Documents de travail

«The Signaling Value of Social Identity»

Auteur

Arnaud Wolff

Document de Travail n° 2022 – 15

Avril 2022

Bureau d'Économie Théorique et Appliquée BETA

www.beta-economics.fr

@beta_economics

Contact :
jaoulgrammare@beta-cnrs.unistra.fr

The Signaling Value of Social Identity

Arnaud Wolff*

April 2022

Abstract

This paper proposes a theory of social identity adoption and expression, which ties the choice of social identity to material and social benefits present in an individual's social environment. I argue that in an environment in which receivers aim at uncovering the sender's motives and commitments, the beliefs and values adopted by an individual can serve as a signal of trustworthiness. In such an environment, individuals are expected to adopt the social identity which will provide them with the greatest amount of (social) benefits. I formalize this choice in a game-theoretic framework, embed in a broader niche selection structure. I argue that the main predictions of the model help illuminate several empirical findings, such as the malleability of beliefs and values, the resistance of beliefs and values to evidence, and the existing correlation between beliefs and values and individual-level traits such as personality.

Keywords: Social Identity, Beliefs, Values, Trustworthiness, Social Incentives
JEL Codes: C72, C73, D83, D91

*Bureau d'Economie Théorique et Appliquée (BETA) - Université de Strasbourg - 61 Avenue de la Forêt Noire, 67000 Strasbourg - France. e-mail: arnaudwolff@unistra.fr. This paper has benefited from the kind and helpful comments from Jocelyn Donze, Frédéric Koessler and Gisèle Umbhauer.

Introduction

Individuals are very often eager to publicly share their beliefs and values on sensitive topics (e.g., "I believe that restricting immigration is wrong"), as well as their memberships in social groups (e.g., "I am a Social-Democrat"). Such *identity signals* tend to transmit information to others about the signaler's membership in a subset (or group) of individuals (Smaldino and Turner, 2021).

The reason why individuals tend to attach value to their *social identity*, and why they are so eager to publicly express it, has been a matter of intense debate since the genesis of social identity theory (Tajfel, 1974, Tajfel and Turner, 1979). Psychological theories tend to emphasize the psychological benefits that individuals reap from their perceived group memberships (Oakes and Turner, 1980, Lemyre and Smith, 1985, Brewer, 1991). In line with psychological theories, theoretical work on social identity tends to emphasize the psychological benefits that individual derive from their social identity. For instance, Akerlof and Kranton (2000) proposed that individuals suffer disutility from not following group norms and prescriptions. Shayo (2009) assumes that agents derive utility from the status of their social group and disutility from their perceived distance (in terms of attributes) from other group members. Similarly, Akerlof (2016) assumes that social identity is linked to esteem, with agents caring about how their group is esteemed. In their model of identity management, Bénabou and Tirole (2011) also argue that individuals behave in such a way as to enhance their self-views, striving to reduce cognitive dissonance. This work therefore tends to further stress the psychological benefits (e.g., self-image or group status) that individuals derive from their choice of social identity.

Another line of research aims at understanding the *material and social* benefits that individuals derive from valuing and expressing their social identity. For instance, Yamagishi and colleagues have argued that social identity is valued because it serves as a signal for coordination and cooperation opportunities (Yamagishi, Jin, and Kiyonari, 1999, Yamagishi and Kiyonari, 2000). The idea is that social identities clarify the boundaries of cooperative relationships, and ensure that everyone is coordinated about these boundaries (Pietraszewski, 2020). Similarly, Smaldino and colleagues argue that the adoption and expression of social identity allows similar individuals to coordinate and cooperate with each other (Smaldino, 2019, Smaldino and Turner, 2021). Finally, Carvalho (2016) argues that social identity can be seen as a club good, with identity-based organization helping to reduce the free-rider problem in collective action. This research tradition therefore aims at uncovering the material and social benefits (and costs) that underlie the important role that social identity plays in people's lives.

The present paper follows this research tradition by proposing a theory of social identity adoption and expression. While researchers often take individual preferences as given, I aim to link the choice of a social identity to (material and social) incentives in an individual's social environment. Therefore, I argue in this paper that one can better understand an individual's

choice of social identity by understanding the incentives that they face. Second, I try to clarify why individuals are usually eager to (publicly) express their social identity. Building on Loury (1994)'s work, I argue that the choice of a social identity often reveals information about an individual's willingness to cooperate. In an environment in which receivers aim at uncovering the sender's (not readily observable) motives and commitments, beliefs and values adopted by the sender are evaluated against beliefs and values adopted by other senders whose motives and commitments may already be known. In such a context, by adopting specific beliefs and values, senders *pool* (respectively *separate*) from others who adopted similar (respectively dissimilar) beliefs and values and whose motives and commitments are publicly known. The choice of a social identity then essentially signals to others the sender's social commitments. This choice, as described above, is expected to be a function of the *material and social* benefits that the sender is expected to derive from cooperating with different groups of individuals.

In order to highlight the trade-off in the choice of social identity, I formalize this choice by having a *sender* play two repeated Asynchronous Prisoner's Dilemma (APD), each with a specific *receiver* (defined as a group of individuals having adopted a given social identity), with both games being preceded by a signaling stage in which the sender has to choose which social identity to adopt. This strategic interaction is embed in a broader niche selection structure (Smaldino, Lukaszewski, von Rueden, and Gurven, 2019). The main idea is that players can condition their strategy in the repeated APD on the sender's decision in the signaling stage. Importantly, the value of the sender's continuation probability across both games is assumed to be a function of the (differential) benefits that the sender might reap from cooperating with each receiver. I show that by adopting a given social identity in the signaling stage, the sender can signal high continuation probability in the repeated APD, therefore reassuring the receiver(s) that the sender will be around in the future to reciprocate favors. Therefore, under appropriate conditions, the adoption of a given social identity can function as a signal of trustworthiness.

The model makes several predictions. First, the choice of social identity is expected to follow social and material incentives. Second, individuals will often want to hold on, and defend their social identity. Third, individuals are expected to balance the benefits from cooperating with different groups of individuals when deciding which social identity to adopt. Fourth, individual-level traits (such as personality and cognition) will be correlated with specific beliefs and values. The existing empirical evidence, reviewed in the last section of the paper, appears to support the main predictions of this model.

Discussion of the Main Argument

The argument advanced here is that social identity—defined as the process of affiliation (or identification) with a given social group—can act as a signal of *trustworthiness*, defined as the expectation of future reciprocation (Jordan, Hoffman, Bloom, and Rand, 2016, Jordan

and Rand, 2017). That is, the idea is that individuals will, under certain circumstances, strategically adopt the beliefs, ideologies and values (i.e., the social identities) that signal their cooperative intent to others in the communities or networks in which they find themselves. According to the argument presented in this paper, then, individual values (or preferences) do not shape the choice of a social identity. Rather, it is the existing incentives in the individual's social environment that influence which social identity (beliefs and values) she will express and adopt.

The question immediately arises as to how beliefs and values might signal trustworthiness. After all, beliefs and values are internal states that can not be observed by others. My focus here is on social identity as outwardly expressed, and I will therefore consider as a signal the *expression* of one's beliefs and values. Yet, this view need not be inconsistent with the idea that these beliefs and values are internally (deeply) felt (Brewer, 1991), given that one can expect individuals to internalize those beliefs and values that they are incentivized to hold and express (Melnikoff and Strohminger, 2020, Schwardmann, Tripodi, and van der Weele, Forthcoming).

While the argument in this paper is that the public expression of one's beliefs and values can signal trustworthiness, this need not necessarily be so. We might actually think of a world in which beliefs and values are completely uncorrelated with intentions to cooperate, where people freely exchange ideas, debate, and argue without making any inferences about their interlocutor's trustworthiness. In fact, this is an equilibrium of Loury (1994)'s "expression game", which is a game played between a sender (e.g., a politician) and a receiver (e.g., an audience). At such an equilibrium (an equilibrium of the game can be seen as a mutually agreed upon *convention*), senders are sincere and truthfully express their ideas, and receivers accept messages for their truth value. Yet, multiple equilibria exist. Another equilibrium (convention) of this game has receivers *read between the lines* of the messages sent by senders, trying to judge their motives and commitments (which are unobservable internal states, and which might differ from those of the receivers), and senders, taking this into account, carefully craft their messages in order not to alienate their audience. At such an equilibrium, messages sent by senders are evaluated against messages sent by other senders whose motives and commitments may already be known. Receivers do not uniquely judge statements for their truth value, but rather wonder what *kind* of person would express herself in that way. This commonly known inference process implies that messages sent carry meaning above-and-beyond their immediate truth value: by sending specific messages, senders *pool* (respectively *separate*) from others who sent similar (respectively dissimilar) messages and whose motives and commitments are publicly known.

