: Essays in Honour of Werner Güth.* Palgrave Press, Basingsoke, 2004.

[9] C. F. Camerer, T.-H. Ho, and J. K. Chong. A cognitive hierarchy theory of one-shot games. *Quarterly Journal of Economics*, 119:861–898, 2004.

[10] C. F. Camerer, E. Johnson, S. Sen, and T. Rymon. Cognition and framing in sequential bargaining for gains and losses. In K. Binmore, A. Kirman, and P. Tani, editors, *Frontiers of Game Theory.* MIT Press, Cambridge, 1994.

[11] C. F. Camerer, G. Loewenstein, and D. Prelec. Neuroeconomics: Why economics needs brains. *Scandinavian Journal of Economics*, 106:555–580, 2004.

[12] C. F. Camerer, G. Loewenstein, and D. Prelec. Neuroeconomics: How neuroscience can inform economics. *J. Econ. Lit.*, 43:9–63, 2005.

[13] F. Castelli, C. Frith, F. Happé, and U. Frith. Autism, asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain*, 125:838–849, 2002.

[14] K. Chong, C. F. Camerer, and T. Ho. A cognitive hierarchy theory of games and experimental analysis. In R. Zwick, editor, *Experimental Business Research, volume II.* Kluwer Academic Press, 2005.

[15] M. Costa-Gomes and V. Crawford. Cognition and behavior in two-person guessing games: An experimental study. University of California at San Diego, Economics Working Paper Series 2004-11, Department of Economics, UC San Diego, Sept. 2004. available at http://ideas.repec.org/p/cdl/ucsdec/2004-11.html.

[16] M. Costa-Gomes and G. Weizsäcker. Stated beliefs and play in normal-form games. University of California Sand Diego Working Paper, 2004.

[17] M. V. Costa-Gomes, V. Crawford, and B. Broseta. Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69:1193–1235, 2001.

[18] D. J.-F. de Quervain, U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck, and E. Fehr. The neural basis of altruistic punishment. *Science*, 305:1254–1258, 2004.

[19] M. Dufwenberg and U. Gneezy. Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior*, 30:163–182, 2000.

[20] S. Dupont, V. Bouilleret, D. Hasboun, F. Semah, and M. Baulac. Functional anatomy of the insula: new insights from imaging. *Surg. Radiol. Anat.*, 25:113–119, 2003.

[21] N. I. Eisenberger, M. D. Lieberman, and K. D. Williams. Does rejection hurt? and fmri study of social exclusion. *Science*, 302:290–292, 2003.

[22] C. Farrer and C. Frith. Experiencing oneself vs. another person as bein the cause of an action: the neural correlates of the experience of agency. *Neuroimage*, 15:596–603, 2001.

[23] G. R. Fink, H. J. Markowitsch, M. Reinkmeier, T. Bruckbauer, J. Kessler, and W. D. Heiss. Cerebral representation of one's own past: neural networks involved in autobiographical memory. *Journal of Neuroscience*, 16:4275–4282, 1996.

[24] H. Gallagher, J. Anthony, A. Roepstorff, and C. Frith. Imaging the intentional stance. *Neuroimage*, 16:814–821, 2002.

[25] H. Gallagher and C. Frith. Functional imaging of “theory of mind”. *Trends in Cognitive Sciences*, 7:77–83, 2003.

[26] T. Gilovich and V. Medvec. The illusion of transparency: biased assessments of others’ ability to read emotional states. *Journal of Personality and Social Psychology*, 75:332–346, 1998.

[27] H. Gintis. Towards a unity of the human behavioral sciences. In S. Rahman, J. Hsymons, D. M. Gabbay, and J. P. van Bendegem, editors, *Logic, Epsitemology, and the Unity of Science.* Kluwer Academic Press, 2004.

[28] P. Glimcher and A. Rustichini. Neuroeconomics: The concilience of brain and decision. *Science*, 306:447–452, 2004.

[29] P. W. Glimcher. *Decisions, uncertainty and the brain: The new science of neuroeconomics.* MIT Press, Cambridge, 2003.

[30] J. Goeree and C. Holt. A model of noisy introspection. *Games and Economic Behavior*, 46:365–382, 2004.

[31] R. Gonzales and G. Loewenstein. Effects of circadian rhythm on cooperation in an experimental game. Carnegie Mellon Working Paper, 2004.

[32] J. Greene and J. Haidt. How (and where) does moral judgment work? *Trends in Cognitive Science*, 6:517–523, 2002.

[33] D. Grether, C. Plott, D. Rowe, M. Serano, and J. M. Allman. an fmri study of selling strategy in second price auctions. Caltech working paper no. 1189, 2004.

[34] A. Gunnthorsdottir, K. McCabe, and V. L. Smith. Using the machiavellianism instrument to predict trustworthiness in a bargaining game. *Journal of Economic Psychology*, 23:49–66, 2002.

[35] T. Hedden and J. Zhang. What do you think i think you think? theory of mind and strategic reasoning in matrix games. *Cognition*, 85:1–36, 2002.

[36] E. Hill and D. Sally. Dilemmas and bargains: Autism, theory of mind, cooperations and fairness. Working Paper, University College London, 2004.

[37] M. Hsu, M. Bhatt, R. Adolphs, D. Tranel, and C. F. Camerer. Neural systems responding to degrees of uncertainty in human decision-making. *Science*, 310:1680–1683, 2005.

[38] S. Huettel, A. Song, and G. McCarthy. *Functional Magnetic Resonance Imaging.* Sinauer Associates Inc., 2004.

[39] E. J. Johnson and C. F. Camerer. Thinking backward and forward in games. In I. Brocas and J. Castillo, editors, *The Psychology of Economic Decisions, Volume 2: Reasons and Choices*. Oxford University Press, 2004.

[40] E. J. Johnson, C. F. Camerer, S. Sen, and T. Rymon. Detecting failures of backward induction: Monitoring information search in sequential bargaining. *Journal of Economic Theory*, 104:16–47, 2002.

[41] G. Loewenstein, D. Moore, and R. Weber. Misperceiving the value of information in predicting the performance of others. *Experimental Economics*, 9(3):281–295, 2006.

[42] K. McCabe, D. Houser, L. Ryan, V. Smith, and T. Trouard. A functional imaging study of cooperation in two-person reciprocal exchange. *Proceeding of the National Academy of Sciences*, 98:11832–11835, 2001.

[43] K. McCabe and V. L. Smith. Neuroeconomics. In L. Nadel, editor, *Encyclopedia of Cognitive Sciences*. MIT Press, 2001.

[44] E. K. Miller and J. D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neurosciences*, 24:167–202, 2001.

[45] R. Nagel. Inraveling in guessing games: An experimental study. *American Economic Review*, 85:1313–1326, 1995.

[46] M. L. Platt and P. W. Glimcher. Neural correlates of decision variables in parietal cortex. *Nature*, 400:233–238, 1999.

[47] M. Rabin. Incorporating fairness into game theory and economics. *American Economics Review*, 83:1281–1302, 1993.

[48] J. K. Rilling, D. A. Gutman, T. R. Zeh, G. Pagnoni, G. S. Berns, and C. D. Kilts. A neural basis for social cooperation. *Neuron*, 35:395–405, 2002.

[49] A. G. Sanfey, J. K. Rilling, J. A. Aronson, L. E. Nystrom, and J. D. Cohen. The neural basis of economic decision-making in the ultimatum game. *Science*, 300:1755–1758, 2003.

[50] W. Schultz. Multiple reward signals in the brain. *Nat. Rev. Neurosci.*, 1:199–207, 2000.

[51] M. Siegal and R. Varley. Neural systems involved in 'theory of mind'. *Nature Reviews Neuroscience*, 1:463–471, 2002.

[52] T. Singer and E. Fehr. The neuroeconomics of mind reading and empathy. *American Economic Review*, 95(2):340–345, 2005.

[53] T. Singer, S. Kiebel, J. Winston, R. Dolan, and C. Frith. Brain responses to the acquired moral status of faces. *Neuron*, 41:653–662, 2004.

[54] D. Stahl and P. Wilson. Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization*, 25:309–327, 1994.

[55] J. Weinstein and M. Yildiz. Finite-order implications of any equilibrium. MIT working paper number 04-06, http://ssrn.com/abstract=500202, 2004.

[56] P. Zak. Neuroeconomics. *Philosophical Trasactions of the Royal Society*, In press.

[57] P. Zak, R. Kurzban, and W. Matzner. The neurobiology of trust. *Annals of the New York Academy of Sciences*, 1032:224–227, 2004.

Appendix B

B.1 Methods

Pairs of subjects were recruited on campus at Caltech through SSEL lab recruiting software.[1] One subject performed the tasks in the scanner, as the row player, while the other performed them in an adjacent room, as the column player. These three tasks were given in a random order for each game to control for order effects.

In the scanner each subject proceeds through a series of screens (like Figure 2.1) one at a time, at their own pace. They press buttons on a box with 4 buttons to record their responses (choosing a row strategy in C and 2B tasks, and a column strategy across the bottom of the screen in B tasks). After each response is recorded, there is a random lag from 6-10 seconds with a "fixation cross" to hold their visual attention in the center of the screen. The entire set of tasks took from 7 to 15 minutes.

