Identity and the Theory of the Firm

Jeffrey V. Butler

This version: March 2017

Abstract

The concept of identity is increasingly entering economists' discourse on a wide range of topics. In this chapter I detail the historical development of the theory of social identity outside of economics, touching on key economically-relevant insights and results. I then describe how identity has been modeled by economists, outlining two key strands of the literature—a preference-based model of identity due to Akerlof and Kranton and a beliefs-based model developed by Bénabou and Tirole. Next, I describe one way identity can be applied to the field of industrial organization in particular: it may help to shed light on fundamental aspects of the classical theory of the firm. In the concluding section, opportunities for future research are highlighted.

1 Introduction

Categorization allows us to make sense of a complex world by organizing and structuring an increasingly bewildering array of stimuli and concepts (Bodenhausen et al., 2012). This categorization compulsion may be evolutionarily adaptive, as for most of human history survival has substantially depended upon the ability to construct a meaningfully predictable world (cf. Bartlett (1932)). Applied to the natural world, taxonomies abound and make basic scientific inquiry possible. Applied to the social world, the question naturally arises: how do I categorize myself? The answer to this fundamental question of social categorization turns out to have important consequences for human behavior and hence for economic theory.

In this chapter, I will provide an overview of what has become known as Social Identity Theory (SIT), first outlining pioneering early work by social psychologists and then detailing how SIT has been incorporated into economics through formal models. In doing so, I hope to: i) highlight how the concept of social identity has enriched economists' understanding of human behavior in general; ii) describe which potentially economically important aspects of social identity are still missing from our economic models; and ultimately iii) detail how SIT can contribute to the field of industrial organization in particular.

2 Classical Social Identity Theory

2.1 Early results: ingroup bias

The social psychological research on identity makes a distinction between *personal identity* and *social* identity as two distinct parts of an individual's self-concept. Personal identity refers to self-knowledge that derives from an individual's unique attributes (Haslam and Ellemers, 2005), i.e., that portion of one's self-concept that differentiates an individual from other *individuals*.[1] Social identity, on the other hand, is a concept that concerns itself with the (social) categories individuals use to describe themselves and others and to place themselves in a particular social context. The groups or categories in question are quite general and abstract, beyond the realm of the small-group face-to-face interactions that have been the purview of, e.g., the formal study of organizational behavior. In particular, they "... need not depend upon the frequency of intermember interaction, systems of role relationships, or interdependent goals ..." (Turner and Tajfel, 1986, p. 15). Instead, a social group or category as defined by SIT is a purely mental, social and emotional construct, a ...

> ... collection of individuals who perceive themselves to be members of the same social category, share some emotional involvement in this common definition of themselves, and achieve some degree of social consensus about the evaluation of their group and of their membership in it (*ibid*).

Notice that this definition requires both self-perception (which category do I belong to?) but also social perception and social reasoning (which category do others place me in?) This ostensibly places the categorization process in the realm of strategic interaction, amenable to game theoretical analysis. As we will see, however, while a shared consensus about categories' relative standings (status or social esteem) is central to social identification processes, a formal treatment of how such a consensus might arise as the equilibrium of a well-defined game has eluded serious inquiry by SIT and economic theory alike.

With this definition in mind, the body of results and conjectures built upon consideration of large-scale social categories, SIT, began as an attempt to understand the age-old puzzle of discrimination. Viewing prejudicial stereotypes as a particular manifestation of social categorization, the question arose: does social categorization *cause* prejudice? Specifically, in their seminal study, Tajfel et al. (1971, p. 151) asked "... can the very act of social categorization ... lead to intergroup behaviour which discriminates against the outgroup and favours the ingroup?" This central question

[1] Hogg et al. (2004) describe it thusly: "Personal identity is a self-construal in terms of idiosyncratic personality attributes that are not shared with other people ('I') or close personal relationships that are tied entirely to the specific other person in the dyadic relationship ('me' and 'you'). Personal identity has little to do with group processes, although group life may well provide a context in which personal identities are formed (e.g., friendships and enmities)."

of causation was posed against the backdrop of an alternative view that had been gaining ground at the time, that it is intergroup conflict of interests and competition for limited resources that *causes* intergroup discrimination (Campbell, 1965; Sherif, 1961). It is worth noting, and expanding upon below, that the question posed was one about sufficiency (*can* categorization cause prejudice) not one of necessity (*must* categorization cause prejudice).

To appreciate the resolution of this early causation question by SIT theorists, it will be helpful to understand an experimental paradigm that yielded overwhelming evidence in support of the notion that social categorization can *cause* intergroup discrimination, the so-called "minimal group paradigm" (MGP). Five criteria were put forth to define an MGP experimental design (Tajfel et al., 1971, pp. 153–154):

1. There should be no face-to-face interaction whatever between the Ss, either in the ingroup or in the outgroup or between the groups.

2. Complete anonymity of group membership should be preserved.

3. There should be no instrumental or rational link between the criteria for intergroup categorization and the nature of ingroup and outgroup responses requested from the subjects.

4. The responses should not represent any utilitarian value to the subject making them.

5. A strategy of responding in terms of intergroup differentiation (i.e., favouring the ingroup and detrimental to the outgroup) should be in competition with a strategy based on other more 'rational' and 'utilitarian' principles, such as obtaining maximum benefit for all. A further step in this direction would be to oppose a strategy of maximum material benefit to the ingroup to one in which the group gains less than it could, but more than the outgroup.

6. Last but not least, the response should be made as important as possible to the Ss. They should consist of real decisions about the distribution of concrete rewards (and/or penalties) to others rather than of some form of evaluation of others.

In the prototypical MGP experiment, participants are divided into two groups based on a trivial distinction meant to convey as little true information about group members as possible. Each participant is then asked to allocate a valued resource, typically money, to *other* participants anonymously. The only knowledge subjects have at their disposal when making their allocation decisions is each of these other participant's group affiliation. Across experimental conditions, in a procedure that would come to feel familiar to experimental economists studying other-regarding preferences

thirty years later, the relative costs of allocating points to each group are varied across treatments in order to isolate participants' underlying motivations.

One celebrated result that arises from many MGP experiments conducted across decades of research[2] is that allocations are indeed discriminatory and that discrimination is motivated by a desire to maximize the relative advantage of one's own ingroup over the outgroup.[3] Because discrimination occurs "... in a situation devoid of the usual trappings of [real-world] ingroup membership and of all the vagaries of interacting with an [real-world] outgroup ..." (Tajfel et al., 1971, p. 172) MGP experiments surprisingly provided an affirmative answer to the discrimination question above. Moreover, because such great pains are taken in MGP experiments to ensure group affiliation conveys little decision-relevant information, the results suggest that social identification affects *preferences* independently of any effects on beliefs about groups, such as effects attributable to statistical discrimination or Beckerian preferences for interacting with certain types of people.

Unfortunately, the notion that simple ingroup bias—treating members of one's own social category better than non-members—is an unavoidable consequence of social categorization overshadowed lessons from subsequent decades of theoretical, experimental and empirical research into social identification processes. Indeed, one could be easily forgiven for identifying the entire contribution of SIT as documenting ingroup bias. In point of fact, while eminently reproducible, simple ingroup bias should be properly viewed as an artefact of particular aspects of the MGP paradigm, most importantly the equal-but different social-groups context and the restriction of decision-makers' choice sets. Much like dictator games being used by experimental economists to document the existence of other-regarding preferences, the MGP served its purpose in showing that individuals' preferences depend on social categorization.

Being so precisely designed, however, the MGP is not capable of providing more nuanced evidence about the myriad ways social identification affects behavior. Indeed, in (even some early) experiments introducing only slight deviations from the criteria defining an MGP design, simple ingroup bias did not obtain. Turner (1978) found that ingroup bias disappears when the allocation decisions include the opportunity to keep some money for oneself, while Mummendey et al. (1992) found that changing the resource being allocated from a good, such as money, to a "bad," such as an

[2] An early summary of existing results is provided by Turner and Tajfel (1986). For a more recent overview see Brown (2000).

