Affective Decision-Making and Moral Sentiments[*]

Erik O. Kimbrough[†] **Alexander Vostroknutov**[‡][§]

February 10, 2022

Abstract

We propose a new modeling framework to study affective decision-making, which produces many notorious ”irrationalities” in human behavior. Building on biologically-inspired models of reinforcement learning, we provide a description of a boundedly-rational affective agent who holds mood-dependent beliefs and exhibits prospect-theory-like behaviors in situations with uncertainty. By construction, affective agents possess personal and social identities and desire to achieve higher social status, which allows them to cooperate with other like-minded agents. We show how moral sentiments, indicative of adherence to identity-based norms, emerge to strengthen cooperation even further. The model of affective decision-making and moral sentiments complements our previous work on moral reasoning among rational agents (“A Theory of Injunctive Norms”). Together, the two models constitute an attempt at a comprehensive framework to study norm-driven human behavior.

JEL classifications: D01, D81, D91
Keywords: evolution, bounded rationality, reinforcement learning, identity, status, social norms

[*] We would like to thank Alexandros Karakostas and the participants of the seminars at Maastricht University, University of Queensland, Vrije Universiteit Amsterdam, IMT Lucca for helpful comments and suggestions. All mistakes are our own.

[†] Smith Institute for Political Economy and Philosophy, Chapman University, One University Drive, Orange, CA 92866, USA. email: ekimbrou@chapman.edu.

[‡] Department of Economics (MPE), Maastricht University, Tongersestraat 53, 6211LM Maastricht, The Netherlands. e-mail: a.vostroknutov@maastrichtuniversity.nl.

[§] Corresponding author.

> *"The man who is unhappy will, as a rule, adopt an unhappy creed, while the man who is happy will adopt a happy creed; each may attribute his happiness or unhappiness to his beliefs, while the real causation is the other way round."*
>
> ∼ Bertrand Russell, The Conquest of Happiness (1930), p. 89

1 Introduction

Under a hedonic interpretation of utility theory, our decisions are driven by how they *will* make us feel. Utility maximization is prospective such that decisions depend only on the (expected) consequences of available choices. However, introspection alone is enough to reveal that our decisions also depend on how we *do* feel, in the moment. A person's evaluation of the same consequence can fluctuate over time depending on his recent experiences, his mood, his relationships, and etc. For example, A's decision whether to eat at a particular restaurant may depend on whether he spilled a drink last time he was there, whether he's in the mood for a quiet meal, or whether he might run into someone he's trying to avoid. None of these factors directly affect the deliciousness of the restaurant's taco salad, but they can nevertheless color how it is evaluated as a prospect.

In short, our choices depend on our *affect*. In this paper, we introduce a model of affective decision-making that is designed to capture the process whereby our affect is influenced by the choices we've made (and outcomes we've experienced) in the past and, by coloring our evaluation of future prospects, also shapes the choices we make about the future. We ground our model in insights from evolutionary biology and neuroscience, building on biology-inspired models of reinforcement learning. Affective decision-makers are modeled as boundedly rational reinforcement learners who, via experience, develop and update beliefs about the affective value of various features of their environment. Their choices then depend not only on consequences but also on the cumulative affective values of environmental features in which those consequences are embedded.

Strikingly, our simple model reproduces many notorious "irrationalities" identified by behavioral economists in individual choice problems such as over- or under-confidence, motivated reasoning, procrastination, addiction, and etc. Moreover, our model provides an intuitive account of "moral sentiments" in which an agent's willingness to cooperate with, trust, punish, or otherwise interact with others depends on the affective value the agent assigns to their observable features. That value is based on past experience with them and with others who share those features. Agents in our model prefer to interact with those who have similar experiences and who "see the world the same way" in the sense of having similar beliefs about the affective value of environmental features. Thus, our reinforcement-learning agents develop something that closely resembles notions of shared identity.

In the spirit of Arthur J. Robson, we root our model in biology to address concerns about the proliferation of *ad hoc* behavioral theories of decision-making. Like Robson (2002) we accept the large body of evidence that challenges the standard economic view of rational decision-making, and we share his view that accepting the "biological evolution of preferences, beliefs, and rationality" can facilitate a "unified treatment" of this evidence. Employing an "evolutionary basis helps to maintain constraints on economic theory" and also to maintain the key advantage of economics as a social science, namely that of "being based on overarching theory."

In this paper we thus propose a new idea about how beliefs, preferences, and some notion of rationality can be theoretically grounded in evolution and biology. While Robson delved mostly into the "preferences" part of this trio, we suggest a model of a boundedly-rational *affective agent* whose beliefs, preferences, and optimization behavior are all strongly influenced by specific biological constraints. It is important to note that this framework is not an attempt at replacing the standard rational economic actor with some modified version that is "less rational" in some sense. Rather, we embrace the dichotomy, traceable at least back to Plato, between emotion and reason, which naturally distinguishes the affective and the rational component of human decision-making.