At the end of the experiment 1 of the 24 tasks was chosen at random and subjects were paid according to their payoffs in the games at a rate of $0.30 a point, if a choice task was picked, or were given $15 for a correct answer to the belief tasks. All payments were in addition to a $5 showup fee.

Subjects in the scanner were debriefed after the experiment to control for any difficulties in the scanner and to get self-descriptions as to their strategies. The most common strategy described was a hybrid between cooperation and self-interest where they acted largely to maximize their own payoffs, but would cooperate if a small loss to herself would result in a large gain to the other player.[2] Some subjects seemed empirically more cooperative than others, but we treated all subjects similarly in our analysis.

To do the scanning, we first acquired a T1-weighted anatomical image from all row players. (This is a sharper-resolution image than the functional images taken during behavior so that we can map areas of activation onto a sharper image to see which brain areas are active.) Functional images

[1] Since Caltech students are selected by the admissions committee, for their unusual analytical skill, they are hardly a random sample. Instead, their behavior is likely to overstate the average amount of strategic thinking in a random population. This is useful, however, in establishing differential activation of regions for higher-order strategic thinking since the subjects are likely to be capable of higher-order thinking in games that demand it.

[2] Subjects reporting this strategy included some who had taken one or more classes in game theory and were familiar with the concept of Nash equilibrium.

were then acquired while subjects in the scanner played with subjects outside the scanner. They were acquired with a Siemens 3T MRI scanner using a T2-weighted EPI (TR = 2000msec TE = 62 ms, 34 (32 for smaller heads) 3mm slices), 32-34 slices depending on brain size. The slice acquisition order was (2, 4, 6, . . ., 1, 3, 4, . . .). Data was acquired with one functional run per subject.

Data were analyzed using SPM2. Data were first corrected for time of acquisition, motion-corrected, coregistered to the T1-weighted anatomical image, normalized to the MNI brain and smoothed with an 8mm kernel. The data were then detrended using a high-pass filter of periods greater that 128 seconds and an AR(1) correction.

For each analysis the general linear model was constructed by creating dummy variables that were "on" from the stimulus onset time until the decision. When a subtraction is measured, the difference between activation in task A and task B, the dummy variable is +1 when the task A stimulus is present and -1 when the task B stimulus is present. These dummy variables were convolved with the standard hemodynamic response function. Standard t-tests were used to determine whether coefficient on one dummy variable is greater than that on another. Data from all the subjects were combined using a random-effects model. The cross-subject regressions regress t-statistics of treatment affects across voxels against behavioral measures of strategic IQ.

B.2 Data Tables

The following tables summarize the order of games and tasks, raw choices of subjects and all regions seen in the scans discussed in the main text.

Game	CGCB transform	Task Order	Game Type
1	2A(-10, -5; AA-BB)	C, B, 2B	row player has dominant strategy
2	3A(-20, +10)	2B, C, B	column player has dominant strategy
3	5A(+15, -13; A-C)	2B, B, C	3 × 2 game, 3 steps of dominance for row player
4	5B(-7, +11, B-C)	B, 2B, C	3 × 2 game, 3 steps of dominance for row player
5	6A(-17, -3; AA-BB)	C, 2B, B	2 × 3 game, 2 steps of dominance for row player
6	6B(+7, +0; AA-CC)	B, C, 2B	2 × 3 game, 2 steps of dominance for row player
7	9A(+19, +19; A-C)	C, B, 2B	row player has a dominant strategy
8	9B(0, 0)	B, C, 2B	column player has a dominant strategy

Table B.1: Order of games, transformation from original CGCB games, and order of tasks for each game. Note: In the "CGCB tranform" column, in notation Gx $(r, c; Y-Z)$ Gx denotes name and letter, r and c are constants added to original CGCB payoffs to transform them to experimental currency payoffs we used, and $Y-Z$ denotes original rows or columns that are switched to create our matrices. Example: Our game 3 (see text, Figure 2.1) is CGCB game 5A with 15 added to all row payoffs, 13 subtracted from all column payoffs, and rows A and C switched. In game 6 there waws a math error in one cell: for (B, AA) in our game we added 6 instead of 7 to the corresponding cell in CGCB, this did not change the strategic structure of the game.

	A		B		C		D		AA		BB		CC		DD	
#	New	C	New	C	New	C	New	C	New	C	New	C	New	C	New	C
1	.25	.21	.75	.79	n/a	n/a	n/a	n/a	.61	.69	.39	.31	n/a	n/a	n/a	n/a
2	.50	.86	.50	.14	n/a	n/a	n/a	n/a	.61	.92	.39	.08	n/a	n/a	n/a	n/a
3	.31	.21	.56	.79	.13	.00	n/a	n/a	.39	.23	.61	.77	n/a	n/a	n/a	n/a
4	.25	.14	.63	.71	.13	.14	n/a	n/a	.44	.46	.56	.54	n/a	n/a	n/a	n/a
5	.44	.79	.56	.21	n/a	n/a	n/a	n/a	.22	.38	.17	.00	.61	.62	n/a	n/a
6	.50	.36	.50	.64	n/a	n/a	n/a	n/a	.56	.77	.22	.08	.22	.15	n/a	n/a
7	.38	.08	.00	.00	.06	.00	.56	.92	.56	.46	.44	.54	n/a	n/a	n/a	n/a
8	.38	.07	.63	.93	n/a	n/a	n/a	n/a	.11	.08	.00	.00	.17	.00	.72	.92

Table B.2: Frequency of strategy choices A-D and AA-DD in our study vs. Costa-Gomes et al (2001) data. (CGCB data denoted "C"; n/a. denotes strategies that did not exist in a particular game)

	Choice (C)			Belief			2^{nd} Order Belief (2B)		
	25%	50%	75%	25%	50%	75%	25%	50%	75%
Game 1	11.4	20.4*	26.2	11.3	12.5	18.3	5.78	8.58	13.7
Game 2	8.87	11	20.9	6.58	7.75	13.5	14.5	22.3*	25.5
Game 3	8.58	10.7	16.3	9.61	11.2	20.2	16.8	25*	42.8
Game 4	2.91	7.83	15	11.4	16.6*	32.9	6.08	10.8	23.9
Game 5	18.6	24.9*	37.3	6.55	11.6	16.7	7.92	10.1	23.9
Game 6	8.1	9.5	13.4	19.6	25.2*	42.8	4.61	6.54	15.1
Game 7	17.6	25.5*	42	6.08	9.23	14.1	6.58	10	17.3
Game 8	6.17	8.05	12.2	15.8	20.9*	26	5.67	11.1	13.8

Table B.3: Distributions of free response times (25^{th}, 50^{th} – median– and 75^{th} percentiles) in seconds across tasks and games. Note: * Denotes task which was presented first (e.g., the 2B task was first in game 3). Response times are typically about twice as long for the first task presented.

Comparison	Significance Threshold	Area	x	y	z	cluster size	t-stat
Choice > Belief (all games, all subjects)	$p = .001$	R Occipital Lobe	9	-78	9	202	6.77
		Cingulate Gyrus	-3	-12	33	24	5.12
		L Dorsolateral	-27	48	9	14	4.74
		ACC	6	42	0	33	4.62
		Frontal Insula	-42	12	-18	31	4.60
		R Cerebellum	9	-42	-27	17	4.49
		R Insula	36	12	-3	6	4.10
2^{nd} Order Belief > Belief (Out of Equilibrium Games Only)	$p = .001$	L Insula	-42	0	0	3	4.44
		Inferior Frontal Gyrus	45	33	0	8	4.85
	$p = .002$	L Insula	-42	0	0	9	4.44
		Inferior Frontal Gyrus	45	33	0	13	4.85
Choice-task activity *nagatively* correlated with SIQ (games w/ dominant strategies excluded)	$p = .0005$	L Insula	-42	6	-3	12	5.34
		BA 11	-24	45	-15	6	5.47
		R Cerebellum	9	-78	-18	6	5.28
Choice-task activity *positively* correlated with SIQ (games w/ dominant strategies excluded)	$p = .001$	Precuneus	3	-66	24	13	4.90
	$p = .05$	Caudate	12	0	15	11	2.52
		Precuneus	3	-66	24	312	4.90
		R Occipital/Cerebellum	18	-87	-21	33	3.61
		Precentral Gyrus	-42	-18	42	45	2.90
		Occipital Gyrus	-27	-63	-12	12	2.35
		L Occipital	-36	-84	-15	6	2.28
		R Occipital	48	-69	36	13	2.24
Choice > Belief (in equil.)	$p = .001$	Ventral Striatum	-3	21	-3	20	5.80
Choice > Belief (out of equil.)	$p = .005$	ACC	6	42	0	13	3.17
		ACC	15	42	0	13	3.33
	$p = .001$	Paracingulate	15	36	33	39	5.93
		L Dorsolateral	-30	30	6	14	4.85
		R Occipital	12	-75	-6	19	4.84
		R Occipital	30	-60	9	12	4.73

Table B.4: Coordinates (x, y, z), cluster sizes (k), and t-statistics for subtractions and activity-behavior correlations reported in the text. R and L denote right and left hemispheres, respectively.