[3] Tajfel et al. (1971, p. 173) summarize their findings as follows: "When the Subjects have a choice between acting in terms of maximum utilitarian advantages to all (the Maximum Joint Profit, MJP, strategy) combined with maximum utilitarian advantage to members of their own group (the Maximum Ingroup Profit, MIP, strategy) as against having their group win on points at the sacrifice of both these advantages (the Maximum Difference, MD, in favour of the ingroup strategy), it is the winning that seems more important to them. It is clear from the analysis of the findings that this is a deliberate strategy adopted for their choices, although they are aware of the existence of the alternative strategies."

unpleasant noise, also eliminated simple ingroup bias.[4] Moreover, several studies found substantial *outgroup* bias in near-minimal contexts (Branthwaite and Jones, 1975; Jost et al., 2004).[5]

2.2 SIT theorists re-group: beyond ingroup bias

Tajfel and co-authors, at times, took great pains to avoid describing SIT in terms of ingroup bias. In language that foreshadows the later elaboration of the "uncertainty reduction" hypothesis (discussed below) Tajfel et al. (1971, p. 153) state the point of their seminal work more in terms of behavioral prescriptions or social norms than the necessity of ingroup bias:

> An undifferentiated social environment makes very little sense and provides no guidelines for action. Whenever alternative guidelines for action are lacking, unclear or confusing, and some form of intergroup categorization can be used, it will give order and coherence to the social situation while at the same time enabling the individual to act in a way which has been sanctioned as 'appropriate' in many other situations. This is an aspect of intergroup conduct which is ... present in all intergroup situations.

Tellingly, putting MGP results in this broader context, Turner and Tajfel (1986, p. 17) assert that

> ...in the paradigm of the minimal group experiments, the intergroup discrimination can be conceived as being due not to conflict over monetary gains, but to differentiations based on comparisons made in terms of monetary rewards. Money functioned as a dimension of comparison (the only

[4]Turner produced this lack of ingroup bias in two separate experiments featuring a decisionmaker allocating a valued resource, money or points, between himself or herself and two other participants. These experiments are discussed in Turner (1975): "In fact self-other but not intergroup competition was found; subjects treated outgroupers and ingroupers in the same fashion ... since the number of subjects in the experiment was small and also since instrumental rather than social competition could be used to explain the lack of intergroup effect, a further experiment was run in which self-other non-monetary choices were required ... the prediction was confirmed that ... there would be self-other but not inter-group discrimination."

[5]An important and well documented context which often yields outgroup bias is the context of social stratification. Turner and Tajfel (1986, p. 11) describe "decades of research" demonstrating that minority or subordinate group members "... have frequently tended to derogate the in-group and display positive attitudes toward the dominant [out-]group." One economically relevant manifestation of bias in favor of high status (near-minimal) outgroups might be the phenomenon documented in Ball et al. (2001) where experimental participants were randomly assigned to a "high status" or "low status" group. The authors found that in a market exchange context, prices significantly favored the high status group, being significantly higher when high status sellers faced low status buyers and significantly lower when low status sellers faced high status buyers. This pattern is consistent with low-status group members exhibiting a preference for the other (high-status) group in what was essentially a money allocation decision—how to distribute potential gains from exchange.

> one available within the experimental design), and the data suggest that larger absolute gains that did not establish a difference in favor of the in-group were sacrificed for smaller comparative gains, when the two kinds of gains were made to conflict.

In particular, the notion of preferences for "fairness" which has recently captured economists' attention, factored heavily into the interpretation of early SIT findings (Tajfel et al., 1971, p. 173–174):

> All the choices in the experiments can be conceived as tending to achieve a compromise between F[airness] and other variables; ... with some exceptions in the results of the pilot experiment, all choices hover around distances not too far from the point of maximum fairness. ... In the Introduction we discussed the importance of the generic 'groupness' norm in the determination of the Ss' behaviour. All our results show that another social norm, that of fairness, is also powerful in guiding their choices and that the pattern of data can best be understood as showing a strategy in which a compromise between these two norms is achieved whenever possible.

Summarizing thirty years of SIT research, Brown (2000, p. 763) proclaims: "As should be clear by now SIT is a theory about intergroup differentiation rather than outgroup derogation."

Ushering in a change of focus from merely establishing that preferences are affected by social identification processes to investigating *how* and under what circumstances social identity impacts behavior, Turner and Tajfel (1986) collect lessons learned from early SIT, particularly results based on MGP and near-MGP experiments, and combine them with new conjectures to clarify, expand and reformulate what SIT says about human cognitive processes, motivations, beliefs and behavior. This led to a closely related body of work collectively referred to as "Social Categorization Theory" (SCT) focused more on intragroup social and cognitive processes than on purely intergroup behavior which had been the domain of early SIT. However, SIT and SCT stem from common roots (e.g., Turner was Tajfel's student) and remain tightly linked so that for the purposes of this chapter I will consider SCT as subsumed within SIT rather than referring to the union of SCT and SIT as "the social identity approach," as is common among social psychologists.

That being said, as a point of departure Turner and Tajfel (1986, pp. 15-16) refine their definition of social identities to be

> ... cognitive tools that segment, classify, and order the social environment ... provide a system of orientation for self-reference ... create and define the individual's place in society ... [and] provide their members with an

> identification of themselves in social terms. These identifications are ... relational and comparative: they define the individual as similar to or different from, as "better" or "worse" than, members of other groups. It is in a strictly limited sense, arising from these considerations, that we use the term social identity. It consists ... of those aspects of an individual's self-image that derive from the social categories to which he perceives himself as belonging.

In order to advance SIT, researchers turned their attention from documenting the fundamental nature of social categorization to focus on the categorization process itself asking, for example, how individuals "choose" among myriad possible social categorizations, how this process is coordinated and why individuals choose particular identity categories for themselves. From the definition above, it should be clear that social identification is to be thought of as closely related with and motivated by considerations of social status/hierarchy. The question of how individuals choose among possible social categorizations hinges upon the underlying conscious or subconscious motivations for social identification. Theoretical and empirical research suggests two motivations are primary: self-esteem and uncertainty reduction.

Self-esteem has the longest and most well established history as a motivation underlying social identification processes. This conjecture is summarized in the " self-esteem hypothesis," according to which "... social identity and intergroup behavior are guided by the pursuit of evaluatively positive social identity, through positive intergroup distinctiveness, which, in turn, is motivated by the need for positive self-esteem..." (Hogg and Terry, 2000, p.124). Thus, individuals seek to attain "positive distinctiveness" in their group's relationship to other groups, which feeds back into self-esteem through a process of "depersonalization" whereby the distinction between self and ingroup becomes blurred. SIT researchers have long posited that (social) identification leads individuals to actually perceive themselves in terms of a shared social identity rather than in terms of the idiosyncratic characteristics differentiating them from other individuals (personal identity) (Turner et al., 1987). Recent research using a response-time methodology supports this conjecture and suggests that individuals literally internalize active social identities as part of their self-concept at a subconscious level (Smith and Henry, 1996). Early MGP results documenting a desire to maximize intergroup differences are consistent with the self-esteem hypothesis. Additional, more direct, support for the self-esteem hypothesis stems from a meta-analysis of the relationship between self-esteem and ingroup bias (Aberson et al., 2000). In this meta-analysis, the authors document a strong positive relationship between an individual's level of self-esteem and the degree to which he or she identifies with the ingroup, as well as with measures of ingroup bias.

The second motivation driving social categorization is uncertainty reduction writ large: "... social identity processes are also motivated by a need to reduce subjective uncertainty about one's perceptions, attitudes, feelings, and behaviors and,

ultimately, one's self-concept and place within the social world" (Hogg and Terry, 2000, p.124). Uncertainty is a generally aversive phenomenon associated with unease and fear (Mullin and Hogg, 1999) and reduced subjective well-being (Graham et al., 2010).[6] SIT theorists have noted that the goal of uncertainty reduction can "...rather well be satisfied by the processes of group identification" and, moreover, that despite being conceptually a late-comer to SIT, "... is perhaps a more basic or stronger motivation than the pursuit of positive self-esteem" (Mullin and Hogg, 1999, p. 92). Hogg and Terry (2000, p. 124) assert "... uncertainty reduction, particularly about subjectively important matters that are generally self-conceptually relevant, is a core human motivation [that] ... renders existence meaningful and confers confidence in how to behave and what to expect from the physical and social environment ...."

According to SIT, the uncertainty reduction motivation manifests itself through a combination of depersonalization—cognitively blurring the lines between the self and the (salient) ingroup—and the categorization process itself. In particular, categorization leads people to "... cognitively represent groups in terms of prototypes—fuzzy sets of interrelated attributes that simultaneously capture similarities and structural relationships within groups and differences between the groups, and prescribe group membership-related behavior" (Hogg et al., 2004, p. 253). Prototypes describe and prescribe perceptions, attitudes, feelings, and behaviors for all group members (Hogg and Terry, 2000). Through depersonalization, "... people come to see themselves and other category members less as individuals and more as interchangeable exemplars of the group prototype" (Hornsey, 2008, p. 208). Because prototypes are a shared and consensual feature of categories that provide support and validation for one's self-concept, beliefs and behaviors, it is *prototypes* that ultimately reduce uncertainty (Hogg and Terry, 2000). Consequently, categorizations yielding prototypes that are better suited to reducing uncertainty could be more likely to survive and be shared among particular populations. Summing this up, Hogg and Terry (2000, p. 124) assert that

> ... [since] uncertainty is better reduced by prototypes that are simple, clear, highly focused, and consensual, and that, thus, describe groups that have pronounced entitativity ... [and] are very cohesive ... [they] provide a powerful social identity. Such groups and prototypes will be attractive to individuals who are contextually or more enduringly highly uncertain, or during times of or in situations characterized by great uncertainty.