Such a convention, coupled with the observation that different social groups have associated different beliefs, ideologies and values, constitutes the starting point of the argument. In fact, the beliefs, ideologies and values that constitute the core tenets of a given social group often are in conflict with those of other groups, such as when one group favors immigration while the other acts to prevent it, when one group strives to extend rights to disadvantaged minorities while the other favors the status quo, or when one group favors free speech while

the other expects speech restrictions on sensitive matters. That is, individuals belonging to different social groups often have worldviews that are at variance, beliefs and values that would lead to different policies, and moral views that prioritize different issues (Jacoby, 2014, Van Bavel and Pereira, 2018). By deciding to join a given social group, with associated core tenets, individuals then indirectly affiliate with a set of beliefs and values that might define them as individuals. Ultimately, then, adopting specific packages of beliefs, ideologies and values, or taking a specific stance on controversial matters in an environment in which receivers judge messages with respect to those sent by others whose motives and commitments are publicly known, amounts to *siding* with others sending similar messages, therefore displaying a specific set of motives and commitments to audience members.

Recent experimental evidence has shown that people seem to treat cues of agreement and disagreement with specific (politicized) statements as markers of *social* (and not just *epistemic*) commitment to a specific group or coalition (Pietraszewski, Curry, Petersen, Cosmides, and Tooby, 2015), implying that we might find ourselves at the above-described equilibrium. *Knowing this*, the actor, who decides to adopt and profess a certain belief or opinion, sides with others sending similar messages, therefore signaling to audience members which group she wants to affiliate with (and, ultimately, who she intends to cooperate with). These signals are *credible* in the sense that taking a public stance in favor of a given group (e.g., by adopting and broadcasting its core tenets) will inevitably reduce the trust members of other groups will be willing to grant. These commonly known reduced outside options will have the alternative effect of increasing an individual's perceived trustworthiness among members of the social group she has pledged to join (Park and van Leeuwen, 2015, Williams, 2021).

Therefore, the argument is that individuals are expected to strategically adopt the beliefs and values that signal their trustworthiness to others in their communities. Importantly, this is expected to be important only in those situations in which the problem of trust has not yet been resolved, and only at the above described equilibrium (convention). When trustworthiness is guaranteed, the expression of social identity might serve other purposes, such as coordination with similar others (Smaldino, 2019). Now, the problem of trust is important enough for the above-described mechanisms to be at work in a wide range of situations. The next section describes a game-theoretic model aimed at formalizing the process of social identity adoption, given that the above-described argument does not make predictions about which social identity an individual is expected to adopt.

Game-Theoretic Analysis of the Choice of Social Identity

This section contains a game-theoretic analysis of the choice of social identity, aimed at describing the main incentives underlying the adoption of specific packages of beliefs and values. The games (asynchronous repeated Prisoner's Dilemmas) are played between a *sender*

and two *receivers* (audiences). The games are embed in a broader niche selection structure.

Model Setup

Niche Selection. Society is characterized by a set N of individuals, a set J of social niches, and a set G of social groups. In this paper, I follow Smaldino, Lukaszewski, von Rueden, and Gurven (2019) in defining ***social niches*** as particular sets of incentives, *means* to extract resources (material, social, etc.) from the environment. The concept of a social niche can be seen as related to the concept of *focus* (or *foci*) developed by Feld (1981, p.1016), which he defines as "a social, psychological, legal, or physical entity around which joint activities are organized (e.g., workplaces, voluntary organizations, hangouts, families, etc.)", although *foci* encompass a wider range of entities (e.g., a family can not be chosen and therefore can not be considered as a social niche). The principal idea is that different behavioral (or personality) profiles are differentially suited to different niches.[1] While niches are characterized by traits that "remain fixed because of the niche's intrinsic ecological and social characteristics" (Smaldino, Lukaszewski, von Rueden, and Gurven, 2019, p.1281), ***social groups*** are collections of individuals that share some core beliefs, values, ideologies, and/or norms of conduct. Therefore, throughout the paper, I will use the term *social group* (or *group*) quite generally in order to refer to ideologically, socially or culturally defined groups, such as "Liberals", "Conservatives", "Animal Activists", "Christians", "Climate Deniers", "Flat Earthers", etc.

Every social niche $j \in J$ has an associated *ideal trait profile* γ_j, that is, an associated vector of traits such that an individual endowed with this exact trait profile would be optimally suited for this niche (Smaldino, Lukaszewski, von Rueden, and Gurven, 2019). Therefore, let $\gamma_j = (\gamma_{j_1}, \gamma_{j_2}, ..., \gamma_{j_p})$, with each γ_{j_p} being a bounded random variable whose value represents niche j's ideal value for the p^{th} trait. The ideal trait profile γ_j that characterizes each social niche j can then be thought of as a description of the incentives faced by individuals in different niches, insofar as individuals will be incentivized to join (respectively depart) niches whose ideal trait profile is similar (respectively dissimilar) to their own trait profile. Each social niche j is populated by a set $N_j \subset N$ of individuals. For simplicity, we assume that all members of a given social niche j belong to social group $g_j \in G$.

In this paper, we take the perspective of a focal individual, the sender s, who starts the game embedded in a community, subsequently called the sender's *home community* m. Members of m are assumed to belong to the social group g_m, with associated beliefs, ideologies and values, which have been transmitted to the sender. Therefore, the sender starts the game belonging to the social group $g_m \in G$, hence with a given *social identity* I_{g_m}. The natural interpretation is to consider the sender as having *learned* the beliefs, ideologies and values

[1] The quintessential example of a social niche in an industrialized society is a professional occupation (scientist, broker, operative, craftsman, teacher, engineer, etc.); it can also be a sporting (coach, educator or team player), artistic (band member or independent artist), political (activist, candidate or party member), or social (volunteer) activity. These particular activities can be seen as ideally suited for different *types* of individuals, and it is expected that individuals will choose to join the social niches that fit their behavioral (personality) profile the most.

(associated to g_m) of the members of the community in which she has grown and developed, with the game being first played while the sender still finds herself in her home community. More generally, this model is expected to apply to any situations in which the sender faces the choice of moving from one (home) community to another, for reasons that are independent of the sender's beliefs and values. Members of the sender's home community constitute the first type of audience (or *receiver*), r_m.

In the first stage, the sender, being embedded in her home community, chooses which *social niche* $j \in J$ she wants to join. The sender s has an associated *trait profile* θ_s, which can be characterized as a vector of P individual behavioral and cognitive characteristics. These together can be said to represent the individual's *cognitive ability* and *personality*.[2] Therefore, let $\theta_s = (\theta_{s_1}, \theta_{s_2}, ..., \theta_{s_p})$, with each θ_{s_p} being a bounded random variable whose value represents individual s's endowment for the p^{th} trait. Importantly, in this first stage, the sender does *not* choose a social group. As discussed before, this first assortment is expected to be a function of the intrinsic characteristics of the individual and of those of the social niche (Feld, 1981, 1982, 1997). The first stage assortment is therefore not based in any way on individual beliefs or values, which are expected to emerge endogenously from the mechanisms in the model.

Strategic Interaction. The second stage is a *signaling* stage. Once s has decided to join her preferred social niche j, she starts to interact with other individuals (the set $N_j \subset N$) who themselves decided to join that niche (e.g., the sender might decide to join an orchestra which is filled with other music players; she might decide to join a doctoral program in which she interacts with other students; or, alternatively, she might decide to join a law firm in which she interacts with other lawyers). As described above, these other niche members are assumed to belong to the same social group g_j. This simplifying assumption stems from Bonica (2014)'s observation that there exist large differences in ideological distributions across industries and professional occupations (i.e., social niches or *foci*), with occupations such as academia, entertainment or media being skewed to the left, while occupations such as banking and finance, building and construction or agriculture being skewed to the right.[3] Other niche members constitute the second type of audience (or receiver), r_j. At this stage, the sender has to decide (i) whether to hold on the beliefs, ideologies and values that she has learned in her home community (i.e., whether to hold on her identity I_{g_m}), or (ii) whether to

[2] An individual's personality represents her traits and behavior that are relatively stable across time and contexts. They are largely *innate*, in the sense that they are likely "built up from variation in a large number of [...] basal decision-making parameters. Variations in neuromodulatory systems may underlie the differential tuning of these parameters across individuals" (Mitchell, 2020, p.124).