B.3 Instructions to Subjects

This is an experiment on decision-making. The decisions you make will determine a sum of money you will receive at the end of this experiment. If you read these instructions carefully, you stand to earn a substantial sum of money.

The questions in this experiment will all involve playing “matrix games” . For the duration of the experiment Player 1 will be the “row player” and Player 2 will be the “column player”. You will be shown a series of game that look something like this:

	Player 1’s payoff			Player 2’s payoffs		
	AA	BB	CC	AA	BB	CC
A	15	16	35	6	20	7
B	10	20	30	7	23	10
C	20	17	36	0	7	3

In these games the row player chooses a row and the column player chooses a column. Above, the row player would choose A, B or C and the column player would choose AA, BB, or CC. You will both make these decisions simultaneously and the cell that is determined by your choices determines your payoff. For example: If in the above example the row player had chosen B and the column player had chosen CC The row player: Player 1, would receive 30 points and the column player: Player 2, would receive 10 points. If on the other hand Player 1 had selected C and Player 2 had selected BB the payoffs would be 17 for Player 1 and 7 for Player 2.

In addition to playing the games you will be asked some questions about the games during the course of the experiment. You will be asked what you think the other player will choose, and what you think the other player believes *you* will choose. These questions will be mixed in with the games in a random order so pay close attention to the question at the top of the screen. If you are player 2 (outside the scanner) you may not go back and forth among the questions.

Payment

In addition to playing the games you will be asked some questions about the games during the course of the experiment. At the end of the experiment we will select one game or question and award you for your performance on that game or question. You will earn $15 15 for a correct answer to a question, or $0.30 a point for points earned in the game. In addition you be given a $5.00 show-up fee.

Questions:

1. What is your age?

2. What is your sex? (F/M)

3. Are you left handed or right handed?

4. Have you taken any courses in Economics of Game Theory. If so please list these below.

   (a)

   (b)

   (c)

   (d)

   (e)

5. In game (a) below, if the row player chooses C and the column player chooses AA, what are both players' payofs?

6. Practice games - If you're player 1 choose a row. If you're player 2 choose a column.

		Player 1's Payoffs			Player 2's Payoffs		
		AA	BB	CC	AA	BB	CC
(a)	A	10	12	48	20	19	12
	B	5	30	25	78	42	60
	C	20	13	0	50	7	9
	D	43	16	27	15	10	13

		Player 1's Payoffs			Player 2's Payoffs		
		AA	BB	CC	AA	BB	CC
(b)	A	0	-1	1	0	1	-1
	B	1	0	-1	-1	0	1
	C	-1	1	0	1	-1	0

Chapter 3

Neural activity during bargaining with private information

with Terry Lohrenz, Read Montague, and Colin F. Camerer

3.1 Introduction

Strategic interaction lies at the heart of social science. Game theory provides a precise language to study strategic interaction and provides analytical tools for predicting what will happen when people interact. In a game, players choose strategies given information about the strategies and private information about other players. Strategy choices create outcomes which players are assumed to value, in a way that can be expressed by numerical *utility.*

The central complication in analysis of a game is what players believe other players will do. Beliefs, and the sensible strategy choices that result given different beliefs will lead to different outcomes. The key questions in predicting outcomes are first, what characterizes a "sensible" strategy choice, and second, what determines our beliefs.

Economists have traditionally gotten around these questions using equilibrium concepts. A sensible strategy is simply a best response, i.e., expected utility maximizing response, given beliefs. The question of belief is replaced by that of mutual best response. However equilibrium concepts like Nash equilibrium give mixed results as predictors of human behavior [4, 3].

An important canonical game in economics, political science, biology, and other fields is bargaining. The exchange of goods, either for other goods or for currency, is a basic activity that occurs in any human culture. However this sort of game requires both parties to have a good understanding of the desires of the other.

In this paper we study a simple version of this game with "cheap-talki," a limited round of communication with no direct effect on the outcome of the game. A buyer and seller bargain briefly over the price at which the seller will trade an object with zero cost in one period of bargaining. Since the seller's cost is zero, she could sell at any price. The complication is that the buyer's value for the object, v, is drawn from a uniform distribution of integers from 1 to 10. The buyer knows her value but the seller only knows that the value is uniformly distributed.

The first step in the game is that the buyer "suggests" a price s. This suggestion is costless and has no direct payoff consequences. The seller hears the suggestion s and names a take-it-or-leave-it price p. If the price p is below or equal to the value v ($p \leq v$) the trade takes place at the price p. The seller then earns $p - 0 = p$ and the buyer earns the surplus $v - p$. Notice that the suggestion s does not enter directly into what the buyer and seller earn; it just serves as a communication device which might influence the price the seller demands.

Despite its simplicity, this game is interesting from both an economic view and from the viewpoint of social neuroscience. In the standard analysis (which assumes self-interest and no computational costs), the seller realizes the buyer's suggestion is always designed to get her (the seller) to name a lower price. The seller should therefore ignore s, and name the price which maximizes her expected profit, either 5 or 6 (both prices are equally profitable). Anticipating the fact that the seller ignores s, the buyer should "babble," in game theory jargon, and choose a value of s which is unrelated to v. However, since the buyer's value is always above the seller's cost (zero), a failure to trade is socially inefficient. This game therefore features a conflict between the social desire to always trade, and the individual desires to get the best price.

In experiments of this general type, buyers generally "over-communicate" (there is a measurable correlation between the true state and the signal sent) and sellers are sensitive to the buyers' suggestions, contrary to the prediction of theory. The gains from exchange that result from this pattern are higher for the two players together than if there was babbling by the buyer and ignorance of the suggested price by the seller [14, 17, 3, 2].

Our bargaining game is of interest in social neuroscience too. Compared to other species, humans are highly social and have created complex social architectures for trade, transmission of cultural practices and information, and institutional rules such as laws and social norms. These practices are presumably supported by neural architecture for decoding and creating facial expressions, processing abstract signals of who can be trusted, and creating reasonable "models" of others' behavior: understanding both their desires and how those desires are linked to their actions. While there are some fMRI studies of social exchange in games involving trust [10, 16] and randomization [8], there are no studies of the sort of strategic communication that results from bargaining.

An important open question is why and when over-communication occurs: Are one or both sides making a strategic mistake? Are they being pro-social and cooperative, knowing that choosing an informative value of s and responding to it will increase how much they make together? Or is there something even more basic going on, involving how our brains are wired to process communication from another person.

3.2 Background

We chose to study bargaining because it provides a natural strategic environment for subjects, bargaining over an object is a familiar situation to many adults; and because the incomplete information in the problem, namely the private values for the buyer and sellers, provide a rich environment to explore belief formation. In addition bargaining is a well-understood problem theoretically. Myerson and Satterthwaite discussed the *inherent* inefficiency of bilateral bargaining using any incentive compatible mechanism [12].

We chose to study cheap talk in these games for several reasons. First, experiments show that cheap talk will increase the average efficiency of bilateral bargaining games [17]. Second, the core purpose of cheap talk is to manipulate another player's beliefs. Crawford and Sobel [6] characterized the Bayes-Nash equilibria of a general class of sender-receiver games. In these games there is some randomly chosen state of the world that is revealed to the "sender," the sender is then allowed to send some signal s to a "receiver." After seeing this signal the receiver takes an action $a \in A$ that determines the outcome of the game.

They predict that the level of information transmission is proportional to how well aligned the player's preferences are in games similar to the task we study here [6], but as mentioned above, in bargaining games the buyer's interests are almost directly opposed to the seller. In addition they show that any equilibrium is equivalent to one of the class described below:

- Let R be the possible states of the world and S be the set of possible signals. There exists some signaling function $\sigma\, R \rightarrow S$ such that the sets $\sigma^{-1}(s)$ create a partition over R.
- The receiver has some response function $\rho\, S \rightarrow A$ such that $\rho(s)$ maximizes the receiver's expected payoff conditional on the fact that the state of the world is an element of $\sigma^{-1}(s)$.

The allows us to use the number of sets in the partition induced by σ^{-1} as a measure of how informative the equilibrium is. For example if there is only one element in this partition the equilibrium is completely uninformative, while if elements of the partition each contain exactly one element the equilibrium is completely informative, i.e. the sender communicates the true state perfectly.

3.2.1 Neuroscience

From the neuroscience perspective, we chose this problem because it allows us to explore the basis of deviations from "rational" behavior in social situations away from a moral context. In the past few years there have been a number of imaging studies examining the neural correlates of social emotions such as trust and resentment [16, 7, 15]. These have focused largely on the *emotional* factors of decision making. This paper focuses more on the constraints neural function may place on social reasoning divorced from moral implications. Cheap talk is a form of strategic information transmission, which means that it may include strategic deception. The ability to deceive is non-trivial from a neural perspective since it requires the deceiver to understand that another person may have a different perception of the world than their own. Children do not usually develop this ability until they are about 4 years old. From the perspective of the person receiving information, responding to possible deception may be difficult since it requires a decision maker to ignore some or all of the information given. Economists almost implicitly assume that all information from another decision maker is suspect until proven otherwise (i.e., through reputation building or shared goals), taking this sort of neural perspective may indicate that the brain's default is to trust information.