The tradeoff between these two motivations, self-esteem and uncertainty reduction, may speak to an early controversy in SIT — the repeated finding that low status groups may sometimes acquiesce to their socially ascribed low status and favor

[6]Further highlighting the intimate relationship between uncertainty and fear, psychologists posit a causal relationship running from fear to uncertainty aversion (Loewenstein et al., 2001) and economists have subsequently documented this causal relationship (Guiso et al., 2013).

(high status) outgroups along various dimensions (cf. Spears and Manstead, 1989). Essentially, if challenging existing shared categorizations increases uncertainty at the expense of self-esteem, this tradeoff may sometimes dictate acquiescing to placement in a low status category (Jost and Kramer, 2003). There is evidence that low status groups do not simply give up, however, but rather attempt to enhance self-esteem by changing the dimension of comparison to margins that favor the ingroup, even if conceding that on other dimensions the outgroup is better. In this way, the self-esteem motive may spill over into beliefs-formation: SIT posits that "[after] having defined themselves in terms of that social categorization, individuals seek to achieve or maintain positive self-esteem by positively differentiating their in-group from a comparison out-group on some valued dimension" (Haslam and Ellemers (2005, p. 43) quoted in Ashforth et al. (2008, p. 335)). Thus, intergroup differences may come to be accentuated along particular dimensions and the particular dimension of accentuation may be constrained by shared social assessments about which margins a group is undeniably superior or inferior on with the consequence that intergroup differences may be perceived as more stark than they actually are. Such a pattern would be in line with a more general and widely-documented phenomenon of motivated beliefs (Festinger, 1957; Akerlof and Dickens, 1982).[7]

Before moving on, it is worth noting three implications of this new formulation of SIT that may not be immediately apparent and which warrant further investigation. The first relates to depersonalization. One possible implication is that the more one identifies with the group and, through depersonalization, internalizes the group as self, the more likely one is to act in accordance with the group's beliefs, norms and values, and generally to act in "group-typical ways" (Van Knippenberg, 2000, p. 358). Essentially, group-contingent prescriptions come to serve as internalized personal behavioral or moral norms guiding behavior. (cf. Akerlof and Kranton, 2000, 2010). A second implication is that social identification may interact with beliefs-generalization processes, such as social projection or "false consensus," as introspection becomes more relevant vis-á-vis ingroup members than outgroup members. There is a long-standing lack of consensus among SIT researchers about whether the ingroup or the outgroup, or neither, is generally perceived as more homogeneous (see the discussion in Brown (2000)). The answer is important for considering the economic (game-theoretic) consequences of SIT. Considering how identification interacts with more well-established beliefs biases could shed light on this open question. Finally, notice that depersonalization can reconcile early ingroup bias results with puzzling findings, both in the early research and continuing into the present day, showing that ingroup bias disappears when own-earnings enter into the mix of allocation decisions. Essentially, allocating money to oneself could be equivalent to allocating money to the ingroup if self and ingroup are synonymous.

[7]See also the discussions in Epley and Gilovich (2016) and Bénabou and Tirole (2016).

3 Identity and Economics

3.1 Theoretical Models of identity

3.1.1 Identity as a component of preferences

Identity was introduced into the economic discourse through a series of papers by Nobel laureate George A. Akerlof and Rachel Kranton, beginning with Akerlof and Kranton (2000) and culminating a decade later with a book by the same authors (Akerlof and Kranton, 2010). In this body of research, Akerlof and Kranton, hereafter AK, develop a theoretical framework that distills much of the preceding SIT literature into an economic model of *preferences*. Mirroring the distinction in SIT between personal identity and social identity, AK posit a model in which an individual's overall utility stems from two sources. The first source is standard economic preferences, i.e., idiosyncratic preferences over the consumption of goods and services implied by one's own and all others' actions. The second source of utility concerns (social) identity. In symbols, utility is given by:

$$U_j = U_j(a_j, a_{-j}, I_j) \tag{1}$$

Here, a_j and a_{-j} refer to individual j's actions and the vector of all others' actions, respectively. I_j, which refers to the portion of utility stemming from identity or self-image concerns, in turn, depends on several factors:

$$I_j = I_j(a_j, a_{-j}, c_j, \epsilon_j, P_j) \tag{2}$$

Person j's identity I_j depends, first of all, on j's assigned social category, c_j. The social status of a category is given by the function $I_j(\cdot)$, and a person assigned a category with higher social status may enjoy an enhanced self-image. In this model, categories and the status relationships among them are exogenously given. Identity utility further depends on the extent to which j's characteristics ϵ_j match the ideal of j's assigned category, indicated by the prescriptions P. Finally, identity depends on the extent to which j's own and others' actions correspond to prescribed behavior, also indicated by P. AK refer to increases or decreases in utility that derive from I_j as gains or losses in identity. In particular, AK assume that individuals gain identity by more closely living up to their category's prescriptions. In relation to SIT described above, one can think of P_j as capturing the "prototype" associated with j's category that prescribes ideal behavior and traits.

As a concrete example, to illustrate what parts of SIT this model captures and what portions are left unmodeled, consider the specific instance of this framework developed in Akerlof and Kranton (2002) where AK apply their identity model to the case of American high schools. The standard part of utility consists of a student's future earnings, which is the product of an exogenously-given wage, w, and marketable skills, k_i. Marketable skills are, in turn, the product of costly effort e_i—a choice

variable—and ability n_i, a trait, so that $k_i = e_i n_i$. Ability is distributed randomly uniformly in the high school population: $n_i \sim U[0,1]$. The cost of effort is increasing and convex, $\frac{1}{2}(e_i)^2$. Overall, then, the standard part of utility is $w \cdot k_i - \frac{1}{2}(e_i)^2 = w \cdot n_i e_i - \frac{1}{2}(e_i)^2$. In addition to ability, students have some level of "looks," l_i, where looks are distributed randomly uniformly and independently of ability: $l_i \sim U[0,1]$.

Building upon sociological, ethnographical and social psychological research, AK further posit that in a typical high school, three identity categories are particularly salient—the L(eading crowd), N(erds) and B(urnouts). Belonging to a particular category carries with it a direct (identity) utility of I_c, where research suggests $I_L > I_N > I_B$. Two of these categories have associated with them a prescription about characteristics, P: L-members should be as attractive as possible ($l_i = 1$), while N-members should be as smart as possible ($n_i = 1$). All three categories prescribe an ideal effort level, $e(N) > e(L) > e(B)$. Identity utility depends upon an individual's chosen or ascribed category, c_i, together with how closely the individual matches the category's prescribed characteristics and effort level. A member of the leading crowd, $c_j = L$ with looks l_i and chosen effort level e_i derives a level of identity $I_L - t(1 - l_i) - \frac{1}{2}(e_i - e(L))^2$, where t is a parameter capturing "how difficult it is for students with different ascriptive characteristics to fit in a group." The loss in identity from not matching prescribed behavior, $e(L)$, is quadratic in the distance from chosen behavior and ideal behavior for mathematical convenience.

Overall utility is the weighted average of standard economic utility and identity utility. With $p \in [0,1]$,

$$U_i(e_i; l_i, n_i, L) = p[wk_i - \frac{1}{2}(e_i)^2] + (1-p)[I_L - t(1-l_i) - \frac{1}{2}(e_i - e(L))^2]$$
$$U_i(e_i; l_i, n_i, N) = p[wk_i - \frac{1}{2}(e_i)^2] + (1-p)[I_N - t(1-n_i) - \frac{1}{2}(e_i - e(N))^2]$$
$$U_i(e_i; l_i, n_i, B) = p[wk_i - \frac{1}{2}(e_i)^2] + (1-p)[I_B - \frac{1}{2}(e_i - e(B))^2]$$

Individuals seek to maximize overall utility through their choice of identity and effort level. Consider the (realistic) assumption that in American high schools identity concerns are paramount, i.e., $p \approx 0$. Intuitively, the choice of identity has two considerations. First of all, each identity carries with it a fixed, direct, utility consequence which can be thought of as reflecting a social consensus about which categories carry more social esteem. On the other hand, identity membership imposes prescriptions through which an individual may gain or lose identity utility depending on the comparison between chosen behavior and prescribed behavior or one's own traits and prescribed "ideal" traits. For example, if an individual's "looks" are far from the looks prescribed by the leading crowd category ($l_i << 1$), then it is possible that the direct utility gain from being in the leading crowd (I_L) is more than swamped by the (indirect) utility loss from falling short of the category-ideal level of looks ($-t(1-l_i)$).