[3] Audiences (r_m and r_j) are therefore assumed to be homogeneous in terms of beliefs, ideologies and values. In reality, disagreement exists in individual social networks, and individuals are therefore unlikely to encounter ideologically homogeneous audiences (Huckfeldt, Mondak, Hayes, Pietryka, and Reilly, 2013). One way to interpret this assumption is to consider that there is a social group *predominantly* represented among both audiences (receivers), and that the associated beliefs, ideologies and values are enforced by community members. Yet, adding different social groups, more or less equally represented among members of one's community, while potentially closer to reality, is not expected to alter the model's main predictions.

adopt the beliefs, ideologies and values that prevail among the members of the social niche she as decided to join (i.e., whether to adopt a new identity I_{g_j}). There are of course cases in which the social group that is most represented in the home community is the same as in the social niche the sender has decided to join, in which case this model does not bring interesting insights. We will, in this paper, focus on cases in which I_{g_m} differs from (or conflicts with) I_{g_j}.

In the third, *partner choice* stage, the sender s plays a repeated Asynchronous Prisoner's Dilemma (APD) with both receivers, r_m and r_j. She therefore plays two games: game m with r_m, and game j with r_j. In the first round of their respective games, receivers decide whether to *Cooperate (C)* in the PD, which implies accepting the sender, or *Defect (D)*, which implies rejecting the sender. Accepting the sender amounts to invest in the relationship with the sender, providing her with a benefit k_i with $i \in \{m, j\}$ at a cost c, with $k_i - c > 0$, while rejecting the sender amounts to refuse to invest in a relationship with her. If the sender is rejected, the game ends, with both players earning payoffs equal to 0. If the sender is accepted, we move on to the second round with probability δ_i, at which point the sender decides whether to *Cooperate (C)* or to *Defect (D)*. Reciprocating (cooperating) also costs c to the sender, and provides benefit b to the receiver, with $b - c > 0$, while defecting amounts to bestowing no benefit to the receiver (and paying no cost).

δ_i is meant to capture the idea that the sender might not be there to reciprocate the favor bestowed by the receiver. Hence, δ_i captures the continuation probability of the sender, which is fixed throughout the game.[4] For simplicity, we assume that the sender s can either have a *high (h)* or *low (l)* continuation probability, with $0 < \delta_i^l < \delta_i^h < 1$. Importantly, for our purposes, δ_i need not be the same across both games (i.e., it need not be the case that $\delta_m = \delta_j$). While δ_i can be thought of as exogenously given, it can also be seen as describing the *incentives* faced by the sender (Jordan, Hoffman, Bloom, and Rand, 2016, Jordan and Rand, 2017): a low continuation probability can realize due to insufficient exposure to mechanisms incentivizing cooperation (e.g., direct or indirect reciprocity, institutions, etc.), while a high continuation probability can stem from high enough exposure to such mechanisms. We will, in this paper, take this latter perspective, by endogenizing the value of δ_i across both games.

Both games are repeated until the relationship between the receiver and the sender is terminated, either because one of the players has defected (played D), or because the sender is not around anymore (i.e., $(1 - \delta_i)$ realizes). That is, we assume that defection from either r_i or s effectively terminates the interaction, reflecting the idea that the other player can not be trusted to cooperate in the future. Finally, if the sender decides not to cooperate with any receiver i, she gets benefit $\bar{\omega} = 0$, which can be considered as the value of her outside option (normalized to zero for simplicity). The structure of the repeated APD between s and receiver r_i is shown in Figure 1. N represents Nature, and the payoffs are such that r_i's payoff is noted first, and s's payoff is noted second.

[4] We assume that the continuation probability of the receivers is 1, such that receivers are always guaranteed to be there for another round.

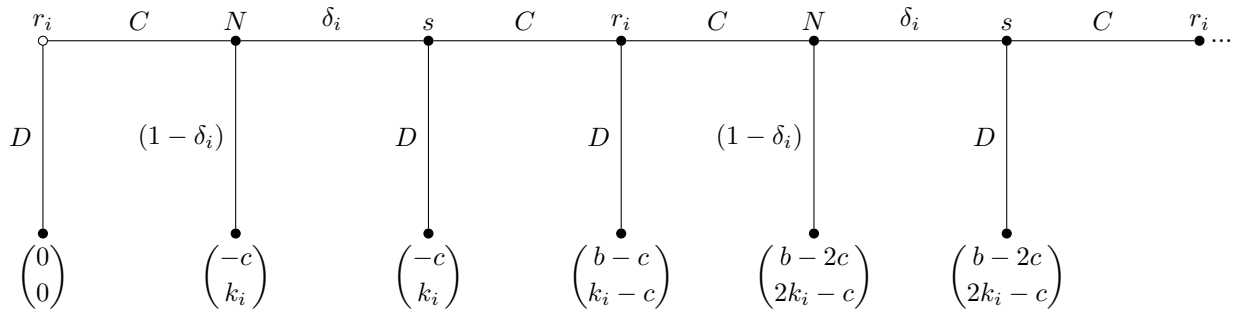

Figure 1: Repeated Asynchronous Prisoner's Dilemma in the *Partner Choice* stage.

Model Resolution

Partner Choice Game. We start by seeking Subgame-Perfect Nash Equilibria (SPNE) of the repeated APD.

Can cooperation (i.e., both players playing C throughout the game) be a SPNE of the game? We start by investigating whether it can be beneficial for r_i to play C when s plays C throughout the game. If r_i decides to play C at all of his decision nodes (a strategy we will call $ALLC_{r_i}$), given that s always plays C (plays $ALLC_s$), then his expected payoff $E[ALLC_{r_i}|ALLC_s]$ is:

$$\begin{aligned}
E[ALLC_{r_i}|ALLC_s] &= (1-\delta_i)(-c) + \delta_i(1-\delta_i)(b-2c) + \delta_i^2(1-\delta_i)(2b-3c) \\
&\quad + \delta_i^3(1-\delta_i)(3b-4c) + ... \\
E[ALLC_{r_i}|ALLC_s] &= -c(1-\delta_i)(1+2\delta_i+3{\delta_i}^2+...) + b\delta_i(1-\delta_i)(1+2\delta_i+3{\delta_i}^2+...) \\
E[ALLC_{r_i}|ALLC_s] &= \frac{-c(1-\delta_i)}{(1-\delta_i)^2} + \frac{b\delta_i(1-\delta_i)}{(1-\delta_i)^2} \\
E[ALLC_{r_i}|ALLC_s] &= \frac{(-c+b\delta_i)}{(1-\delta_i)}.
\end{aligned}$$

If r_i decides to play D at the first node, then his payoff will be equal to 0. Therefore, for r_i to be willing to play $ALLC_{r_i}$ when s plays $ALLC_s$, it needs to be the case that:

$$\begin{aligned}
\frac{(-c+b\delta_i)}{(1-\delta_i)} &\geq 0 \\
b\delta_i &\geq c \\
\delta_i &\geq \frac{c}{b}.
\end{aligned}$$

It follows that as long as $\delta_i \geq \frac{c}{b}$, then r_i is incentivized to play $ALLC_{r_i}$ when s also plays $ALLC_s$. Importantly, if $\delta_i \geq \frac{c}{b}$ holds, then r_i is incentivized to play C in every subgame of the game (as long as s also cooperates).