One well-known task in both psychology and neuroscience is the "Stroop Task." In this task subjects are shown the names of colors in different colored ink. At the beginning of each trial the subjects are told to either read the word or name the color of the ink. There are essentially two types of stimuli: congruent stimuli where the word and color match, e.g., red in red ink; and incongruent stimuli where the word and color as mismatched, e.g., blue in green ink. MacDonald et al. found that the dorsolateral prefrontal cortex (DLPFC) was differentially activated by the color-naming task, the more cognitively difficult task, than the word-reading task implying that this area may be instrumental in implementing cognitive control, whereas they found that the anterior cingulate cortex (ACC) was differentially activated by incongruent versus congruent stimuli. We find significant correlations to both of these areas during our task, supporting the hypothesis that departures from rational behavior may be the result of automated information processing in the brain.

The DLPFC has also been implicated in cognitive control in the ultimatum game. Sanfey et al. found that low offers elicited activity in the DLPFC, and Knoch et al. found that disrupting activity in the right DLPFC (but not the left DLPFC) reduces a subjects' rate of rejection for unfair offers [15, 11]. These results suggest acceptance of the offer may be the more basic or automatic response to ultimatum offer and subjects need to exert cognitive control to reject offers as a means of punishing the proposer[1].

Other neural regions that seem consistently active over many types of decision-making tasks are the orbital frontal cortex (OFC) and amygdala. Hsu et al. found that risky or ambiguous situations elicit activity in both of these areas, the the activity increasing as the level of information decreased. In addition, patients with damage to the orbitofrontal area showed significantly less risk and ambiguity aversion than patients with comparable damage to other areas.

For the purposes of this study we will be particularly interested in two systems. The first one involves the Anterior Cingulate Cortex (ACC) and the DLPFC, and is instrumental for implementing cognitive control and resolving conflicting response impulses. The second is a system that includes various parts of the basal ganglia (sub-cortical areas of the brain, these are some of the oldest regions of the brain from an evolutionary perspective) and the OFC. This second system is thought to be necessary for assigning reward values to objects or decisions.

3.3 The Bargaining Task

Subjects are recruited in pairs and read detailed instructions. Following experimental economics conventions, all the details of the protocol (except the specific values of the buyer in each stage) are known by both players. This convention means that it is possible for players to arrive at an economic equilibrium, in theory, simply by reasoning and introspection (so if they deviate from equilibrium, it is not because they were confused or misled). It also means that noise from perceptions about the nature of the experiment and the information of the other player are minimized, so we can study neural activity that arises solely from perceptions of the other players' hidden value and strategy.

At the beginning of each round the players were told whether they were the "buyer" or the "seller" in that round. Buyers were told their random private value for a hypothetical good, distributed identically and independently across trials over integers 1 through 10. Sellers never had any intrinsic value for the good (as if they were producing perishable goods at zero marginal cost), to simplify the game. In this simple structure, trades should always take place since the buyer's value is always above the seller's cost of zero (i.e., trade is *always* efficient). Subjects switched roles every five

[1] This interpretation is more congruent with anthropological studies showing that societies where people have less social interaction are less likely to reject low offers in the ultimatum game [9].

rounds.

The task had two stages at which subjects make choices affecting the game:

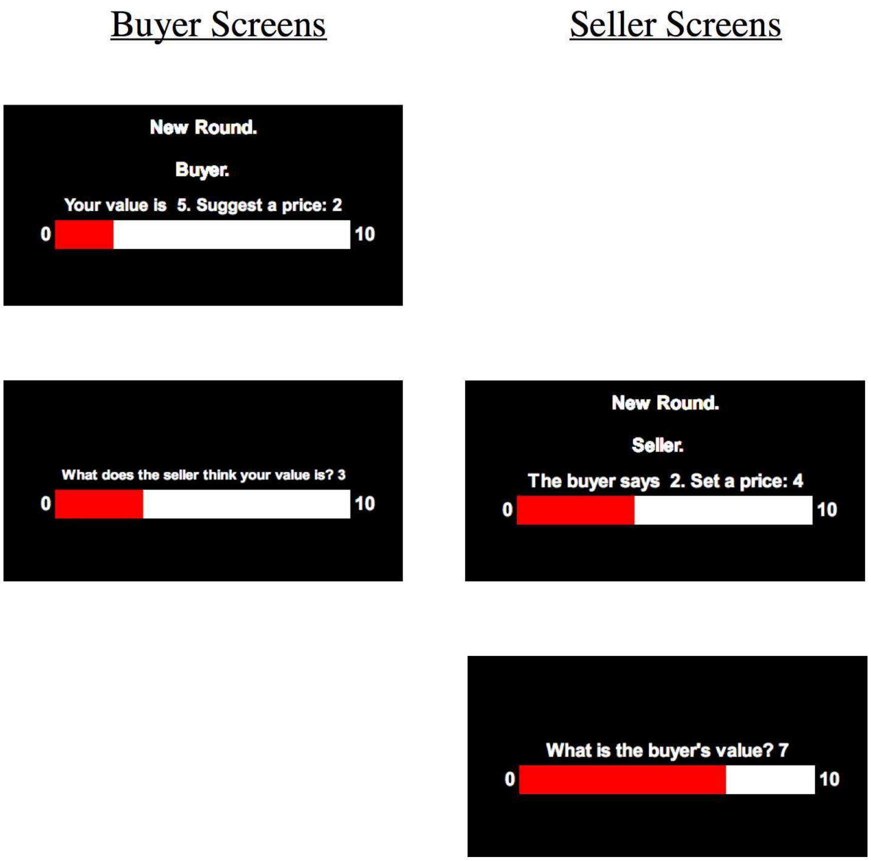
- Stage 1: The buyer is informed of her private value, v and asked to "suggest a price." We call the buyer's suggested price s.

- Stage 2: The seller sees s and is asked to choose a price p. If $p \leq v$ the trade occurred in that round; the buyer receives the difference between the value and the price (the surplus in economic terms) $v - p$ and the seller receives the net profit p.

In addition we asked subjects directly about their beliefs. For the seller this was always a first-order belief: What do you think the buyer's value is? However for the buyer this is always a second-order belief, i.e., a belief about someone else's beliefs: What does the seller *think* your value is? Subjects were rewarded for their answers to these questions according to their accuracy, getting one point for the correct answer and a decreasing payoff as their answers diverged from the correct answer. This way, we incentivized honest reporting, but the payoffs for belief elicitation were far outweighed by the payoffs to the game itself.

Because we wanted to image brain activity of both the buyer and seller in a pair simultaneously, we used a partner protocol in which two subjects play each other repeatedly. Games like this are sensitive, in theory, to what two partners know about the history. This reputational history-dependence is a huge source of complication in analyzing brain activity since the signal-to-noise ratio is so low in fMRI. It is therefore usually necessary to pool many trials which are informationally similar, making feedback impractical. Therefore neither player received any direct feedback about the outcome of each trial. Subjects were not even told whether a trade had taken place. This allows us to treat each trial as approximately independent, and to analyze behavior as if they are participating in a one-shot game.

Each subject performed this task 60 times, 30 as a buyer and 30 as a seller, switching roles every 5 rounds. The role switching allows us to compare behavior as buyer and seller within each subject, which is statistically advantageous. They were not informed of their total earnings until the very end of the experiment. Since the sellers never received information about the buyer value, subjects never learned directly about their opponent's suggested pricing strategy. Similarly, since the buyers never learned whether a trade occurred, they could not learn directly about the seller's pricing strategy.



Buyer Screens

Seller Screens

New Round.

Buyer.

Your value is 5. Suggest a price: 2

What does the seller think your value is? 3

New Round.

Seller.

The buyer says 2. Set a price: 4

What is the buyer's value? 7

Figure 3.1: Screens seen by the buyer and seller during one trial

However, sellers could conceivably make inferences based on the *range* of suggestions they received over course of the experiment, a possibility we investigate below.

3.3.1 Predictions

A theoretical analysis of the one shot game shows that if a seller's strategy is at all sensitive to suggested prices, high-value buyers will all pool with the low-value buyers, transmiting minimal information and making the seller's strategy non-optimal (see appendix for formal details). So the equilibrium prediction is that suggested price should be completely unrelated to value ($Prob(v|s) = .1$ for all v, s), and the seller will choose p maximize $u(p)(11-p)/10$, for a risk-neutral seller this implies that he should always choose $p = 5$ or $p = 6$.

Keep in mind that in the theory above, subjects are forming *correct* beliefs about what other players will do and maximizing their own personal gains given their beliefs. As a useful benchmark contrast, suppose both players were trying hard to cooperate and earn the most money together from the task (e.g., suppose they had planned beforehand to split their earnings equally, so they want to maximize the total gain). Since the seller always wants to sell the good to maximize their joint gains, the seller should try to always name a price the buyer can afford. One way to do this is to always name a price of 1. Another way to do this is for the buyer to make a suggestion which is communicates the value perfectly, i.e., the suggestion correspondence has a functional inverse $f(s) = v$, and for the seller to choose $p \leq f(s)$. Such patterns could emerge, in theory, but neither emerged in this experiment, instead we observe partial information revelation of the sort you might expect if the buyer and seller incentives were somewhat aligned.