Formally, AK show that in the standard economic case without identity ($p = 1$), the average effort and skill acquisition are completely determined by the market wage and equal to $\frac{w}{2}$ and $\frac{w}{3}$, respectively. In stark contrast, when identity is the only motive ($p = 0$), both effort and skill acquisition are completely determined by the identity parameters $I_c, e(c)$ and t. In particular, when $I_L > I_N$ there will be some high-ability "high-looks" students who choose to identify with the leading crowd. Because the effort level prescribed by the leading crowd category is lower than the prescribed level of effort for nerds ($e(N) < e(L)$), these high ability handsome students choose to acquire a lower level of skills than their nerd-identifying classmates of similar ability. Another important margin in this specification is the ease with which students can "fit in", t. When t is high, for a fixed level of ability or looks, being a non-burnout is strictly more costly than identifying with either of the other two groups *ceteris paribus* than if t were low. Consequently, students are more likely to identify as burnouts and, because $e(B) < e(L) < e(N)$, skill acquisition is lower on average. Thus, AK highlight both societal and individual consequences to identity. Differences in social esteem among identity categories may cause individuals to invest less in behaviors that are given less prominence in category-specific prescriptions. When categories are particularly socially esteemed, their prescriptions may exert a strong enough influence to affect societal welfare—here, by potentially reducing the average level of human capital acquisition.

In summary, AK construct a model of individual behavior in which identity enters directly into the overall utility function, overall utility essentially being a weighted average of two distinct components: standard economic preferences and identity-based preferences for, e.g., living up to identity-contingent ideals. In this sense, AK's model is one in which identity enters directly into preferences. Since identity itself depends on the social contextual factors, this creates a model of preferences in which context matters. Their model can capture important features of SIT. The self-esteem hypothesis can be captured by positing an order with respect to the levels of direct identity utility associated with each possible identity category. The model posited by AK fulfills the (intra-individual) uncertainty reduction hypothesis through prescriptions, P, which represent category-specific ideal attributes and behaviors. Moreover, if categorizations come to be common knowledge, i.e., shared social constructs, inter-individual uncertainty is also reduced. By knowing the portion of a counterparty's overall utility defined by his identity, an individual should be better able to predict that counterparty's behavior. This could be termed a strategic-uncertainty reduction byproduct which, while obviously important, does not feature prominently in SIT. What is lacking from AK's framework, and from SIT in general. is a formal consideration of how categories, and the status relationships between them, are formed.

3.1.2 Identity as a signalling problem

The primary alternative in economics to conceptualizing identity as a purely preference-based affair comes from another Nobel laureate, Jean Tirole, together with frequent co-author Roland Bénabou (Bénabou and Tirole, 2006a,b). Their theory, hereafter referred to as the BT framework or simply BT, "... explicitly treats identity ... as beliefs about one's deep preferences or 'values' and emphasizes the self-inference process—defining oneself by one's actions—through which it operates" (Bénabou and Tirole, 2006a, p. 1). The BT framework captures both the self-esteem motive and the uncertainty reduction motive. The self-esteem motive is clearly present in their assertion that "... self image (as, e.g., a caring, honorable, smart, or hard-working person) has consumption value ... precisely because certain self-views are more pleasant ... to have than others ... " and that " ... people invest substantial resources in trying to achieve, maintain and defend these beliefs" (*ibid*). The uncertainty reduction motive is echoed in BT's statement that "... a strong sense of self may provide clear priorities and directions that help mobilize energy and make better decisions"(*ibid*).

The BT framework has several components. The first component is an (exogenously given) set of "types" as in a standard incomplete information game setting. Types here specify some preference parameter—the "deep preferences" or values mentioned above. Because types are linked to preferences, (observed) actions can be informative about an individual's (unobserved) type. Consequently, others' beliefs about one's own type may be affected by one's actions whenever these are observed. The second component of BT's framework is an individual's "social image" or reputation: others' beliefs about that individual's type conditional on all observed actions, where beliefs are updated according to Bayes' rule. The third key component of the BT framework is the assumption that, unlike traditional treatments of reputation, here social image enters directly into an individual's overall utility function. That is to say, individuals have direct preferences about how they are perceived by others, whereas in a classical reputation model such beliefs can only have instrumental value. The last component of BT, the component that permits an identity interpretation of their framework, is that individuals themselves are uninformed about their own types even though types are used in determining actions. This last assumption allows an individual's own actions to serve as a signal to herself about her values and deep preferences, i.e., about her identity.

The most accessible specification of their model is presented in Bénabou and Tirole (2006b), where the authors consider social image motivations for pro-social behavior, such as contributing to a public good, and describe how their analysis can be re-interpreted through the lens of identity (Bénabou and Tirole, 2006b, p. 1657):

> When making a decision affecting others' welfare, an individual will often engage in a self-assessment: "How important is it for me to contribute to the public good? How much do I care about money? What are my real values?" Later on, however, this information may no longer be perfectly

> "accessible" in memory—in fact, there will often be strong incentives to recall it in a self-serving way. Actions, by contrast, are much easier to remember than their underlying motives, making it rational to define oneself partly through one's past choices: "I am the kind of person who behaves in this way."

Here, BT explicitly decompose overall utility into two parts, as do AK above and as does SIT more generally. BT's decomposition is different from both of these previous theories, however. The first part of overall utility is labeled "intrinsic" or direct preferences which, unlike AK, may directly incorporate non-pecuniary motivations such as altruism or distributional social preferences. The second part of overall utility is preferences over reputation. It is this second part of utility that can be thought of as corresponding to an identity component of preferences.

Specifically, denote by $a \in A \subseteq \mathbb{R}$ the individual's action, such as his or her level of contribution to a public good. Denote by ν_a and ν_y the individual's "intrinsic valuations" for the action a and for money, respectively. These are the deep preferences or values mentioned above. Denote by y the monetary incentives associated with choosing the action a. The first part of an individual's overall utility—the intrinsic or direct benefit of choosing action a—can be written $(\nu_a + \nu_y y)a - C(a)$, where $C(a)$ is the direct cost (e.g., disutility of effort) of a. An individuals' type is given by $(\nu_a, \nu_y) \in \mathbb{R}^2$. Types are independently distributed according to a continuous joint distribution function with finite mean. An individual's type is more-than-private information—unknown to others as well as to the individual himself/herself—even though type features directly into the individual's optimization problem. It is this last assumption that allows action to be informative of type to the individual herself.

The second part of overall utility is the beliefs component, alternatively labeled reputation or social image. Here, $E[\nu_a|a, y]$ is the ex-post belief about the individual's intrinsic value for pro-sociality (e.g., altruism) obtained through updating according Bayes' rule when observing a being chosen in a decision situation with material incentive rate y. The ex-post belief about the individual's intrinsic value for money, $E[\nu_y|a, y]$, is defined similarly. Assume that individuals prefer to appear pro-social and to not appear greedy in the sense of valuing money highly. These assumptions can be incorporated by positing a value for reputation given by $R(a, y) = \gamma_a E[\nu_a|a, y] - \gamma_y E[\nu_y|a, y]$, with $\gamma_a \geq 0, \gamma_y \geq 0$ being weights capturing how bad appearing greedy, or how pleasant appearing pro-social, feels. Finally, suppose for simplicity, the second component enters overall utility with weight $0 \leq x \leq 1$ which captures the decision weight given to reputation relative to the intrinsic net costs associated with a. The individual's problem is to maximize overall utility given by:

$$\max_a \{(\nu_a + \nu_y y)a - C(a) + x(\gamma_a E[\nu_a|a, y] - \gamma_y E[\nu_y|a, y])\}$$

This maximization problem yields a first order condition equating the marginal (intrinsic and image) benefits of a to the marginal intrinsic and image costs.

$$C'(a) + x\gamma_y \frac{\partial E[\nu_y|a,y]}{\partial a} = \nu_a + \nu_y y + x\gamma_a \frac{\partial E[\nu_a|a,y]}{\partial a}$$

As a further simplification, define $\mu_a = x\gamma_a$ and $\mu_y = x\gamma_y$ and assume they are common knowledge in the population, including the individual himself. One can interpret these as the weight given to appearing intrinsically pro-social or money-loving, respectively. The first-order condition can be re-written:

$$C'(a) = \nu_a + \nu_y y + \mu_a \frac{\partial E[\nu_a|a,y]}{\partial a} - \mu_y \frac{\partial E[\nu_y|a,y]}{\partial a}$$