We now investigate the conditions under which s would be willing to play $ALLC_s$ from her first decision node on, when r_i plays $ALLC_{r_i}$. If s decides to always play $ALLC_s$, when

r_i plays $ALLC_{r_i}$, then her expected payoff $E[ALLC_s|ALLC_{r_i}]$ is:

$$E[ALLC_s|ALLC_{r_i}] = (1-\delta_i)(2k_i - c) + \delta_i(1-\delta_i)(3k_i - 2c) + \delta_i^2(1-\delta_i)(4k_i - 3c) + ...$$
$$E[ALLC_s|ALLC_{r_i}] = k_i(1-\delta_i)(1 + \delta_i + \delta_i^2 + ...) + (-c + k_i)(1-\delta_i)(1 + 2\delta_i + 3\delta_i^2 + ...)$$
$$E[ALLC_s|ALLC_{r_i}] = \frac{k_i(1-\delta_i)}{(1-\delta_i)} + \frac{(-c+k_i)(1-\delta_i)}{(1-\delta_i)^2}$$
$$E[ALLC_s|ALLC_{r_i}] = k_i + \frac{(-c+k_i)}{(1-\delta)}.$$

If s instead decides to play D at her first decision node, then her payoff will be equal to k_i. For s to be willing to play $ALLC_s$ when r_i plays $ALLC_{r_i}$, then it needs to be the case that:

$$k_i + \frac{(-c+k_i)}{(1-\delta)} \geq k_i$$
$$\frac{(-c+k_i)}{(1-\delta)} \geq 0$$
$$k_i \geq c.$$

It follows that as long as $k_i \geq c$ is satisfied, then s is willing to play $ALLC_s$ from her first decision node on, as long as r_i plays $ALLC_{r_i}$. As before, if $k_i \geq c$ holds, then s is incentivized to play C in every subgame of the game (as long as r_i also cooperates). We can conclude that as long as $\delta_i \geq \frac{c}{b}$ and $k_i \geq c$ are satisfied, then both players playing C at every decision node is a SPNE of the game. On the other hand, if $\delta_i \geq \frac{c}{b}$ and/or $k_i \geq c$ do not realize, then the only SPNE of the game is for both players to play D at all their decision nodes (a strategy we will call $ALLD$). This is due to the fact that if the above conditions are not simultaneously satisfied, then at least one player will never play C. But if one player always plays D, then the other is never incentivized to play C, given that they would pay the costs of cooperation c without receiving any future benefits.

In fact, there are only two SPNE of the game: either (i) both players play D at all their decision nodes, or (ii) both players play C at all their decision nodes, when $\delta_i \geq \frac{c}{b}$ and $k_i \geq c$ both realize. To see why combinations of C and D can not be part of a SPNE of the game, first note that this could only happen when $\delta_i \geq \frac{c}{b}$ and $k_i \geq c$ both realize, given that we have determined that the only SPNE of the game when $\delta_i \geq \frac{c}{b}$ and/or $k_i \geq c$ do not realize is for both players to play $ALLD$. Assume that s plays D at one of her decision node x. The logic of SPNE therefore requires that r_i also plays D at his $(x-2)$ decision node, otherwise he will pay the costs of cooperation c without any further benefits. Similarly, if r_i plays D at one of his decision node x (which is not the initial node), then the logic of SPNE requires s to play D at her $(x-1)$ decision node. Therefore, if a player plays D at a decision node x, then both players necessarily play D until this decision node is attained at a SPNE of the game. Now, assume that s plays C at one of her decision node y. Given that $\delta_i \geq \frac{c}{b}$ holds,

then r_i will also play C at his $(y-2)$ decision node. Similarly, if r_i plays C at one of his decision node y (which is not the initial node), then s will also play C at her $(y-1)$ node, given that $k_i \geq c$ holds. Therefore, if a player plays C at a decision node y, then both players necessarily play C until this decision node is attained at a SPNE of the game, when $\delta_i \geq \frac{c}{b}$ and $k_i \geq c$ both realize. It follows that at a SPNE of the game, either players play ALLD, or they play ALLC (when $\delta_i \geq \frac{c}{b}$ and $k_i \geq c$ both hold).

The results of our analysis have shown that cooperation can be sustained at equilibrium if and only if the receiver is sufficiently confident that the sender will be there in the future to reciprocate the favor (i.e., if and only if δ_i is sufficiently large). Yet, how can the receiver be confident that the sender will be around in the future? Alternatively, how can the sender convince the receiver that she will be around in the future? The key idea is that receivers can condition their strategy in the partner choice stage to the sender's strategy in the signaling stage. Recall that the situation described is one in which receivers carefully monitor the sender's beliefs and values in order to infer her underlying motives and commitments. Hence, the sender's strategy in the signaling stage will inevitably influence the receiver's strategy in the partner choice stage. Therefore, what should be the sender's strategy in the signaling game, knowing that her choice might bring herself cooperative partners, while alienating some others?

Signaling Stage. This section will be dedicated at determining the optimal choice of social identity for the sender. This requires a description of how the values of the sender's continuation probabilities are set across both games.

We assume, for convenience, that receivers never accept senders that adopt *conflicting* identities. Admittedly, this is an extreme case. As described by Williams (2021), adopting a conflicting identity usually *reduces the probability* that members of an another social group might be willing to cooperate with us, but does not completely set it to zero. This assumption is made in order to simplify the analysis and focus on the trade-off in the choice of social identity. Therefore, the sender's strategy in the signaling stage ultimately amounts to *choose* one receiver with whom to cooperate over another, given that her choice of identity I_{g_i} will either attract, or alienate, some receiver. In particular, if s adopts I_{g_m}, then she knows that only r_m might be willing to cooperate with her. Alternatively, if s adopts I_{g_j}, then she knows that only r_j might be willing to cooperate with her. Hence, if the sender's continuation probability δ_i captures her likelihood of staying around in the future, then the sender's choice of identity in the signaling stage can be seen as analogous to *signaling* her continuation probability δ_i across both games.

In this paper, we assume that the sender's continuation probability across both games is a function of the benefits that she might reap from cooperating with r_m and/or r_j. In particular, if the sender is not enough exposed to mechanisms incentivizing cooperation, then we set $\delta_i = \delta_i^l < \frac{c}{b}$, and cooperation can not stabilize. On the other hand, if the sender is enough exposed to mechanisms incentivizing cooperation, then we set $\delta_i = \delta_i^h > \frac{c}{b}$, and cooperation can stabilize. Now, what determines whether $\delta_i = \delta_i^l$ or δ_i^h? The answer, it is

assumed, lies in the differential benefits that the sender might reap from cooperating with r_m and/or r_j.

In order to determine the value of the sender's continuation probability δ_i across both games, we start by *assuming* that $\delta_i = \delta_m^h = \delta_j^h$. That is, we assume that it is equally likely that the sender might be around in both games, reflecting her *choice* to cooperate with r_m and/or r_j. If $\delta_i = \delta_m^h = \delta_j^h$, we know that the sender is expected to hold on the beliefs, ideologies and values of her home community (i.e., hold on I_{g_m}) if the benefits generated from cooperating exclusively with r_m are (i) greater than her benefits from cooperating with r_j, and (ii) greater than the value of her outside option. This leads to the following two conditions:

(i) $k_m \geq k_j$,

(ii) $k_m \geq \delta_i c$.

Condition (i) simply requires that the benefits that she reaps from cooperating with her home community members are greater than the benefits that she reaps from cooperating with other social niche members. Condition (ii) is necessarily satisfied at a cooperative equilibrium, which requires $k_m \geq c$. Therefore, if condition (i) and (ii) are satisfied, we set $\delta_m = \delta_m^h > \delta_j = \delta_j^l$, given that the benefits from cooperating exclusively with r_m are greater than the benefits from cooperating with r_j. The sender is then incentivized to cooperate exclusively with r_m, which translates into a higher continuation probability in her game with r_m, which she can signal by holding on the identity I_{g_m}.

Second, by again assuming $\delta_i = \delta_m^h = \delta_j^h$, we know that the sender is expected to adopt the identity I_{g_j} associated to other social niche members if the benefits generated from cooperating exclusively with r_j are (i) greater than her benefits from cooperating with r_m, and (ii) greater than the value of her outside option. This leads to the following two conditions:

(i) $k_j \geq k_m$,

(ii) $k_j \geq \delta_i c$.

Condition (i) realizes if the benefit that the sender reaps from cooperating with social niche members is greater than the benefits that she reaps from cooperating with home community members ($k_j > k_m$). Condition (ii) necessarily realizes at a cooperative equilibrium. Therefore, if condition (i) and (ii) are satisfied, we set $\delta_j = \delta_j^h > \delta_m = \delta_m^l$, given that the benefits from cooperating exclusively with r_j are greater than the benefits from cooperating with r_m. By the same reasoning as before, the greater incentives to cooperate with r_j translate into a higher continuation probability in game j, which she can signal by adopting the identity I_{g_j}.[5]

[5] $\delta_m = \delta_m^l$ and $\delta_j = \delta_j^l$ can not simultaneously realize, given that we have assumed that the sender can always *decide* to cooperate either with r_m or r_j, without constraints. This implies that the gains from cooperating with r_m or r_j can always realize, and these gains are always greater than the gains from defecting. The adoption of a social identity I_{g_i} is therefore always a truthful signal of high continuation probability in the present setup.