While subjects were not given any feedback about the outcome of each trial, both the buyer and seller are aware of the buyer's history of suggestions over the course of the experiment. This raises the question, do sellers pay attention to the buyer's "reputation" and, if they do, do buyers anticipate this by revealing information. We consider two ways in which the history of suggestions might create repeated game effects. First, we consider purely strategic models where sellers attempt to glean information from not only the current suggestion, but the entire history of suggestion. Second, we will look at non-standard preferences that include a taste for retaliation or fairness.

One simple way to model information extraction from a buyer's history is to look at an extension of the cognitive hierarchy model [4]. Assume that level-0 buyers use a strategy of the form $s = \max(1, \lfloor \alpha v \rfloor)$, with α drawn randomly from $[0, 1]$. Higher-level sellers assume that any given level-0 player uses the same α throughout the experiment, so they can infer their opponents type from the suggestions they send. Level-n buyers choose signals to maximize their sum payments for the current period and the n following period (sellers do not consider future payoffs since their actions, unlike buyer actions, will not affect buyer beliefs in the future).

First let us consider the level-1 sellers. Before the first round they have a uniform prior on the value α that their opponent uses to generate suggestions. At the beginning of the first round they will receive a signal from the buyer, s. The seller can update their prior on the value of α using this signal and Bayes Law. For example, if $s = 8$, the seller will update their perceived probability distribution on α so that they believe that $P(\alpha < .8) = 0$. Sellers then best respond to s given their new updated prior. In each subsequent round the seller will update their prior again based on the new signal. Here we can make two different assumptions. First, the sellers may use the entire history of signals to generate a new prior. In this case if the seller ever sees a particularly high signal, like $s = 1$, beliefs will always be such that α must be high and sellers will always be credulous. Second, sellers may have a limited memory, and thus only update the original uniform prior on the value of α using the current signal s_0 and the signal from the previous round s_{-1}. Here, the effects of a high signal are transient.

Level-1 buyer behavior is uninteresting. Since they believe that the sellers are credulous, they will always choose a signal $s = 1$. However, Level-2 buyers will anticipate the fact that their cheap talk signal s will affect the seller's behavior. When the buyer believes that seller memory is unlimited, they can build up their credibility, i.e., manipulate the sellers to believe that $\alpha = 1$, but signaling 10 just once. They will signal $s = 1$ in all subsequent rounds. If, on the other hand, buyers believe that sellers have a limited memory, there will be an inverse relationship between the buyer's value v and the signal chosen s. In essence, buyers build credibility when they have low values by sending high signals in order to gain more surplus when they have high values. The average best response by a buyer given their previous signal s_{-1} and current value v-are shown in Figure 2 below.

A model where sellers have retaliatory preferences, i.e., they enjoy punishing buyers who send low signals, yields buyer behavior similar to CH level-2 buyers since sellers never know if their punishment is justified. Buyers will still use rounds where they have low values, and therefore little potential payoff, to raise their average signal and fend off retaliation by the seller. When they have higher values, they will send low signals to get better prices.

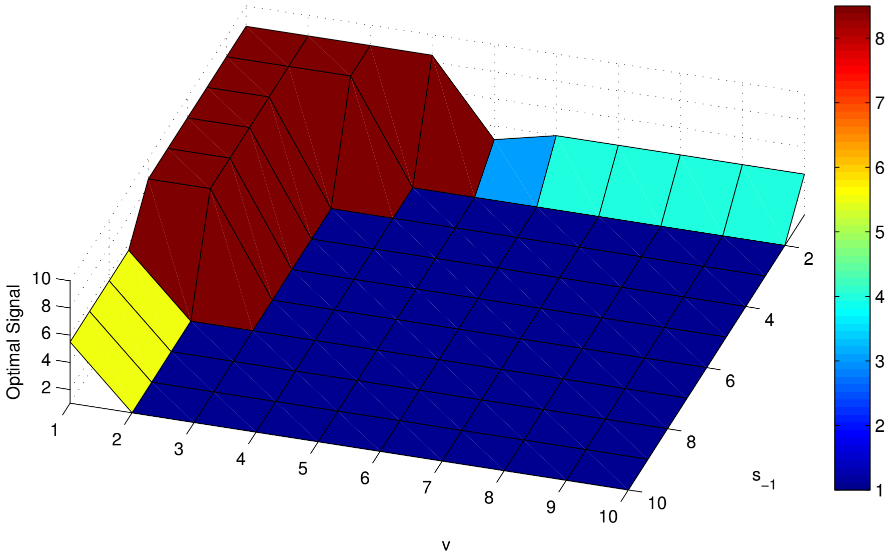


Figure 3.2: The average best cheap talk signal (sometimes the best response set has multiple values) for a level-2 buyer given $\tau = 1.5$ over all possible values v, and previous signals, s_{-1}, when the seller has a limited memory

Model	Predicted Buyer Behavior	Predicted Seller Behavior
Nash	Babbling	$p = 5$ or 6
Cognitive Hierarchy with full memory		
level-0	$s = \max(1, \lfloor \alpha v \rfloor)$	$p = s$
level-1	$s = 1$ for all v	$argmax_p(p \cdot Pr(v \geq p \vert \{s_t\}))$ where $\{s_t\}$ is the series of signals until the current time
level-2	$s = 10$ early in the experiment, $s = 1$ from there on	$argmax_p\left(p \cdot \left(Pr(l=1 \vert \{s_t\})\frac{11-p}{10} + Pr(l=0)Pr(v \geq p \vert \{s_t\})\right)\right)$
Cognitive Hierarchy with limited memory		
level-1	$s = 1$	$argmax_p(p \cdot Pr(v \geq p \vert s_0, s_{-1}))$
level-2	$s = 10$ for low v, $s = 1$ otherwise	Similar to above but priors depend only on recent information.

Table 3.1: Predicted buyer and seller behavior using a cognitive hierarchy model

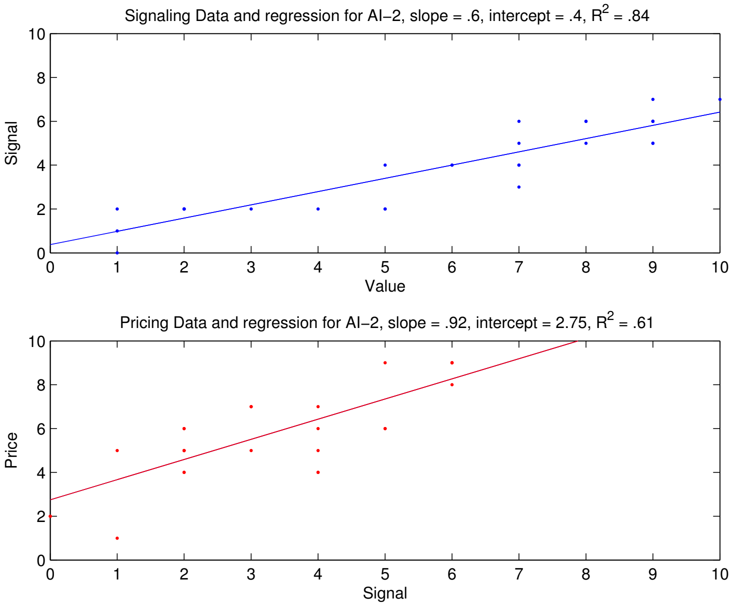


Figure 3.3: Suggested prices sent by a single subject as buyer, and prices set as seller. (Points are jittered by adding random noise so that identical points are plotted separately).

3.4 Behavioral Results

In a study like this which is aimed at two audiences, it is always helpful if the behavioral patterns themselves are surprising and of interest to readers who are not interested in neural activity. Roughly speaking, buyers' suggested prices are often remarkably revealing of their values. Buyers not only sent suggestions which are statistically informative (in the sense that s is realted to v), but they seemed to use linear strategies resulting in bins similar to those predicted by Crawford and Sobel for games where player's incentives are partially aligned. Furthermore, sellers seem to anticipate informativeness of price suggestsions and set their prices p based on the suggestions s also using roughly linear strategies. Interestingly, however, a single player's strategies in their role as buyer were not generally best responses to their own strategies as a seller (and vice versa).

Figure 1 reports all the data from a single subject. The top graph shows the subject's suggested prices s against her valuations v. The bottom graph shows the prices chosen by that same subject, p against the suggested prices she saw s. Both look roughly linear and have much more statistical association than predicted by equiliibrium theory. As a buyer, she seems to "shave" prices by suggesting a fraction of the value. As a seller, she seems to follow a markup strategy of adding a small constant to the buyer's suggestion.

Looking at these behavioral data for each subject suggests that they can be characterized by three parameters in each role: The slope and intercept of the regression line of suggested prices s on values v (as buyer) or prices p on suggestions s (as seller), and the associated R^2 of each regression. The buyer-regression slope is a measure of information revelation while the seller-regression slope is a measure of information-sensitivity.