From this last equation it should be clear that "... observing someone's choice of a reveals the sum of his three motivations to contribute (at the margin): intrinsic, extrinsic, and reputational" (Bénabou and Tirole, 2006b, p. 1658). Because all three of these motivations jointly determine action, however, inferring the level of either of the intrinsic parameters, ν_a or ν_y, from actions is a potentially complicated signal extraction problem involving differential equations. To get a flavor for the solution to the optimization problem, put aside the social image component of identity for the moment and to make things even simpler, suppose there are no material incentives for the pro-social act ($y = 0$). Behavior is then governed by the relationship between $C'(a)$ and ν_a. For any a, only individuals with types $\nu_a \geq C'(a)$ choose $a' \geq a$ so that choice is very informative about type. Consequently, observing a high level of a should increase ex-post beliefs about ν_a.[8]

This last observation suggests how the self-esteem motive for identity is captured by BT's framework. If society and the individual reach consensus somehow on the idea that types with higher intrinsic pro-sociality (ν_a) should be accorded more social esteem, then an individual obtains an increase in (identity) utility from actions which signal a higher ν_a, like choosing a higher level of a in our simplified example above. The uncertainty reduction motivation is also achieved through a consensus on what types are generally assigned higher social status. The motivation to behave in a way that increases ex-post beliefs about being a "desirable" type together with the idea that the beliefs-updating process is well understood (e.g., Bayesian) given a full description of the decision situation acts as a *de facto* set of behavioral prescriptions. Notice that, as with AK, ultimately the set of social identities (here, types) together with the status relationships among these identities is exogenous to the model. Once these factors are agreed upon, however, identification processes play a potentially substantial role in determining behavior.

[8]When $y > 0$, however, there is a tradeoff. Some types that would not have chosen a high level of a without material incentives y because ν_a was not sufficiently large will now choose a high a level because of material incentives. Thus, the signal extraction problem becomes more difficult—financial incentives inject noise into the problem, so that ex-post beliefs are less responsive to the level of a. This is BTs explanation for "crowding out."

3.2 Experimental and empirical evidence

Shortly after economists began theorizing about social identity, experimental economists began testing the basic tenets of this theory. As with SIT in social psychology, the first tests focused on how fundamental social identification processes were. They asked whether near-minimal identity categorizations were a fundamental enough determinant of preferences to affect behavior. As did SIT, many of these studies documented a resulting simple ingroup bias (Chen and Li, 2009; Eckel and Grossman, 2005; Charness et al., 2007; Guala et al., 2013). Butler et al. (2013) show that, consistent with depersonalization, simple ingroup bias may result from internalizing ingroup members' preferences (empathy) rather than their earnings *per se*, which had been the favored explanation for these early results (cf. Chen and Li (2009)).

Experimental economists have also asked what accounts for (simple) ingroup bias: innate preferences or group-contingent variation in norm-concern? Harris et al. (2010) show that ingroup bias increases when there is opportunity for third-party punishment, suggesting that ingroup bias is a social norm in the context of equal-but-different groups. This interpretation is strengthened by two other papers: Hertel and Kerr (2001) find that priming specific norms that imply ingroup bias (e.g., group loyalty) increases ingroup bias and the expectation of others' ingroup bias, while Jetten et al. (1997) find the effect of norm-priming on ingroup bias is stronger for those who identify more strongly with their group. Finally, it is worth noting that several of these early studies by economists find no effect of equal-but-different social identities on behavior at all (see, e.g., Güth et al. (2008) as well as many of the treatments in Charness et al. (2007) and Eckel and Grossman (2005)).

Mirroring the development of SIT in social psychology, a second wave of interest in identity has begun to investigate *how* identity changes preferences and, in particular, focuses on patterns not consistent with simple ingroup bias. Hargreaves-Heap and Zizzo (2009) show that individuals intrinsically value their ingroup affiliation, while Kranton et al. (2016) document substantial individual heterogeneity in group-contingent behavior. McLeish and Oxoby (2007) find that individuals are more likely to levy costly punishment on ingroup members for breaches of normative behavior than on outgroup members, a pattern consistent with "parochial altruism"'(Bernhard et al., 2006) or, more classically, "amoral familism" (Edward, 1958). Butler (2014) finds a similar result in a similar equal-but-different-group setting—individuals are more likely to positively (negatively) reciprocate pro-social (anti-social) behavior from ingroup members than from outgroup members. However, when a status difference is imposed between groups, these patterns change qualitatively. High (low) status group members are more (less) reciprocal in general. Taken together, this suggests that individuals expect better behavior from ingroup members when groups are equal-but-different, perhaps because of motivated beliefs (e.g., social projection) coupled with identification processes, and so view ingroup transgressions as more serious. The expectations explanation is made more plausible by a long tradition of expecting

more from groups with higher social esteem, a phenomenon generally referred to as *noblesse oblige*. Moreover, results from a handful of other studies inducing uninformative status differences in the lab are also generally consistent with higher status groups behaving more in line with normative prescriptions (Ball and Eckel, 1998; Kumru and Vesterlund, 2008; Willer, 2009). The fullest expression of this second-wave interest is perhaps Chang et al. (2015), where the authors take the AK model literally and elicit group-contingent prescriptions to investigate whether knowing these prescriptions can improve economists' predictions of individual behavior. Cohn et al. (2014) examine the lasting impact of a particular social identity and show that banking industry employees behave less trustworthily, i.e., are less honest, when primed with their banking industry identity.

It should be clear that most of the existing experimental economics literature has adopted an AK-style framework for investigating identity. The beliefs-based approach due to BT has proven more difficult to test and perhaps consequently the effects of identity on beliefs has garnered less attention. This is an important oversight since the SIT literature suggests that identity may affect beliefs in predictable ways and the effect of identity on behavior through beliefs processes may be even more enduring than identity's effects on preferences. One such pattern is the "contrast effect" mentioned above whereby a desire for "positive distinctiveness" leads individuals to exaggerate intergroup differences on valued dimensions. Another example of how social identity may affect behavior indirectly through belief-formation processes relies on guilt or disappointment aversion. If behavior depends on group-contingent prescriptions, an individual's social identity may affect others' expectations about that individual's behavior. Several studies have shown that many individuals have an aversion to disappointing others by falling short of others' expectations (Butler et al., 2016b; Charness and Dufwenberg, 2006). By affecting others' expectations, therefore, social identity may ultimately affect behavior by determining the threshold defining disappointment.

Finally, an important open question is where social identities come from. Who defines the set of categories to which individuals may belong? Here it is important to realize that prototypes need not be information-based, separating them from the closely-related concept of stereotypes. To fulfill their goal of uncertainty reduction, they need only be commonly-held. Any coordinating device, for instance, might suffice. A fruitful avenue of research to pursue may be to conceptualize identity categories as an equilibrium phenomenon in a suitably defined game in which they serve the role of pure coordination devices. It seems likely that pervasive stories, narratives and mythologies are an important source of feasible prototypes and categorizations which may also clearly serve a (cultural) coordination role. Apropos of this, and as a segue into our consideration of Industrial Organization, it is worth noting that the potential for identity to enhance coordination has been demonstrated in a general (laboratory) game (Chen and Chen, 2011) as well as in the specific context of a (laboratory) game framed to induce a within-firm coordination context (Butler et al.,

2016a).

4 Identity and Industrial Organization

A major focus of industrial organization has been the "theory of the firm," which concerns itself with questions related to why some transactions take place within a firm rather than at arm's length through an impersonal market transaction. Historically, this branch of IO has asked what factors limit or shape the size and scope of firms. Understanding the advantages of production within firms has been a puzzle from the beginning (Coase, 1937, p. 390):

> The price mechanism (considered purely from the side of the direction of resources) might be superseded if the relationship which replaced it was desired for its own sake."

While Coase goes on to dismiss this justification for the existence of firms as fanciful, identity has the potential to speak directly to precisely this motivation. The point Coase raises fundamentally concerns employee preferences which, as we have seen, may be malleable by outside actors—including the firm. Most Americans spend a majority of their waking hours under the influence of their employers in some way, either by being physically present in the firm or mentally present while working away from the firm's grounds. Consequently, the firm may be a particularly plausible candidate for "identity entrepreneur," an actor which can create social categories for employees to adopt, define the status relationships among these categories or, most intriguingly, affect the prototypes associated with each identity category. Through these channels, the social identification processes we have discussed so far may enhance the efficiency of within-firm production relative to arm's-length production in several ways.