To summarize, if the benefits from cooperating with r_m are greater than the benefits from cooperating with r_j, the sender s is expected to adopt the social identity I_{g_m}, which (truthfully) signals high continuation probability in her repeated interaction with r_m, and therefore can stabilize cooperation between the two players. On the other hand, if the benefits from cooperating with r_j are greater than the benefits from cooperating with r_m, the sender s is expected to adopt the social identity I_{g_j}, which similarly (truthfully) signals high continuation probability in her repeated interaction with r_j. In the present framework, adopting a given social identity therefore signals high continuation probability in the repeated APD, and can convince receivers to cooperate with the sender.

Niche Selection. In the first stage, the sender s is expected to choose the social niche whose ideal trait profile is the closest (in terms of distance) from her trait profile.[6] This first choice is therefore devoid of any strategic consideration. One can write the (Euclidean) distance d_{sn} between the sender s's trait profile and niche j's ideal trait profile in the following way:

$$d_{sj} = \sqrt{\sum_{p=1}^{P} (\theta_{sp} - \gamma_{jp})^2}$$

Let $\bar{v} : J \to \mathbb{R}_+$ be a function which gives, for each niche $j \in J$, a prospective value $\bar{v}(j) \in \mathbb{R}_+$ to the sender. More specifically, let $\bar{v}(j) = \frac{1}{d_{sj}}$.[7] Given that the distance d_{sj} between the sender's trait profile and the niche's ideal trait profile is fixed and given, the sender is expected, *ex ante*, to choose the niche j^* which satisfies $\underset{j \in J}{max}\, \bar{v}(j)$. As a matter of example, one can imagine that a creative, conscientious and curious individual will be particularly fit to do scientific research, that a musically gifted individual might stand out in an orchestra, or that a natural analytical thinker might excel at chess.[8]

Equilibrium Specification

The following Propositions describe the conditions underlying our two main equilibrium strategy profiles of interest, which characterize the circumstances under which the sender s will be willing to adopt social identity I_{g_m} (Proposition 0.1) or I_{g_j} (Proposition 0.2). The sender s is considered to be Player 1, the receiver r_m Player 2, and the receiver r_j Player

[6] In reality, individuals belong to different social niches, and are therefore exposed to a potentially wide variety of audiences. As a matter of example, individuals can simultaneously belong to a professional occupation, a book club, a sports club, a musical band and/or an online gaming community. While the formalization of this more realistic state of affairs (with n receivers instead of two) adds complexity, I expect the model's main predictions to remain the same.

[7] We assume that d_{sj} never takes a value of 0.

[8] For evidence that personality traits influence occupational choice, see Cobb-Clark and Tan (2011), De Fruyt and Mervielde (1999), Wells, Ham, and Junankar (2016); for evidence that personality traits influence the activities that individuals indulge in, see Carlo, Okun, Knight, and de Guzman (2005) or Ickes, Snyder, and Garcia (1997). For evidence that a fit between individual traits and the niche's ideal trait profile is beneficial, see Denissen, Bleidorn, Hennecke, Luhmann, Orth, Specht, and Zimmermann (2018).

3. A strategy for the sender in this game must specify (i) which social niche j she decides to join, (ii) which signal to send (or, alternatively, which social identity to adopt) in the signaling stage (either I_{g_m} or I_{g_j}), and (iii) whether to play ALLC or ALLD in her game with r_m and r_j (e.g., $ALLC_mALLD_j$, implying that the sender plays ALLC with r_m but ALLD with r_j, and written C_mD_j for convenience). An example of a strategy profile for the sender would be $\{j^*, I_{g_m}, C_mD_j\}$, where the sender would choose the social niche j^*, adopt the identity I_{g_m}, and cooperate with r_m but defect with r_j. A strategy for r_i in this game must specify whether to play ALLC or ALLD as a function of the sender's decision in the signaling stage. An example of a strategy for the receiver would be $ALLD_mALLD_j$ (written D_mD_j for convenience), where the sender would defect no matter what signal has been sent by the sender.

Proposition 0.1. *The strategy profile* $\{\{j^*, I_{g_m}, C_mD_j\}, C_mD_j, D_mC_j\}$ *is a SPNE of the game if the following conditions are satisfied:*

(i) $j^* \in \underset{j \in J}{max}\, \bar{v}(j)$,

(ii) $k_m \geq k_j$,

(iii) $k_m \geq c$,

(iv) $\delta_m^h \geq \frac{c}{b}$.

At this equilibrium strategy profile, the sender adopts the identity I_{g_m} of receiver r_m, and plays ALLC (cooperates) only with r_m. Receiver m cooperates with the sender s if and only if the sender adopts I_{g_m}, while receiver j refuses to invest in a relationship with the sender if the sender adopts I_{g_j}.

Proposition 0.2. *The strategy profile* $\{\{j^*, I_{g_j}, D_mC_j\}, C_mD_j, D_mC_j\}$ *is a SPNE of this game if the following conditions are satisfied:*

(i) $j^* \in \underset{j \in J}{max}\, \bar{v}(j)$,

(ii) $k_j \geq k_m$,

(iii) $k_j \geq c$,

(iv) $\delta_j^h \geq \frac{c}{b}$.

At this equilibrium strategy profile, the sender adopts the identity I_{g_j} of receiver r_j, and plays ALLC (cooperates) only with r_j. Receiver j cooperates with the sender s if and only if the sender adopts I_{g_j}, while receiver m refuses to invest in a relationship with the sender if the sender adopts I_{g_m}.

Discussion of the Main Predictions

This section is dedicated to a discussion of the main predictions of the model developed in the previous section. Some of these predictions will be confronted to empirical data in the last part of the paper in order to determine whether the model can provide new, interesting insights regarding the adoption of relevant aspects of an individual's social identity.

1. The first main prediction is that social identity will follow social and material incentives. This means that if (social and material) incentives in an individual's environment remain stable, then her social identity is expected to remain stable too. On the other hand, a change in incentives (e.g., when one joins a novel environment, or when one's social group modifies its core tenets) is expected to produce a change in the individual's social identity. The argument exposed in this paper is that social incentives principally take the form of long-term mutually beneficial relationships with other members of an individual's *community*. Therefore, if (i) an individual's community remains stable across time, and (ii) the beliefs, ideologies and values adopted by community members remain stable, then the individual's social identity is expected to remain stable too. On the other hand, if (i) an individual's community changes, and/or (ii) the beliefs, ideologies and values adopted by community members change, then the individual's social identity is expected to change too.

2. The second main prediction is that individuals will (often) want to hold on, and defend their social identity. Given that social identity can serve as a signal of trustworthiness (or intention to cooperate), it is expected that individuals will be eager to make their social commitments public in order not to alienate other members of their community, especially in contexts in which the relevant aspects of social identity become particularly salient (such as, for instance, in a polarized environment). In fact, if some beliefs and values have been adopted solely for their signaling value, then one does not expect individuals to modify their social identity when new (potentially conflicting) evidence arrives.

3. The third main prediction is that individuals are expected to balance the benefits from cooperating with different audiences (or communities) when deciding which social identity to adopt.

4. The fourth main prediction is that individual level traits (personality and cognition) will be correlated with specific beliefs and values. To see this, assume that there is only one relevant individual trait in the sender's trait profile θ_s, p. That is, let $\theta_s = (p)$, with p taking one of two values: either $p = \underline{p}$, implying that the sender has a low value for trait p, or $p = \bar{p}$, implying that the sender has a high value for trait p. Moreover, assume that there are only two social niches to join, q and t. Social niche q's ideal trait profile is $\gamma_q = (\gamma_{q_p})$, and social niche t's ideal trait profile is $\gamma_t = (\gamma_{t_p})$. Assume

that $\gamma_{q_p} = \underline{p}$ and $\gamma_{t_p} = \bar{p}$. Let the social group g_q be primarily represented among the members of q, while let the social group g_t be primarily represented among the members of t. The sender s with trait profile θ_s will choose the niche i providing her with the highest prospective value $\bar{v}(j)$, which is, by definition, the niche i that satisfies $\underset{i \in \{q,t\}}{min}\, d_{si}$. Therefore, if s is endowed with $\theta_s = \underline{p}$, then s will choose to join q. Alternatively, if s is endowed with $\theta_s = \bar{p}$, then s will choose to join t. Given that members of q have primarily adopted the identity I_{g_q}, while members of t have primarily adopted I_{g_t}, a correlation will inevitably arise between individual traits (here, p) and social identities, defined as packages of beliefs and values (here, I_{g_q} and I_{g_t}). At the aggregate level, then, we should observe a correlation between specific individual-level traits (e.g., personality traits) and specific beliefs and values.