We can express all three parameters compactly for both roles by plotting each subject's buyer and seller regression parameters in a two-dimensional "intercept-slope" space, in the figure below, with intercepts on the x-axis and slopes on the y-axis. Since each subject has two sets of parameters (because they played both buyer and seller roles) their parameters are connected by a chord. Red circles represent seller regressions (p on s) and blue circles represent buyer regressions (s on v). The sizes of the circles for each subject role are proportional to the R^2 statistic for the regression. The particular subject whose full suggested price and price regressions are graphed in Figure 3.1 is the one with the filled circles.

To make the graphs more readable, subjects were divided into four groups based on the R^2 statistics for the two regressions, which represent the behavioral predictability of their strategies. The majority of subjects have substantial predictability in both s and p (R^2¿.30 for both regressions). The most common pattern is like that shown in Figure 1, the red (seller) circle has a positive intercept and slope around 1, which indicates the sellers are marking up the suggested price by a fixed constant, while as buyers, the blue circles typically have slopes around .5 and small intercepts. Note that the presence of a significant markup by the seller indicates that these subjects are not trying to be "fair" by suggesting $s = \frac{v}{2}$, since that would imply that they should also price fairly by choosing $p = s$.

The four subjects shown in the graph on the lower right show little information revelation as buyers (the theorized "babblers"), but as sellers, they still seem to mark up suggested prices. For the 5 subjects in the upper and lower left graphs, nothing systematic is happening.[2] Recall that the simple theory prediction, babbling and ignoring the suggested price, predicts that players' parameter estimates will be centered around slopes of zero, i.e. no information revelation or information sensitivity. In addition it predicts that the seller R^2 statistics should be close to one, since they should employ a constant strategy. The theory clearly does not fit well in general; although there are many cases where slopes are low, we see nothing corresponding to the predicted seller strategies, since these all have very low R^2 statistics. Furthermore, there is striking cross-subject variation which might be explained by imaging. Notice that only one subject shows the inverse correlation between v and s predicted by the CH-model or retaliatory seller preferences.

[2] However, as one might expect, the 3 subjects got some of the lowest payoffs.

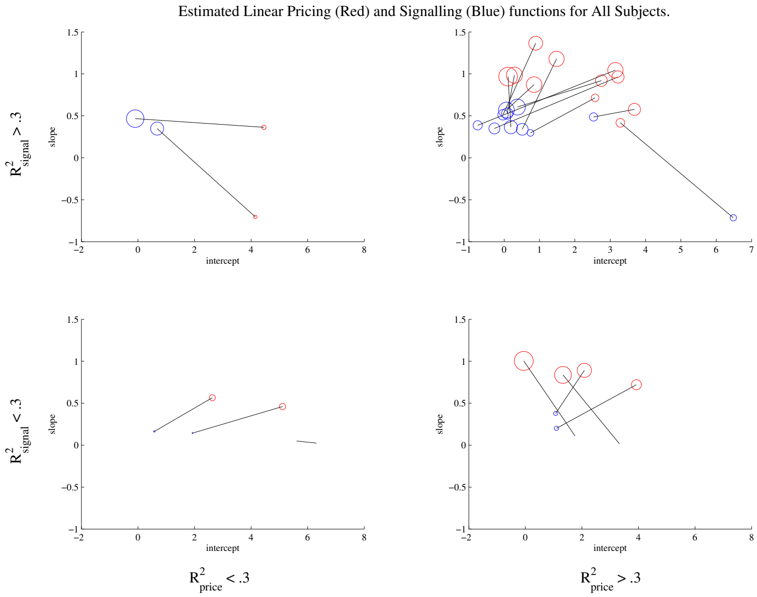


Figure 3.4: Behavioral data for all 20 subjects separated into quadrants according the R^2 statistics for the p on s, R^2_p, and s on v, R^2_s, regressions

The extra information transfer by the subjects resulted in an increased level of efficiency. Equilibrium predicts that $50 - -60\%$ of the feasable trades should take place. We found that 68% of trades took place over all the 20 buyer-seller pairs; in eight of the 20 pairs at least 75% of the feasible trades were made. One pair traded in every round, extracting all of the possible surplus. The efficiency of pairs was weakly correlated with the information revelation and information sensitivity of the buyer/seller pair.

3.5 fMRI results

The central brain imaging dependent variable is the blood oxygen level-dependent (BOLD) signal. This is a coarse measure of brain activity based on how much oxygenated blood flows through a brain area with a short time lag (2–4 seconds). The generic analysis correlates the time series of the BOLD signal in each artificially defined voxel (a 3.4x3.4x4 mm cube) with a matched time series of some sort of regressor. Generally we look for clusters of $k \geq 5$ contiguous voxels which each have BOLD activity correlated with the regressor at a low p-value (typically $p < .001$). There are several types of analyses one can do with BOLD-signal data. We will present only our results derived so far which illustrate how these analyses work and what we have learned and might be learned with further analysis.

There are two interesting behaviors in this paradigm. The first is suspicion in Sellers: how sensitive are sellers to the cheap talk signals that they see? The second is deception in the buyers: How closely do the cheap talk signals they send resemble their private values? For both of these behaviors we can observe both within subject parameters: the distance between the information they receive and their choice, $p - s$ for the sellers and $v - s$ for the buyers; and between subject parameters: the regression slopes and intercept parameters estimated in the previous section. We found areas correlating both of these measures.

3.5.1 Correlates to Suspicion

A between-individual analysis takes some behaviorally derived parameter for each person (or, perhaps each pair) and regresses the brain activity linked to some task, such as the onset of a stimulus, on this parameter. Any such correlation permits a statement of the form "higher activity in area A is linked to the propensity to exhibit behavior B." These analyses are quite difficult because brain activity is measured with a time lag and error, and individual-level behavioral parameters are often measured with error as well.

There are two interesting brain areas whose activity at the time of price choice is correlated with the seller's information sensitivity (the slope in Figure 1) in a between-individual analysis. Activity in the anterior cingulate (ACC) is *decreasing* in slope and activity in the right temporal pole is *increasing* in slope (Figures 4 and 5 show sections and scatterplot).

The cingulate is typically active during cognitive conflict or "executive function." A class of tasks which reliably produce cingulate activity are called "Stroop tasks" (after the pioneering psychologist Colin Stroop). Stroop gave subjects words printed in ink colors which sometimes matched the word and sometimes did not. The task is then to rapidly name the ink color, but not the word. When subjects see the word they often mistakenly say the word rather than name the ink color. The child's game "Simon says" is also a Stroop task; for an American in England, looking to the right side for oncoming traffic (rather than the familiar left side) is a Stroop task too. More generally, a Stroop task requires a decision maker to override a rapid, highly practiced, automatic response. The fact that ACC activity is higher when sellers are less sensitive to the buyer's price suggestions hints that ignoring a suggested price is like overriding an automatic response in a Stroop task.

The temporal pole region, whose activity is positively correlated with seller's information sensitivity, is known to be active in "theory of mind" tasks, in which people must form judgments about what others intend, it is particularly associated with detecting biological motion, or listening

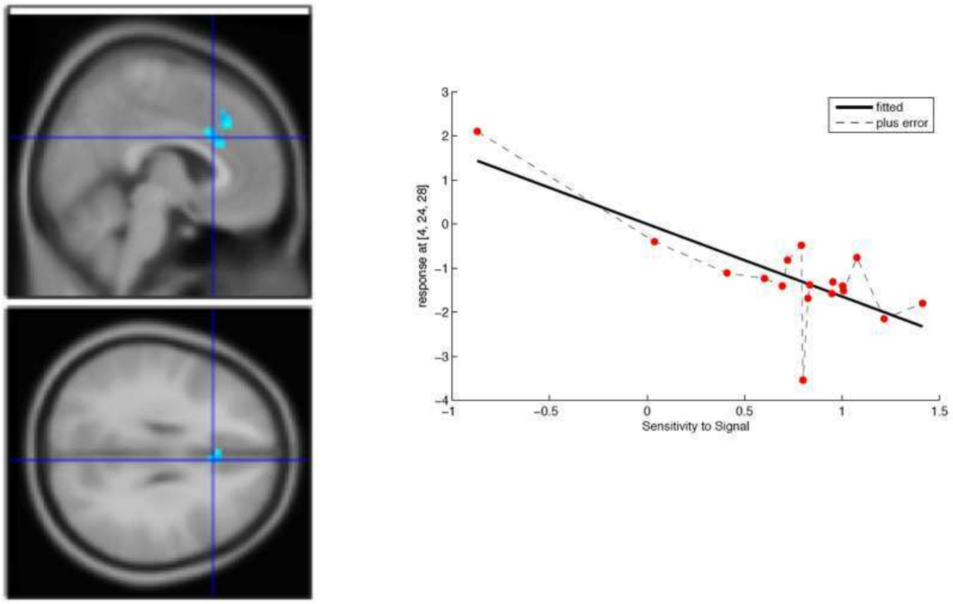


Figure 3.5: Activity in the anterior cingulate cortex at the point of price choice is negatively correlated with a subject's information sensitivity. Activation is shown overlain on the MNI average brain on the left, the plot of relative activation against information sensitivity is shown on the right.