4.0.1 Reducing efficiency losses due to incomplete contracts

The most obvious potential channel is that, through depersonalization, an employee who adopts a firm-created social identity will tend to internalize group incentives, such as monetary rewards. One simple way to model this internalization is by assuming an individual is more likely to empathize with fellow group members. Denoting by $\mathbb{I}_{[I_i=I_j]}$ an indicator function that takes the value one if i and j share the same social category and zero otherwise:

$$U_i(a_i, a_j, I_i, I_j) = U_i(a_i, a_j) + \mathbb{I}_{[I_i=I_j]} \times \alpha U_j(a_i, a_j) \tag{3}$$

In words, Equation 3 says that i's overall utility places some weight, α, on j's preferences only if j is a fellow group member.

As our first example consider a standard principal/agent situation and further suppose that the principal and agent share the same identity category (e.g., they are both members of a firm that has succeeded in instilling a salient firm identity). In order to avoid infinite regress, assume for simplicity that the principal is a purely selfish money-maximizer while the agent cares about identity and internalizes the principal's preferences. A reduced form model of this situation would be functionally equivalent to a situation in which the agent has distributional social preferences (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002)), i.e., the agent puts some weight in his utility function on the principal's earnings. Suppose the agent's distributional preferences take the form of "inequality aversion" as in (Fehr and Schmidt, 1999), so that $U_i(x_i, x_j) = x_i - \beta_i|x_i - x_j|$. In words, individual i gains utility from higher own earnings but loses utility when earnings are unequal, with the tradeoff between these two motives parameterized by the weight β_i. This case has been analyzed by Kőszegi (2014) in his consideration of how a wide range of results from behavioral economics may be incorporated into contract theory.[9] In the discussion that follows, I rely heavily on Kőszegi (2014).

Consider a fixed-wage (incomplete) contract environment. In this type of contract, the P(rincipal) offers the A(gent) a fixed wage ($w \geq 0$) and the agent chooses a level effort, $e \geq 0$, entailing a cost $c(e) > 0$. Assume effort and production are one-to-one related, so that the effort level is also the material production in dollar terms. P's material payoff is the value of production minus the wage bill, $x_P = e - w$, while A's own earnings are given by $x_A = w - c(e)$.

If the firm must hire As on the spot market and offer a fixed-wage contract, we can think of this as a one-shot interaction. In this case, if all As have purely selfish preferences, i.e., they care only about their own earnings, there is a unique subgame perfect outcome of this (one-shot) contracting situation: A exerts no effort and, knowing this, P offers $w = 0$. Contrast this with the joint-earnings maximizing outcome which features the marginal cost of effort being equal to its marginal production, i.e., $c'(e) = 1$.

Next, consider the case where A is a firm-member who partially internalizes P's earnings. In particular, as suggested above, suppose that we can model A as being averse to inequality. What will the optimal contract look like in this case? We can start from A's effort choice conditional on w. For any degree of inequality aversion, $\beta_A > 0$, A will never choose an effort level so high that $e - w > w - c(e)$: If s/he did, reducing effort would increase own earnings and reduce inequality, both of which increase A's utility. So, we can restrict attention to effort choices yielding $e - w \leq w - c(e)$. Now, assume that A cares at least a moderate amount about inequality so that $\beta_A > \frac{1}{2}$. It can be shown that $e - w < w - c(e)$ is never optimally

[9]While Kőszegi (2014) considers a case where both the principal and the agent have inequality averse distributional social preferences, in the process of solving for the optimal contract it is shown that the principal can be treated as a purely selfish own-payoff maximizer.

chosen by A.[10] Consequently, whenever A is able, s/he always chooses an effort level that eliminates inequality: $w - c(e) = e - w$. Re-arranging this expression yields $w = \frac{e+c(e)}{2}$. Knowing this and solving backwards, P maximizes own earnings by solving the following program:

$$\begin{aligned} &\max e - w \\ &\text{s.t. } w = \frac{e + c(e)}{2} \end{aligned}$$

The solution to this problem entails P maximizing $\frac{e-c(e)}{2}$ so that in the optimal contract, $c'(e) = 1$. Consequently, in stark contrast to the spot market setting, even if interactions are still one-shot, A internalizing P's earnings, perhaps because of a shared firm identity, allows P to implement the first-best surplus-maximizing level of effort with the optimal contract. Notice that this would be the case whether or not P cares about A's earnings directly in an inequality-aversion sense. In summary, by reducing inefficiencies associated with contract incompleteness such as moral hazard, even in a one-shot setting, the ability of firms to generate shared social identities may confer an advantage to within-firm production over production achieved through arms-length market transaction.

4.0.2 Reducing efficiency losses due to asymmetric information

Next consider the case where the optimal action may depend on information about the state of the world, but where the most-informed employees are not necessarily those who must make decisions. This could be because of job duties — employees on the plant floor may, by simple proximity and frequency of interaction, be better informed about the state of plant equipment at any point in time. It could also be about expertise. The team of economists at Amazon may have a better idea about where to invest resources to improve click-through rates than Jeff Bezos, even if Bezos were ultimately in charge of this resource allocation decision. A source of inefficiency may occur when preferences of the information-holders and the decision-makers are not aligned. We can think of this situation involving strategic information transmission. In their seminal work, Crawford and Sobel (1982) show that the maximum precision of information that can be transmitted in equilibrium of such games depends on how misaligned preferences are.

[10] There are multiple ways to see this. In the most straightforward way, first, assume that $c'(e) \leq 1$, as will turn out to be the case in the optimal contract. Write down A's utility conditional on being weakly ahead in earnings: $U_A(e) = x_A - \beta_A(x_A - x_P) = w - c(e) - \beta_A(w - c(e) - (e - w)) = w(1 - 2\beta_A) + \beta_A e - (1 - \beta_A)c(e)$. Taking the derivative with respect to e yields $\beta_A - (1 - \beta_A)c'(e)$. Since $\beta_A > \frac{1}{2}$ and $c'(e) \leq 1$, this derivative is always positive so that utility can always be increased by increasing e.

Consider their motivating example: the state of the world is characterized by a random variable uniformly distributed on $[0, 1]$. The S(ender) privately learns the state of the world, m, and sends a costless message to the R(eceiver) who has no direct information about m. R then takes an action, $y \in \mathbb{R}$. Preferences are given by $U_S = -(y-(m+b))^2$ and $U_R = (-y-m)^2$. In this formulation, S and R have different preferences over R's action conditional on the true state of the world, which only S knows. The scalar b measures the extent of preference conflict, since conditional on m, R's utility is maximized by setting $y = m$ while S's utility is maximized at $y = m+b$. Equilibrium consists of a signalling strategy for S and a decision rule incorporating S's signal for R. The primary result of the paper is that, in equilibrium, S's signals constitute a partitioning of the state space and that the finest partition possible in equilibrium is finer when $|b|$ is smaller. Preference conflict therefore limits the informativeness of S's signal achievable in equilibrium. This particular informational friction obviously represents a source of firm inefficiency. The question is, then, whether a firm can reduce this inefficiency by creating a shared social identity among S and R relative to the alternative of market-based transactions at arm's length.

It is easy to see that the answer is "yes." Consider again the case where S and R are employees of the same firm and share a common social identity. Suppose sharing an identity causes S to internalize R's preferences. Assume preference internalization, as above, takes the form of placing some weight, $(1-\alpha) \in (0, 1)$, on other identity-members' preferences. Suppose for simplicity that S internalizes R's preferences, $U_S = \alpha[-(y-(m+b))^2] + (1-\alpha)[(-y-m)^2]$, but that but R does not internalize S's preferences. Since U_S is the average of two quadratic functions, it is also a quadratic function and, what is more, has a bliss point at $y = m + \alpha b$. In this reformulation, there is strictly less preference conflict between S and R than without preference internalization: $|\alpha b| < |b|$. Consequently, the most informative equilibrium with a common firm social identity is more informative than would be possible without a shared identity.

A further point can also be made with respect to informational frictions. The type of Sender/Receiver games just outlined are referred to as cheap talk games. In cheap talk games S knows more than R about what R's surplus-maximizing action is and can costlessly send some message to R. When S's set of possible messages is sufficiently rich, Crawford and Sobel (1982) show that when preferences are not too misaligned then there are equilibria in which S's messages convey some information. In simpler settings, however, for instance where S's set of possible messages are severely constrained, S's messages are totally uninformative in every equilibrium whenever there is *any* degree of preference conflict. As a consequence, mere preference internalization cannot improve firm efficiency due to asymmetric information. However, there may be identity-based firm-level policies that can attack even this source of inefficiency.