Empirical Evidence

This section is dedicated at confronting the main predictions of the model with existing empirical evidence.

Do Beliefs and Values Respond to (Social and Material) Incentives?

The theory developed in this paper predicts that social identities will follow social and material incentives, principally defined here as mutually beneficial long-term relationships. Therefore, we should observe, in the data, changes in relevant beliefs and values following changes in relevant incentives. There exists significant evidence that individual political (Goren, 2005, Goren, Federico, and Kittilson, 2009, McCann, 1997) and moral (Smith, Alford, Hibbing, Martin, and Hatemi, 2017, Hatemi, Crabtree, and Smith, 2019) values change over time. In fact, the very concept of a *core* value as an internal predisposition has recently been questioned (Connors, 2019). Can we trace, in the data, changes in beliefs and values to changes in incentives?

One strand of evidence comes from panel studies documenting changes in individual beliefs and values to changes in the core tenets of the social group one belongs to (or identifies with). For instance, using quasi-experimental settings, it has been shown that political party members very often realign their views with those of elite party members following changes in elite opinions (Barber and Pope, 2019, Slothuus and Bisgaard, 2021). Also, Gould and Klor (2019) have shown, using a long-run panel study, that changes in individual political beliefs and values closely track changes in the core tenets of the political party individuals identify with. Finally, Egan (2020) has shown that individuals shift their identities following congruent shifts among other members of their political coalitions, while Agadjanian and Lacy (2021), in a related paper, show that individual racial identities converge towards the identity enforced in their political party. A second strand of evidence comes from longitudi-

nal studies linking changes in individual beliefs and values to changes in communities (e.g., following changes in neighborhood, occupation, studies, etc.). For instance, Sinclair (2012) and Martin and Webster (2020) have found that individuals tend to change their party affiliation to match that of other community members when moving to new neighborhoods. Other studies based on field experiments (Levitan and Visser, 2009) or longitudinal research (Lazer, Rubineau, Chetkovich, Katz, and Neblo, 2010, Mayrl and Uecker, 2011) have shown that once individuals move to a new community, composed of members with dissimilar beliefs and values, significant social influence tends to happen, with individuals very often shifting their beliefs and values towards those predominantly held by other members of the community. A third strand of evidence comes from laboratory experiments showing that individual political values usually respond to social cues, with individuals often modifying their expressed identities when interacting with others holding dissimilar views, when encountering an ideologically homogeneous audience, or when receiving information about peer preferences and values (Connors, 2019, Klar, 2014, Levitan and Verhulst, 2016, Mallinson and Hatemi, 2018, Toff and Suhay, 2019, Visser and Mirabile, 2004). These results are consistent with the idea that (relevant) individual beliefs and values are socially enforced, with individuals eager to display their social commitments by updating their identity as a function of the beliefs and values held by other members of their community.

Finally, a fourth, related strand of evidence, links changes in social identities to changes in material incentives. For instance, Cassan (2015) describes how individuals in India have manipulated their caste identity in order to benefit from land reforms in the beginning of the twentieth century. Also, Green (2021) shows that individuals in Sub-Saharan Africa and China tend to switch their ethnic identities in order to reap benefits from patronage. Finally, Antman and Duncan (2015) describe how racial identification in the U.S. is sensitive to potential benefits from affirmative action policies. These findings support the idea that expressed social identities are often tied to incentives, whether material or social. Hence, the argument is that we can garner important insights about the social identities individuals decide to adopt (and express) by understanding the relevant (social and material) incentives that they face.

Are Relevant Beliefs Responsive to Information?

According to the *biased informational exposure* model (Huckfeldt, 2001), members of a given community will share the same beliefs *because* they have access to the same informational sources. This model makes two main predictions: (i) less sophisticated or less informed individuals should be the most responsive, and (ii) access to new informational sources should influence individual beliefs. By contrast, the model developed in this paper predicts that the relevant aspects of social identity principally respond to social incentives; therefore, more or better information is not expected to affect relevant beliefs. Rather, people are expected to actively resist changing their social identity as a way to maintain their social commitments on display.

Several recent studies have found that the same information, provided or supported by different individuals or entities, has different effects upon the receiver. For instance, Druckman, Peterson, and Slothuus (2013) find that subjects reject arguments that they consider strong when they are supported by members of another party. This has been replicated by Bolsen, Druckman, and Cook (2014), who showed that subjects become less supportive of a given policy when members of the opposite party endorse it. Importantly, these effects only arise when subjects are embedded in a polarized environment, suggesting that it is only when cued with their social commitments that subjects distort the way they process and evaluate incoming information, as predicted by the model developed in this paper. Even more telling are findings that individuals do not reduce their support for a candidate after learning (and accepting) that they have been told lies (Nyhan, Porter, Reifler, and Wood, 2019, Swire-Thompson, Ecker, Lewandowsky, and Berinsky, 2020). This tendency to treat information in a biased, *directed* way, has been said to represent evidence of *motivated reasoning* in the political sphere (Flynn, Nyhan, and Reifler, 2017).

Yet, rather than being evidence of motivated reasoning, this *source* effect might stem from the fact that individuals find different sources more or less credible (Druckman and McGrath, 2019, Tappin, Pennycook, and Rand, 2020). Therefore, these findings might not reflect evidence of motivated reasoning, wherein people discard (or argue against) evidence not compatible with their prior beliefs, but might actually stem from accuracy-motivated individuals finding different informational sources more or less credible. Tappin, Pennycook, and Rand (2020, p.85, emphasis in original) conclude that "paradigmatic designs often fall short in identifying a *particular* motivation [...] as causing reasoning, as opposed to other motivations, such as that for accuracy".

Disentangling the effect of trust from the one of motivated reasoning is in fact a complicated task. Nevertheless, there exist studies that might tip the balance toward the motivated reasoning story. For instance, Kahan, Peters, Dawson, and Slovic (2017), in a problem-solving environment that did not include information transmission, find evidence that the most sophisticated individuals (those high in numeracy) are the most polarized when it comes to solving a politically charged problem. Therefore, doing away with the issue of source credibility, they find that individuals use their reasoning in order to arrive at identity-congruent beliefs. Even more interesting, in an environment in which other forms of (accuracy-oriented) updating have been controlled for, Thaler (2021) finds that individuals *attribute trust* instrumentally: they decide to trust (or distrust) an information source depending on whether the information provided is congenial (or not) with already held beliefs, particularly so when issues are politicized. Moreover, if issues of trust and credibility were driving the source effect, then we might reasonably expect that those individuals that are the less sophisticated and the less informed would be more likely to rely on source cues when updating their beliefs and attitudes. In fact, the contrary has been found, with individuals having the greatest cognitive resources often being *more* likely to resort to partisan cues (Bakker, Lelkes, and Malka, 2020). Finally, Frimer, Skitka, and Motyl (2017) find that subjects *consciously* decide to avoid hearing non-congruent opinions by fear of undermining valuable relationships; this

is exactly what is predicted in the model developed in this paper.

Is There A Link Between Individual Traits and Beliefs and Values?

A large literature, at the intersection of behavioral genetics and political science, has shown that there exist significant links between individual-level traits and specific beliefs, ideologies and values (Dawes and Weinschenk, 2020). In particular, personality traits have been shown to be correlated with political beliefs and ideologies (Gerber, Huber, Doherty, and Dowling, 2011), while twin studies have described how ideology is partly heritable (Alford, Funk, and Hibbing, 2005, Hatemi, Hibbing, Medland, Keller, Alford, Smith, Martin, and Eaves, 2010).