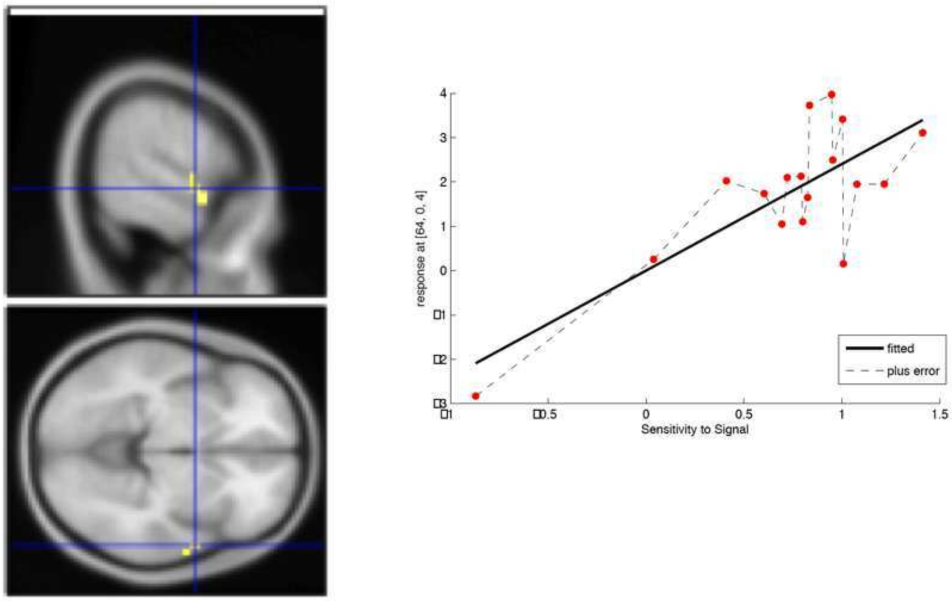


Figure 3.6: Activity in the right temporal pole is positively correlated with subject information sensitivity. Activation is shown on the left, parameters are plotted on the right.

to meaningful speech [5]. The temporal pole and ACC activity together suggest two patterns of behavior different subjects adopt: In one pattern, sellers with high information sensitivity do not perceive a conflict between the suggested price s and the optimal price p (since parametrically, the relation between those two variables is high), so ACC is relatively inactive; instead, the temporal pole is active since these sellers are attributing meaning to these suggested prices. Conversely sellers with low information sensitivity exhibit little temporal pole activity because they are not attributing a lot of meaning to the suggested price. However, because the price they choose and the suggested price are not correlated, the ACC is busy trying to resolve the conflict between the "automatic" response to the suggested price and the "override" of choosing a different price.

We also found interesting activations correlating to the within subject, trial-by-trial parameter, of price market, i.e., the difference $p-s$ between a seller's chosen price and and the cheap talk signal they receive from the buyer. This price "mark up" is correlated with activity in right dorsolateral prefrontal cortex (RDLPFC). This area, like the ACC, is involved in cognitive control and working memory. In fact, as noted above, the RDLPFC is activated when responders in ultimatum games receive low offers compared to high offers [15], and when activity in the region is inhibited using transcranial magnetic stimulation, subjects are far less likely to reject low offers in the ultimatum game [11]. These results from the ultimatum game both supprt the hypothesis that the RDLPFC is a necessary component of a circuit used to inhibit automatic, or prepotent, responses. This, paired with the between subject activation observed in the ACC, suggests that when sellers choose larger markups and ignore the cheap talk signals they receive, they are exerting more cognitive control to override the impulse to use this information as is.

3.5.2 Correlates to Deception

A between-subjects analysis of brain activity correlating to "buyer honesty" as indicated by the slope of the regression of s on v revealed significant negative correlation with the dorsal striatum bilaterally. This is an area that has been implicated in goal-directed behavior in both humans [13] and animals. A recent study by Atallah found that, in rats, this area was crucial to choosing the correct option in an instrumental conditioning task even when learning was intact [1]. When this region of the brain was pharmacologically blocked during a conditioning task, rats were unable to consistently choose a rewarding option. However, when the block was removed, these same rats were able to immediately perform the task correctly. This implies that although they were not choosing the correct target, the rats were actually learning the correct choice and the dorsal striatum is crucial

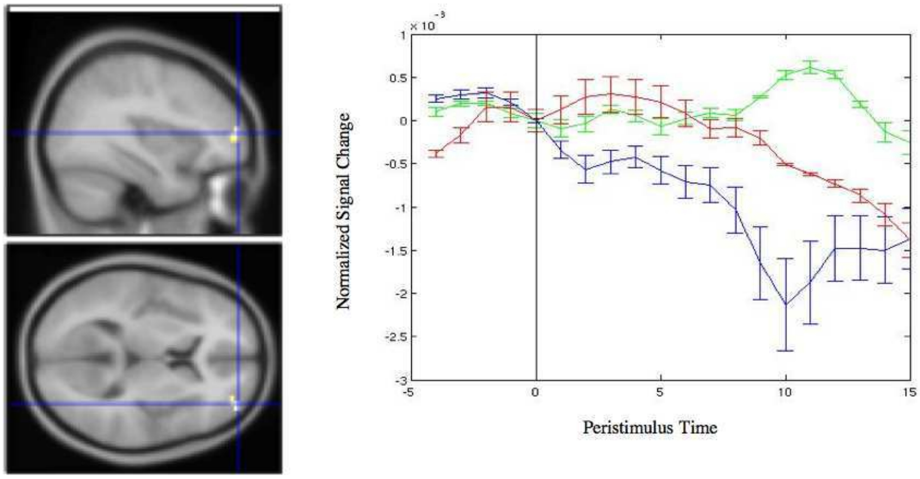


Figure 3.7: The right DLPFC is correlated within subject to the seller's chosen markup and the price decision point. Once again, the cluster is shown on the left on the MNI average brain with peristimulus activation plots for trials divided into the top (red), middle (green), and low (blue) thirds of each subjects range.

to their ability to actually use this information.

The negative correlation of this area with buyer honesty suggests that subjects who revealed less information via cheap talk were somehow more engaged in goal-seeking behavior than those who didn't. Once again this suggests that honesty is a somehow more habitual or automated response requiring less cognitive effort.

All these results support the view that both deception and suspicion are somehow more cognitively effortful that credulity and honesty. In both cases, regions of the brain necessary for effortful behavior, inhibition of a prepotent responses in the RDLPFC for sellers and engaging in goal-directed behavior for buyers, are *negatively* correlated with credulity and honesty (i.e., they are correlated with suspicion and deception).

3.6 Predicting behavior from neural activity and future experiments

One way neural measures might be especially useful for economics is if they can add predictive power to other types of variables. In our study, a candidate prediction of a causal change comes from the well-studied ACC. It is known that scarce cingulate resources can be tied up if subjects have to perform a parallel activity which requires attention or working memory (e.g., remembering a several-digit number). Suppose sellers also have to remember a number while they are deciding

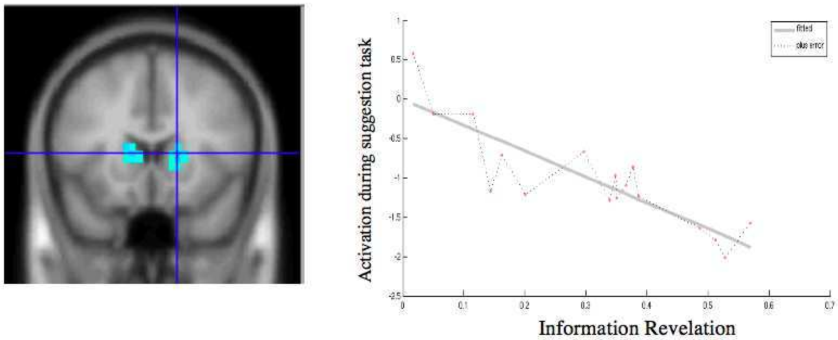


Figure 3.8: Both the left and right dorsal striatum are negatively correlated with subject level "Information Revelation" coefficient (the slope of the regression of s on v), which we use as a proxy for buyer honesty.

what price to choose. If ACC resources are substituted away from the pricing decision and toward number-memory, then the negative correlation between ACC activity and information sensitivity suggests sellers will become *more* information sensitive (i.e., they will respond with prices more closely linked to s). That is, adding a number-memory treatment could change the outcome of the bargaining.

Another candidate prediction comes from the positive correlation between right DLPFC activity and higher price markups (seller prices above buyer-suggested prices). As described above, Knoch et al. (2006) [11] successfully used TMS to disrupt brain activity in right DLPFC and induce players in ultimatum games to accept low offers more frequently. Their interpretation is that the DLPFC is activated by a conflict between the desire to accept money and the desire to enforce social norms by rejecting a low offer. TMS disruption disables this region so behavior reverts to the simpler, more automatic reaction of acceptance. Similarly, one prediction from the DLPFC-markup link we see is that if TMS was used to disrupt activity in the right DLPFC in our task, markups would be reduced.