Consider the cheap talk game investigated in Butler (2014). The game consists of a S(ender) and a R(eceiver). S can be thought of as the seller of a good, perhaps expensive machinery, about which he has private information regarding quality. S

knows whether she has a high-quality, profit-enhancing, machine for sale. R can be thought of as a firm who would like to buy high-quality machinery, as this enhances output, but would like to avoid buying low-quality machinery as this just takes up floor space without contributing to output:

$$U_R(\text{buy } |\text{high quality}) > U_R(\text{walk}|\text{high quality})$$
$$U_R(\text{walk}|\text{low quality}) > U_R(\text{buy } |\text{low quality})$$

For his part, S prefers that R buy machinery irrespective of quality, perhaps because of a seller's commission:

$$U_S(\text{buy}|\text{high quality}) > U_S(\text{walk}|\text{high quality})$$
$$U_S(\text{buy}|\text{low quality}) > U_S(\text{walk}|\text{low quality})$$

Thus, there is some preference misalignment. In terms of messages, S is constrained. She can send one of two messages: "high quality" or "low quality." The game proceeds sequentially: R observes S's message and decides between "buy" or "walk." It can be easily shown that in this game with purely selfish preferences S's messages never convey information, i.e., they are independent of the true quality of the machinery.[11] Moreover, mere preference internalization may leave some conflict in preferences and, consequently, may not be helpful in expanding the set of equilibria beyond uninformative "babbling" equilibria.

However, suppose a firm could inculcate not only a shared identity but also particular values. By shaping the "prototype" associated with firm-employee identity, this could be feasible. In particular, suppose the firm instills a value for honesty so that lying entails a loss in identity that takes the form of a psychic cost $c > 0$. Strikingly, no matter how small the psychic cost of lying is, this intervention rules out the possibility of the babbling equilibrium. Intuitively, in any babbling equilibrium S's messages convey no information so that they have no impact on R's decision to buy or walk. Without loss of generality, suppose S knows he has a high quality machine for sale. Since R's decision is the same irrespective of S's message, the expected benefit of the message "high quality" is the same as the expected benefit of the only other possible message "low quality." But, the message "low quality" entails a strictly higher cost than the message "high quality" since the former is dishonest and engenders the psychic cost c while the latter message is truthful. Since the two messages have the same benefit but strictly different costs, S cannot be indifferent between which

[11] A standard argument proceeds as follows. Think of S as having two types: "high (quality)" and "low (quality)." Since both types strictly prefer that R buys rather than walk away, if there were any messaging strategy for a high-type S that would increase R's likelihood of buying, then a low-type S would optimally mimic this strategy. Hence, the only possible equilibrium is one in which high and low types are indistinguishable, which is an uninformative "babbling" equilibrium.

rules out "pooling" as an equilibrium strategy for S, a necessary component of any babbling equilibrium.

Notice that many successful organizations, including universities, have honesty as a key tenet of their (corporate) values or mission statement. Harvard University's "Statement of Values," for instance, includes the line "Honesty and integrity in all dealings" Summers (2002). Of course, it could be that honest people form successful organizations, not the other way around. On causality, Butler (2014) experimentally shows that creating a social identity that ostensibly instills a value for honesty makes it more likely that messages in cheap talk games are informative and, moreover, that this information is utilized by message-receivers, both of which are necessary to improve firm decision efficiency.

4.0.3 Affecting preferences through status allocation

The last possibility I will outline is the most speculative. Firm organization typically imposes a (*de jure* or *de facto*) hierarchy which defines the relative status among different sub-organizations within the firm. It has been a puzzle to economists that often the same role within a firm may pay vastly different wages depending on its place within the firm hierarchy. The archetypal example would be a receptionist for the office of the company CEO as opposed to a receptionist for a lower-level manager. While there are stories that could be told which may justify a wage differential, ultimately the economic situation is that individuals who perform identical tasks earn vastly different wages.

Identity may provide one way to understand this common phenomenon. Consider a fixed-wage environment like the one we started this section with. Another way to conceptualize that situation would be one in which achieving higher-surplus outcome depends crucially on reciprocity. If the employee reciprocates high wages with high effort, the "gift-exchange" theory of labor relationships (Akerlof, 1982; Gneezy and List, 2006), then the undesirable but unique subgame-perfect equilibrium associated with fixed-wage contracts—lowest effort, lowest wage and lowest total surplus—may be avoidable. Highly reciprocal employees may be induced to exert high effort with relatively high wages. The danger is, of course, that any given wage may not be perceived as particularly generous and either fail to elicit positive reciprocity or, worse yet, elicit negative reciprocity, reducing even intrinsic motivations to exert effort. If a firm could separate strongly reciprocal employees from the weakly reciprocal ones, then a possible efficiency-maximizing strategy would be to pay the highly reciprocal a high wage thereby eliciting high effort while, at the same time, pay weak reciprocators the lowest possible wage since higher wages would not result in more effort.

One way this strategy could be implementable would be by inducing, through identification processes, high and low reciprocity. While both theory and empirics are sparse on this possibility, experimental findings in Butler (2014) suggest that by manipulating the relative status of identity categories firms could affect how wage-

reciprocal employees are. In particular, one plausible interpretation of his results are that high status induces more reciprocal behavior. Some implications for the optimal organization of the firm are straightforward and seem to match stylized facts. For example, for otherwise identical individuals performing identical tasks for a *fixed* wage, since low status makes surplus-enhancing gift exchange less viable, low status individuals should be (and typically are) paid a lower wage. On the other hand, if the same individual were assigned to a category imparting high status (working in the CEO's office), status itself could induce a level of reciprocity making a high wage optimal from the firm perspective through gift exchange.

5 Concluding thoughts

Over the past several decades, the theory of how social categorization and identification processes affect behavior has developed an impressive array of tested and testable implications. While starting in social psychology, the relevant body of research has recently garnered the attention of economists and has continued to develop, largely separately, in both disciplines.

In the economics literature, this has yielded two important classes of theoretical models—one focusing directly on how categories enter preferences (AK) and another investigating how identity may enter indirectly into preferences through beliefs (BT). Both of these models yield important insights which remain largely untested. Both of them miss important insights as well, being almost completely silent on the second wave of SIT which emphasizes how identity colors cognition through, e.g., motivated reasoning. Specifically, while there has been substantial research on how a desire to maintain a specific self-image or identity (as, e.g., a pro-social individual) may affect beliefs about oneself and one's own actions through self-signaling (Bénabou and Tirole, 2006b,a, 2016), consideration of how social categorization affects "intergroup beliefs"—beliefs about own- and other-group members' characteristics and preferences in general, such as the contrast effect mentioned in Section 3.2 above—has thus far not entered the economic discourse.

More importantly, both the economics and social psychology literature fail to consider the wider context and formally model categorization as the equilibrium outcome of strategic interaction. In this fullest specification, the role of identity in societal coordination may be highlighted and mark a complete turn-around from SIT's beginning as a way to rationalize discrimination rather than enhance cooperation. This last point is the most important for IO in particular, as the benefits of cooperation and coordination with respect to within-firm efficiency are a fundamental concern for the field with measurable implications.

References

Aberson, Christopher L, Michael Healy, and Victoria Romero (2000), "Ingroup bias and self-esteem: A meta-analysis." *Personality and social psychology review*, 4, 157–173.

Akerlof, George A (1982), "Labor contracts as partial gift exchange." *The quarterly journal of economics*, 97, 543–569.

Akerlof, George A and William T Dickens (1982), "The economic consequences of cognitive dissonance." *The American economic review*, 72, 307–319.

Akerlof, George A. and Rachel E. Kranton (2000), "Economics and identity." *Quarterly Journal of Economics*, CVX.

Akerlof, George A and Rachel E Kranton (2002), "Identity and schooling: Some lessons for the economics of education." *Journal of economic literature*, 40, 1167–1201.

Akerlof, George A. and Rachel E. Kranton (2010), *Identity Economics: how our identities shape our work, wages, and well-being*. Princeton University Press, Princeton, New Jersey.

Ashforth, Blake E, Spencer H Harrison, and Kevin G Corley (2008), "Identification in organizations: An examination of four fundamental questions." *Journal of management*, 34, 325–374.

Ball, Sheryl, Catherine Eckel, Philip J Grossman, and William Zame (2001), "Status in markets." *The Quarterly Journal of Economics*, 116, 161–188.

Ball, Sheryl and Catherine C Eckel (1998), "The economic value of status." *The Journal of socio-economics*, 27, 495–514.

Bartlett, Frederic C (1932), "Remembering: An experimental and social study." *Cambridge: Cambridge University*.

Bénabou, Roland and Jean Tirole (2006a), "A cognitive theory of identity, dignity, and taboos."

Bénabou, Roland and Jean Tirole (2006b), "Incentives and prosocial behavior." *The American economic review*, 96, 1652–1678.

Bénabou, Roland and Jean Tirole (2016), "Mindful economics: The production, consumption, and value of beliefs." *The Journal of Economic Perspectives*, 30, 141–164.