The model developed in this paper helps explain some empirical findings that are not easily reconcilable with the predominant idea that dispositional traits directly influence the adoption of specific beliefs and values (Funk, Smith, Alford, Hibbing, Eaton, Krueger, Eaves, and Hibbing, 2013). For instance, it helps to explain the finding that "genetic influence [on political attitudes] is manifest only after moving away from the parental home" (Hatemi, Funk, Medland, Maes, Silberg, Martin, and Eaves, 2009, p.1153). If dispositional traits causally influence the adoption of specific beliefs and ideologies, then this effect is expected to be observed from development on. Yet, the observed statistical relationship between genes/psychology and beliefs and ideologies only emerges once individuals leave their home community (Hatemi, Funk, Medland, Maes, Silberg, Martin, and Eaves, 2009, Hufer, Kornadt, Kandler, and Riemann, 2020). This is predicted by the model developed in this paper. During development, psychologically and genetically dissimilar individuals are embedded into families and communities that strongly influence their beliefs and values. This will translate into the *shared environment* explaining most of the variance in beliefs and values before entering into adulthood. But once individuals have left their home, they (often) need not hold on the beliefs of their parents, family or previous community anymore, and depending on the environment (social niche or *foci*) they decide to join (which will be a function of their individual-level traits), they will congregate with genetically and psychologically similar others, and come to adopt new beliefs and ideologies that are associated with the social groups other community members belong to, therefore explaining the finding that the correlation between genes and political attitudes only emerges once individuals have left their home.

Moreover, Gerber, Huber, Doherty, Dowling, and Ha (2010) find that the widely replicated relationship between Conscientiousness and conservative ideology, and between Openness to Experience and liberal ideology, disappears when they restrict their analysis to black Americans. While they interpret this result as suggesting that ideological labels and economic policy have different *meanings* for black Americans, it is perfectly consistent with the idea that due to constraints on the environments they can join, they have not the opportunity to sort themselves according to their own abilities, personalities, and other individual-traits. As a result, individuals with *dissimilar* traits will find themselves in the *same* environments (niches or *foci*), and will likely adopt the same beliefs and ideologies, thereby weakening the

relationship between individual-traits and specific political attitudes.

In a recent review, Dawes and Weinschenk (2020) discuss how researchers are trying to uncover the specific mechanisms and pathways linking individual-level traits (genes or psychological mechanisms) to beliefs and ideologies. The present model suggests that instead of focusing exclusively on individual-level traits as mediators, interesting insights can also be garnered by investigating how similar (in terms of genes, personality, psychological mechanisms, cognition, etc.) individuals might join similar environments—as first described by Feld (1981, 1982)—and as a consequence be exposed to similar incentives to adopt similar beliefs, ideologies and values.

Conclusion

Social identity, defined in this paper as the adoption of (or affiliation with) the core beliefs, ideologies and values of a social group, has been shown to play an integral role in people's lives. While researchers often stress the psychological benefits that individuals derive from their perceived group memberships, the objective of the present paper was to characterize the incentives that underlie the choice and expression of social identity. More precisely, the present paper has proposed a theory of social identity adoption and expression which ties the choice of social identity to material and social benefits present in an individual's social environment, and which aims to explain why individuals are often so eager to make their social identity known to others.

The model presented in this paper makes several predictions, which help illuminate several empirical findings. First, it helps to explain why individual beliefs and values can be so malleable, depending on the social context in which individuals find themselves. Second, it helps to explain why individuals are often emotionally invested in their social identity, and why beliefs and values can become resistant to evidence. Finally, it helps to explain why beliefs and values can come to be correlated with individual-level traits such as personality. Overall, this paper makes the case that to better understand the social identity that individuals decide to adopt and express, it is important to understand the social incentives that they face.

References

Agadjanian, A. and D. Lacy (2021): "Changing votes, changing identities? Racial fluidity and vote switching in the 2012–2016 US Presidential Elections," *Public Opinion Quarterly*, 85, 737–752.

Akerlof, G. A. and R. E. Kranton (2000): "Economics and identity," *The quarterly journal of economics*, 115, 715–753.

Akerlof, R. (2016): """ We Thinking""" and Its Consequences," *American Economic Review*, 106, 415–19.

Alford, J. R., C. L. Funk, and J. R. Hibbing (2005): "Are political orientations genetically transmitted?" *American political science review*, 99, 153–167.

Antman, F. and B. Duncan (2015): "Incentives to identify: Racial identity in the age of affirmative action," *Review of Economics and Statistics*, 97, 710–713.

Bakker, B. N., Y. Lelkes, and A. Malka (2020): "Understanding partisan cue receptivity: Tests of predictions from the bounded rationality and expressive utility perspectives," *The Journal of Politics*, 82, 1061–1077.

Barber, M. and J. C. Pope (2019): "Does party trump ideology? Disentangling party and ideology in America," *American Political Science Review*, 113, 38–54.

Bénabou, R. and J. Tirole (2011): "Identity, morals, and taboos: Beliefs as assets," *The Quarterly Journal of Economics*, 126, 805–855.

Bolsen, T., J. N. Druckman, and F. L. Cook (2014): "The influence of partisan motivated reasoning on public opinion," *Political Behavior*, 36, 235–262.

Bonica, A. (2014): "Mapping the ideological marketplace," *American Journal of Political Science*, 58, 367–386.

Brewer, M. B. (1991): "The social self: On being the same and different at the same time," *Personality and Social Psychology Bulletin*, 17, 475–482.

Carlo, G., M. A. Okun, G. P. Knight, and M. R. T. de Guzman (2005): "The interplay of traits and motives on volunteering: Agreeableness, extraversion and prosocial value motivation," *Personality and Individual Differences*, 38, 1293–1305.

Carvalho, J.-P. (2016): "Identity-based organizations," *American Economic Review*, 106, 410–14.

Cassan, G. (2015): "Identity-based policies and identity manipulation: Evidence from colonial Punjab," *American Economic Journal: Economic Policy*, 7, 103–31.

Cobb-Clark, D. A. and M. Tan (2011): "Noncognitive skills, occupational attainment, and relative wages," *Labour Economics*, 18, 1–13.

Connors, E. C. (2019): "The social dimension of political values," *Political Behavior*, 1–22.

Dawes, C. T. and A. C. Weinschenk (2020): "On the genetic basis of political orientation," *Current Opinion in Behavioral Sciences*, 34, 173–178.

De Fruyt, F. and I. Mervielde (1999): "RIASEC types and Big Five traits as predictors of employment status and nature of employment," *Personnel psychology*, 52, 701–727.

Denissen, J. J., W. Bleidorn, M. Hennecke, M. Luhmann, U. Orth, J. Specht, and J. Zimmermann (2018): "Uncovering the power of personality to shape income," *Psychological Science*, 29, 3–13.

Druckman, J. N. and M. C. McGrath (2019): "The evidence for motivated reasoning in climate

change preference formation," *Nature Climate Change*, 9, 111–119.

Druckman, J. N., E. Peterson, and R. Slothuus (2013): "How elite partisan polarization affects public opinion formation," *American Political Science Review*, 107, 57–79.

Egan, P. J. (2020): "Identity as dependent variable: How Americans shift their identities to align with their politics," *American Journal of Political Science*, 64, 699–716.

Feld, S. L. (1981): "The focused organization of social ties," *American journal of sociology*, 86, 1015–1035.

——— (1982): "Social structural determinants of similarity among associates," *American Sociological Review*, 797–801.

——— (1997): "Structural embeddedness and stability of interpersonal relations," *Social networks*, 19, 91–95.

Flynn, D., B. Nyhan, and J. Reifler (2017): "The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics," *Political Psychology*, 38, 127–150.

Frimer, J. A., L. J. Skitka, and M. Motyl (2017): "Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions," *Journal of Experimental Social Psychology*, 72, 1–12.

Funk, C. L., K. B. Smith, J. R. Alford, M. V. Hibbing, N. R. Eaton, R. F. Krueger, L. J. Eaves, and J. R. Hibbing (2013): "Genetic and environmental transmission of political orientations," *Political Psychology*, 34, 805–819.

Gerber, A. S., G. A. Huber, D. Doherty, and C. M. Dowling (2011): "The Big Five personality traits in the political arena," *Annual Review of Political Science*, 14, 265–287.

Gerber, A. S., G. A. Huber, D. Doherty, C. M. Dowling, and S. E. Ha (2010): "Personality and political attitudes: Relationships across issue domains and political contexts," *American Political Science Review*, 104, 111–133.

Goren, P. (2005): "Party identification and core political values," *American Journal of Political Science*, 49, 881–896.

Goren, P., C. M. Federico, and M. C. Kittilson (2009): "Source cues, partisan identities, and political value expression," *American Journal of Political Science*, 53, 805–820.

Gould, E. D. and E. F. Klor (2019): "Party hacks and true believers: The effect of party affiliation on political preferences," *Journal of Comparative Economics*, 47, 504–524.