We don't know if either of these experiments would have the predicted effect, or whether they are worth doing at all. We offer these suggestions, at this exploratory point, mostly as illustrations of the recipe of using the fMRI to identify regions of activity which can then be influenced by other treatments in a way that potentially changes behavior.

3.7 Conclusion

The behavioral data from this experiment show that subjects have a strong tendency toward both information revelation and information sensitivity, even in situations where economic theory predicts there should be none. The neural data linked to these behaviors suggest that rather than a social or emotional explanation (subjects want to be fair), this information transfer may be the result of a far more basic automated response to information. In fact, some activations seen in this study are very close to those seen in the classical information processing experiment, the Stroop task. These sorts of data suggest that we should reconsider what the "default" response of a decision maker might be to any piece of information.

Bibliography

[1] H. Atallah, D. Lopex-Paniagua, J. Rudy, and R. c. O'Reilly. Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. *Nature Neuroscience*, 10:126–131, 2007.

[2] H. Cai and J. Wang. Overcommunication in strategic information transmission games. *Games and Economic Behavior*, 56(1):7–36, 2006.

[3] C. F. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton University Press, Princeton, 2003.

[4] C. F. Camerer, T.-H. Ho, and J. K. Chong. A cognitive hierarchy theory of one-shot games. *Quarterly Journal of Economics*, 119:861–898, 2004.

[5] F. Castelli, C. Frith, F. Happe, and U. Frith. Autism, asperger syndrom and brain mechanisms for the attribution of mental states to animated shapes. *Brain*, 125:1939–1849, 2002.

[6] V. P. Crawford and J. Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.

[7] D. J.-F. de Quervain, U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck, and E. Fehr. The neural basis of altruistic punishment. *Science*, 305:1254–1258, 2004.

[8] H. L. Gallagher, A. I. Jack, A. Roepstorff, and C. D. Frith. Imaging the intentional stance in a competitive game. *NeuroImage*, 16:814–821, 2002.

[9] J. Henrich, R. Bayd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, R. McElreath, M. Alvard, A. Barr, J. Ensminger, N. S. Henrich, K. Hill, F. Gil-White, M. Gurven, F. W. Marlowe, J. Q. Patton, and D. Tracer. "economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28:795–855, 2005.

[10] B. King-Casas, D. Tomlin, C. Anen, C. Camerer, S. Quartz, and R. Montague. Getting to know you: Reputation and trust in a two-person economic exchange. *Science*, 309(5718):78–83, 2005.

[11] D. Knock, A. Pascual-Leone, K. Meyer, V. Treyer, and E. Fehr. Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, 314(5800):829–932, 2006.

[12] R. Myerson and M. A. Satterthwaite. Efficient mechanisms for bilateral trade. *Journal of Economic Theory*, 29(2):265, 1983.

[13] J. O'Doherty, P. Dayan, J. Schults, R. Deichmann, K. Friston, and R. Dolan. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304:452–454, 2004.

[14] R. Radner and A. Schotter. The seeled-bib mechanism: An experimental study. *Journal of Economic Theory*, 48(1):179–220, 1989.

[15] A. G. Sanfey, J. K. Rilling, J. A. Aronson, L. E. Nystrom, and J. D. Cohen. The neural basis of economic decision-making in the ultimatum game. *Science*, 300:1755–1758, 2003.

[16] D. Tomlin, M. A. Kayali, B. King-Casas, C. Anen, C. Camerer, S. Quartz, and R. Montague. Agent-specific responses in the cingulate cortex during economic exchanges. *Science*, 312(5776):1047–1050, 2006.

[17] K. Valley, L. Thompson, R. Gibbons, and M. H. Bazerman. How communication improves efficiency in bargaining games. *Games and Economic Behavior*, 38:127–155, 2002.

Appendix C

Bayes-Nash equilibria of the one-shot game

Claim C.0.1 *The only Bayes-Nash equilibrium of the one-shot game is one where the buyer's suggestion is uninformative.*

Intuitively this follows from the fact that in this game the buyer and seller's incentives are almost exactly opposed. They both prefer that a trade take place, but given that a trade does take place the game is constant sum. To prove this formally we first need to define a few functions.

The probability mass functions for both player's strategy based on the information they receive. For buyers this is the perfect information v, while for sellers this is the noisy and possibly uninformative suggestion s.

$$\begin{aligned} s(x|v) &= Pr\{suggestion = x|v\} \\ p(x|s) &= Pr\{price = x|s\} \\ q(v|x) &= \frac{s(x|v)}{\sum_y s(x|y)} \end{aligned}$$

The utility functions for the buyer and seller respectively are:

$$\begin{aligned} U^B(s,v) &= \sum_{x<v} p(x,s)u_b(v-x) \\ U^S(p,s) &= u_s(p)\sum_{v\geq p} q(v|s) \end{aligned}$$

where u_s, u_b and increasing, weakly concave functions on $\mathbb{R}$ such that $u_b(0) = u_s(0) = 0$.

Let n be the lowest price chosen by the seller with positive probability over all suggestions and choose s^* such that $p(n|s^*) > 0$. Let S_0 be the set of suggestions such that $p(n|s) > 0$. We will show that any buyer with value $v > n$ will always choose a suggested price in S_0 by induction. First note that since n is the lowest price charged given any suggested price, suggestions in S_0 strictly dominate other strategies for buyers with value $n+1$ since any other price yields a payoff of 0.

Assume that all buyers with values in $[n+1, m-1]$ all only choose suggested prices in S_0. So any $s' \notin S_0$ $q(v|s') = 0$ for all values between $n+1$ and $m-1$, which in turn implies that $Pr(v \geq p|s') = Pr(v \geq m)$ for all $v \in [n+1, m-1]$. In other words, given s', m dominates all prices in $[n+1, m-1]$. By assumption $p(x, s') = 0$ for all $x \leq n$ so we have $p(x, s') = 0$ for all $x < m$. Therefore $U^B(s', m) = 0$ for any $s' \notin S_0$ and buyers with value m will always choose $s \in S_0$. Note that this means that all buyers with value $v < n$ must always choose $s \in S_0$ as well since otherwise there would be some $s \notin S_0$ such that $Pr(v < n|s) = 1$ implying that $y(p|s) > 0$ for some $p < n$.

Lemma C.0.1 *In equilibrium sellers choose prices according to some fixed distribution $p^*(x)$ regardless of the suggested price received.*

We show this using an induction argument like the one above. We need to show that $p(x|s_1) = p(x|s_2) = p^*(x)$ for all $s_1, s_2 \in S_0$ that are sent with positive probability. To see this note that it must be true for $x = n$ since if there is any s_1, s_2 such that $p(n|s_1) > p(n|s_2)$ s_1 dominates s_2 for buyers with value $n+1$. Once again, this implies that $s(s_2|n+1) = 0 \Rightarrow p(n+1|s_2) = 0$ and s_1 dominates s_2 for buyers with value $n+2$. Itterating up we see that s_1 must dominate s_2 for all buyers and $s(s_2|v) = 0$ for all v. Once again, using induction we see that $p(x|s_1) \cong p(x|s_2)$ for all s_1 s_2 sent with positive probability. In other words, any suggested price actually sent by the buyer in equilibrium has no effect on the seller's pricing strategy. Let S_1 be the set of suggestions sent in equilibrium.

Let x be some price sucht that $p^*(x) > 0$. This means that x is a best response to $s^* \in S_1$, specifically it must be at least as good as $x+1$, and $x-1$ so

$$\begin{aligned}
u_s(x)\sum_{v\geq x} q(v|s^*) &\geq u_s(x+1)\sum_{v\geq x+1} q(v|s^*) \Rightarrow \\
u_s(x)q(x|s^*) &\geq (u_s(x+1)-u_s(x))\sum_{v\geq x+1} q(v|s^*) \Rightarrow \\
u_s(x)s(s^*|x) &\geq (u_s(x+1)-u_s(x))\sum_{v\geq x+1} s(s^*|v)
\end{aligned}$$

Since $p^*(x)$ is the same for all $s^* \in S_1$ we can sum over all the s in S_1 and get

$$u_s(x) \sum_{s^* \in S_1} s(s^*|x) \geq (u_s(x+1) - u_s(x)) \sum_{v \geq x+1} \sum_{s^* \in S_1} s(s^*|x).$$

We know that buyers will always choose $s^* \in S_1$, so this is simplified to

$$\frac{u_s(x)}{u_s(x+1) - u_s(x)} \geq 10 - x.$$

Since u_2 in increasing and concave the left side of this inequality is strictly increasing while the right side is strictly decreasing. Also notice that if the inequality is *strict* $p^*(x+1) = 0$, so the seller will mix between at most 2 prices x and $x+1$ (this occuring only when the intersection point x^*, such that $10 - x^* = \frac{u_s(x^*)}{u_s(x^*+1) - u_s(x^*)}$ is an integer). This in turn implies that the seller simply chooses x to maximize $(11 - x)u_s(x)$ the expected utility when the seller has no information.