Bernhard, Helen, Urs Fischbacher, and Ernst Fehr (2006), "Parochial altruism in humans." *Nature*, 442, 912–915.

Bodenhausen, Galen V, Sonia K Kang, and Destiny Peery (2012), "Social categorization and the perception of social groups." *The Sage handbook of social cognition*, 318–336.

Bolton, Gary E and Axel Ockenfels (2000), "Erc: A theory of equity, reciprocity, and competition." *American economic review*, 166–193.

Branthwaite, Alan and Jane E Jones (1975), "Fairness and discrimination: English versus welsh." *European Journal of Social Psychology*, 5, 323–338.

Brown, Rupert (2000), "Social identity theory: Past achievements, current problems and future challenges." *European journal of social psychology*, 30, 745–778.

Butler, Jeff, Danila Serra, and Giancarlo Spagnolo (2016a), "Motivating whistleblowers." unpublished working paper.

Butler, Jeffrey V, , Paola Giuliano, and Luigi Guiso (2016b), "Trust and cheating." *The Economic Journal*, 126, 1703–1738.

Butler, Jeffrey V (2014), "Trust, truth, status and identity: An experimental inquiry." *The BE Journal of Theoretical Economics*, 14, 293–338.

Butler, Jeffrey V, Pierluigi Conzo, and Martin Leroch (2013), "Social identity and punishment." unpublished working paper.

Campbell, Donald T (1965), "Ethnocentric and other altruistic motives." In *Nebraska symposium on motivation*, volume 13, 283–311.

Chang, Daphne, Roy Chen, and Erin Krupka (2015), "Social norms and identity dependent preferences." unpublished working paper.

Charness, Gary and Martin Dufwenberg (2006), "Promises and partnership." *Econometrica*, 74, 1579–1601.

Charness, Gary and Matthew Rabin (2002), "Understanding social preferences with simple tests." *The Quarterly Journal of Economics*, 117, 817–869.

Charness, Gary, Luca Rigotti, and Aldo Rustichini (2007), "Individual behavior and group membership." *The American Economic Review*, 97, 1340–1352.

Chen, Roy and Yan Chen (2011), "The potential of social identity for equilibrium selection." *The American Economic Review*, 101, 2562–2589.

Chen, Yan and Sherry Xin Li (2009), "Group identity and social preferences." *The American Economic Review*, 99, 431–457.

Coase, R H (1937), "The Nature of the Firm." *Economica, New Series*, 4, 386–405.

Cohn, Alain, Ernst Fehr, and Michel André Maréchal (2014), "Business culture and dishonesty in the banking industry." *Nature*, 516, 86–89.

Crawford, Vincent P and Joel Sobel (1982), "Strategic information transmission." *Econometrica: Journal of the Econometric Society*, 1431–1451.

Eckel, Catherine C and Philip J Grossman (2005), "Managing diversity by creating team identity." *Journal of Economic Behavior & Organization*, 58, 371–392.

Edward, Banfield (1958), *The moral basis of a backward society*. Glencoe.

Epley, Nicholas and Thomas Gilovich (2016), "The mechanics of motivated reasoning." *The Journal of Economic Perspectives*, 30, 133–140.

Fehr, Ernst and Klaus M Schmidt (1999), "A theory of fairness, competition, and cooperation." *The quarterly journal of economics*, 114, 817–868.

Festinger, Leon (1957), *A Theory of Cognitive Dissonance*. Stanford University Press.

Gneezy, Uri and John A List (2006), "Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments." *Econometrica*, 74, 1365–1384.

Graham, Carol, Soumya Chattopadhyay, and Mario Picon (2010), "Adapting to adversity: happiness and the 2009 economic crisis in the united states." *Social Research: An International Quarterly*, 77, 715–748.

Guala, Francesco, Luigi Mittone, and Matteo Ploner (2013), "Group membership, team preferences, and expectations." *Journal of Economic Behavior & Organization*, 86, 183–190.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales (2013), "Time varying risk aversion." Technical report, National Bureau of Economic Research.

Güth, Werner, M Vittoria Levati, and Matteo Ploner (2008), "Social identity and trust–an experimental investigation." *The Journal of Socio-Economics*, 37, 1293–1308.

Hargreaves-Heap, Shaun P and Daniel John Zizzo (2009), "The value of groups." *The American Economic Review*, 99, 295–323.

Harris, Donna, Benedikt Herrmann, and Andreas Kontoleon (2010), "What is the nature and social norm within the context of in-group favouritism?"

Haslam, S Alexander and Naomi Ellemers (2005), "Social identity in industrial and organizational psychology: Concepts, controversies and contributions." *International review of industrial and organizational psychology*, 20, 39–118.

Hertel, Guido and Norbert L Kerr (2001), "Priming in-group favoritism: The impact of normative scripts in the minimal group paradigm." *Journal of Experimental Social Psychology*, 37, 316–324.

Hogg, Michael A, Dominic Abrams, Sabine Otten, and Steve Hinkle (2004), "The social identity perspective intergroup relations, self-conception, and small groups." *Small group research*, 35, 246–276.

Hogg, Michael A and Deborah I Terry (2000), "Social identity and self-categorization processes in organizational contexts." *Academy of management review*, 25, 121–140.

Hornsey, Matthew J (2008), "Social identity theory and self-categorization theory: A historical review." *Social and Personality Psychology Compass*, 2, 204–222.

Jetten, Jolanda, Russell Spears, and Antony SR Manstead (1997), "Strength of identification and intergroup differentiation: The influence of group norms." *European Journal of Social Psychology*, 27, 603–609.

Jost, John T, Mahzarin R Banaji, and Brian A Nosek (2004), "A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo." *Political psychology*, 25, 881–919.

Jost, John T and Roderick M Kramer (2003), "The system justification motive in intergroup relations." *From prejudice to intergroup emotions: Differentiated reactions to social groups*, 227–245.

Kőszegi, Botond (2014), "Behavioral contract theory." *Journal of Economic Literature*, 52, 1075–1118.

Kranton, Rachel, Matthew Pease, Seth Sanders, and Scott Huettel (2016), "Group bias, identity, and social preferences." unpublished working paper.

Kumru, Cagri S and Lise Vesterlund (2008), "The effect of status on voluntary contribution." unpublished working paper.

Loewenstein, George F, Elke U Weber, Christopher K Hsee, and Ned Welch (2001), "Risk as feelings." *Psychological bulletin*, 127, 267.

McLeish, Kendra N and Robert J Oxoby (2007), "Identity, cooperation, and punishment." unpublished working paper.

Mullin, Barbara-Ann and Michael A Hogg (1999), "Motivations for group membership: The role of subjective importance and uncertainty reduction." *Basic and Applied Social Psychology*, 21, 91–102.

Mummendey, Amelie, Bernd Simon, Carsten Dietze, Melanie Grünert, Gabi Haeger, Sabine Kessler, Stephan Lettgen, and Stefanie Schäferhoff (1992), "Categorization is not enough: Intergroup discrimination in negative outcome allocation." *Journal of Experimental Social Psychology*, 28, 125–144.

Sherif, Muzafer (1961), *Intergroup conflict and cooperation: The Robbers Cave experiment*, volume 10. University Book Exchange Norman, OK.

Smith, Eliot R and Susan Henry (1996), "An in-group becomes part of the self: Response time evidence." *Personality and Social Psychology Bulletin*, 22, 635–642.

Spears, Russell and Anthony SR Manstead (1989), "The social context of stereotyping and differentiation." *European Journal of Social Psychology*, 19, 101–121.

Summers, Lawence H. (2002), "Harvard university statement of values." URL `http://www.harvard.edu/president/speeches/summers_2002/values.php`.

Tajfel, H., C. Flament, M.G. Billig, and R.F. Bundy (1971), "Social categorization and intergroup behavior." *European Journal of Social Psychology*, 1, 149–177.

Turner, J. (1978), *Differentiation Between Social Groups*, chapter Social Categorization and Social Discrimination in the Minimal Group Paradigm, 101–140. Academic Press, London.

Turner, John C (1975), "Social comparison and social identity: Some prospects for intergroup behaviour." *European journal of social psychology*, 5, 1–34.

Turner, John C, Michael A Hogg, Penelope J Oakes, Stephen D Reicher, and Margaret S Wetherell (1987), *Rediscovering the social group: A self-categorization theory*. Basil Blackwell.

Turner, John C and Henri Tajfel (1986), "The social identity theory of intergroup behavior." *Psychology of intergroup relations*, 7–24.

Van Knippenberg, Daan (2000), "Work motivation and performance: A social identity perspective." *Applied Psychology*, 49, 357–371.

Willer, Robb (2009), "Groups reward individual sacrifice: The status solution to the collective action problem." *American Sociological Review*, 74, 23–43.