Green, E. (2021): "The politics of ethnic identity in Sub-Saharan Africa," *Comparative Political Studies*, 54, 1197–1226.

Hatemi, P. K., C. Crabtree, and K. B. Smith (2019): "Ideology justifies morality: Political beliefs predict moral foundations," *American Journal of Political Science*, 63, 788–806.

Hatemi, P. K., C. L. Funk, S. E. Medland, H. M. Maes, J. L. Silberg, N. G. Martin, and L. J. Eaves (2009): "Genetic and environmental transmission of political attitudes over a life time," *The Journal of Politics*, 71, 1141–1156.

Hatemi, P. K., J. R. Hibbing, S. E. Medland, M. C. Keller, J. R. Alford, K. B. Smith, N. G. Martin, and L. J. Eaves (2010): "Not by twins alone: Using the extended family design to investigate genetic influence on political beliefs," *American journal of political science*, 54, 798–814.

Huckfeldt, R. (2001): "The social communication of political expertise," *American Journal of Political Science*, 425–438.

Huckfeldt, R., J. J. Mondak, M. Hayes, M. T. Pietryka, and J. Reilly (2013): "Networks, interdependence, and social influence in politics," in *The Oxford Handbook of Political Psychology*, Oxford University Press.

Hufer, A., A. E. Kornadt, C. Kandler, and R. Riemann (2020): "Genetic and environmental variation in political orientation in adolescence and early adulthood: A Nuclear Twin Family analysis." *Journal of personality and social psychology*, 118, 762.

Ickes, W., M. Snyder, and S. Garcia (1997): "Personality influences on the choice of situations," in *Handbook of Personality Psychology*, Elsevier, 165–195.

Jacoby, W. G. (2014): "Is there a culture war? Conflicting value structures in American public opinion," *American Political Science Review*, 108, 754–771.

Jordan, J. J., M. Hoffman, P. Bloom, and D. G. Rand (2016): "Third-party punishment as a costly signal of trustworthiness," *Nature*, 530, 473–476.

Jordan, J. J. and D. G. Rand (2017): "Third-party punishment as a costly signal of high continuation probabilities in repeated games," *Journal of Theoretical Biology*, 421, 189–202.

Kahan, D. M., E. Peters, E. C. Dawson, and P. Slovic (2017): "Motivated numeracy and enlightened self-government," *Behavioural Public Policy*, 1, 54–86.

Klar, S. (2014): "Partisanship in a social setting," *American Journal of Political Science*, 58, 687–704.

Lazer, D., B. Rubineau, C. Chetkovich, N. Katz, and M. Neblo (2010): "The coevolution of networks and political attitudes," *Political Communication*, 27, 248–274.

Lemyre, L. and P. M. Smith (1985): "Intergroup discrimination and self-esteem in the minimal group paradigm." *Journal of Personality and Social Psychology*, 49, 660.

Levitan, L. C. and B. Verhulst (2016): "Conformity in groups: The effects of others' views on expressed attitudes and attitude change," *Political Behavior*, 38, 277–315.

Levitan, L. C. and P. S. Visser (2009): "Social network composition and attitude strength: Exploring the dynamics within newly formed social networks," *Journal of Experimental Social Psychology*, 45, 1057–1067.

Loury, G. C. (1994): "Self-censorship in public discourse: A theory of "political correctness" and related phenomena," *Rationality and Society*, 6, 428–461.

Mallinson, D. J. and P. K. Hatemi (2018): "The effects of information and social conformity on opinion change," *PloS one*, 13, e0196600.

Martin, G. J. and S. W. Webster (2020): "Does residential sorting explain geographic polarization?" *Political Science Research and Methods*, 8, 215–231.

Mayrl, D. and J. E. Uecker (2011): "Higher education and religious liberalization among young adults," *Social Forces*, 90, 181–208.

McCann, J. A. (1997): "Electoral choices and core value change: The 1992 presidential campaign," *American Journal of Political Science*, 564–583.

Melnikoff, D. E. and N. Strohminger (2020): "The automatic influence of advocacy on lawyers and novices," *Nature Human Behaviour*, 1–7.

Mitchell, K. J. (2020): *Innate: How the wiring of our brains shapes who we are*, Princeton University Press.

Nyhan, B., E. Porter, J. Reifler, and T. J. Wood (2019): "Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorabil-

ity," *Political Behavior*, 1–22.

Oakes, P. J. and J. C. Turner (1980): "Social categorization and intergroup behaviour: Does minimal intergroup discrimination make social identity more positive?" *European Journal of Social Psychology*, 295–301.

Park, J. H. and F. van Leeuwen (2015): "Evolutionary perspectives on social identity," in *Evolutionary Perspectives on Social Psychology*, Springer, 115–125.

Pietraszewski, D. (2020): "Intergroup processes: Principles from an evolutionary perspective," *Social Psychology: Handbook of Basic Principles*, 373–391.

Pietraszewski, D., O. S. Curry, M. B. Petersen, L. Cosmides, and J. Tooby (2015): "Constituents of political cognition: Race, party politics, and the alliance detection system," *Cognition*, 140, 24–39.

Schwardmann, P., E. Tripodi, and J. J. van der Weele (Forthcoming): "Self-Persuasion: Evidence from Field Experiments at International Debating Competitions," *American Economic Review*.

Shayo, M. (2009): "A model of social identity with an application to political economy: Nation, class, and redistribution," *American Political science review*, 103, 147–174.

Sinclair, B. (2012): *The social citizen: Peer networks and political behavior*, University of Chicago Press.

Slothuus, R. and M. Bisgaard (2021): "How political parties shape public opinion in the real world," *American Journal of Political Science*, 65, 896–911.

Smaldino, P. E. (2019): "Social identity and cooperation in cultural evolution," *Behavioural Processes*, 161, 108–116.

Smaldino, P. E., A. Lukaszewski, C. von Rueden, and M. Gurven (2019): "Niche diversity can explain cross-cultural differences in personality structure," *Nature Human Behaviour*, 3, 1276–1283.

Smaldino, P. E. and M. A. Turner (2021): "Covert signaling is an adaptive communication strategy in diverse populations." *Psychological Review*.

Smith, K. B., J. R. Alford, J. R. Hibbing, N. G. Martin, and P. K. Hatemi (2017): "Intuitive ethics and political orientations: Testing moral foundations as a theory of political ideology," *American Journal of Political Science*, 61, 424–437.

Swire-Thompson, B., U. K. Ecker, S. Lewandowsky, and A. J. Berinsky (2020): "They might be a liar but theyâ€™re my liar: Source evaluation and the prevalence of misinformation," *Political Psychology*, 41, 21–34.

Tajfel, H. (1974): "Social identity and Intergroup Behaviour," *Social Sciences Information*, 13, 65–93.

Tajfel, H. and J. C. Turner (1979): "An integrative theory of intergroup conflict," *The Social Psychology of Intergroup Relations*.

Tappin, B. M., G. Pennycook, and D. G. Rand (2020): "Thinking clearly about causal inferences of politically motivated reasoning: Why paradigmatic study designs often undermine causal inference," *Current Opinion in Behavioral Sciences*, 34, 81–87.

Thaler, M. (2021): "The Fake News Effect: Experimentally Identifying Motivated Reasoning Using Trust in News," *Available at SSRN 3717381*.

Toff, B. and E. Suhay (2019): "Partisan conformity, social identity, and the formation of policy

preferences," *International Journal of Public Opinion Research*, 31, 349–367.

Van Bavel, J. J. and A. Pereira (2018): "The partisan brain: An identity-based model of political belief," *Trends in Cognitive Sciences*, 22, 213–224.

Visser, P. S. and R. R. Mirabile (2004): "Attitudes in the social context: The impact of social network composition on individual-level attitude strength." *Journal of Personality and Social Psychology*, 87, 779.

Wells, R., R. Ham, and P. N. Junankar (2016): "An examination of personality in occupational outcomes: antagonistic managers, careless workers and extraverted salespeople," *Applied Economics*, 48, 636–651.

Williams, D. (2021): "Signalling, commitment, and strategic absurdities," *Mind & Language*.

Yamagishi, T., N. Jin, and T. Kiyonari (1999): "Bounded generalized reciprocity: Ingroup boasting and ingroup favoritism," *Advances in group processes*, 16, 161–197.

Yamagishi, T. and T. Kiyonari (2000): "The group as the container of generalized reciprocity," *Social Psychology Quarterly*, 116–132.