In this view, choices can be understood as resulting from a mixture of the activities of two decision makers, one rational and another affective (somewhat reminiscent of dual-self models, e.g. Fudenberg and Levine, 2006). This paper focuses on the latter. Thus, the model presented here is supposed to complement rational framework, with the affective agent, as one of the dual selves, being responsible for deviations from rationality in human behavior.

We make two assumptions about the biological constraints that define how the affective agent functions. The first one outlines the scope of the affective agent's world and how he makes choices in it. Here, we assume that evolution is a "stingy engineer," who first attempted to design a (human) decision maker from the elements that were already present in the standard mammalian brain architecture (before rationality appeared on the stage). This implies that at the core of affective decision-making lies *biological reinforcement learning* (Sutton and Barto, 1998) which allows the agent to incorporate new information, created by the interaction of the brain and the sensory organs, into the affective values of various features of the world (e.g., objects, humans, concepts) by updating the currently stored values. This also implies that the affective agent does not perform any reasoning in the usual sense of the word, does not build logically-consistent models of reality, and thus only uses the information about the environment remembered in the *affective values* of various features that this environment contains. To reiterate, the word "affective" here refers to how agent *feels* about a feature, given past sensory experiences with it. These affective values can be understood as biologically-inspired "preferences" in Robson's trio, which update via experience.

Our second assumption, built on modern neuroscientific evidence about how the brain encodes value, defines the procedure by which the affective agent chooses from among alternatives. In particular, we assume that all computations that the agent performs boil down to calculating weighted sums of values and choosing the largest of the two values. Both types of computations are very easily implementable with human neurons that are more than capable of adding neural signals, weighting them (through local inhibition or excitation, Dayan and Abbott, 2001), and comparing them (as in drift-diffusion models, e.g. Bogacz, 2007). These mechanisms represent the scope of "rationality" of the affective agent in Robson's terminology.

While these biological prerequisites do not change the nature of choice among *certain* outcomes (the agent still chooses the option that brings the highest affective value), they do give rise to behaviors that differ from rational in decision problems under *uncertainty*. The difference lies in how beliefs about the resolution of uncertainty are constructed. Since the only information that the affective agent possesses is stored in the affective values of the features currently present in his environment, we assume that the agent relies on them to make judgements about how uncertainty will resolve. Specifically, we define the agent's *mood* (or reference point) as the sum of the values of all features currently present in the environment. When choosing under uncertainty, the agent identifies which uncertain outcome is most consistent with his mood (i.e. is the closest in terms of value distance) and simply assumes that this is the outcome that will result when uncertainty is resolved, simplifying the decision problem under uncertainty to one of choice among certain outcomes. We call this process *mood affiliation* (coined by Tyler Cowen on his blog) and the beliefs generated by it the *mood-dependent beliefs*.

As we show below, mood affiliation produces behavior under uncertainty that reflects many "irrational" deviations from expected utility maximization noticed in laboratory experiments. For example, we show how risk-aversion in the gain domain and risk-loving in the loss domain emerge together with other biases inherent to Prospect Theory like over- or under-confidence (Kahneman, 2011). Motivated reasoning in some contexts (Bénabou, 2015) can be seen as a consequence of mood affiliation. Various other phenomena that we discuss in the paper include: procrastination, depression, addiction, and radical discounting of future outcomes (because they cannot be directly felt). More importantly, since the affective agent does not think rationally and does not strategize about what others will do, any decision problem (including games) that such an agent faces is conceptualized as an individual decision problem under uncertainty. This boundedness of the affective agent's rationality makes it very straightforward to predict what he will do in situations of arbitrary complexity, simply because the agent does not take such complexity into account. Thus, our formalization of choice under uncertainty completely specifies the behavior of the agent in any environment as long as the affective values of the features involved in choice can be specified.

After discussing how affective agents behave in decision problems, we then turn to modeling how they interact with one another. Here we build on the robust literature showing that humans are uniquely capable coordinators and cooperators. Humans are an extremely social and extremely pro-social species and have been for a very long time (e.g., De Waal et al., 2006;

Laland, 2018). Ample economic theory exists to show how cooperation and mutually-beneficial coordination can be sustained as equilibria of repeated games among rational agents, but extensive evidence of cooperation in isolated, one-shot interactions remains something of a puzzle. As above, we believe affect is an important piece of the puzzle – generosity comes more naturally when we're in a good mood; often we help someone because we *like* them; often we like someone because we are like them; etc. Thus, we believe it is reasonable to understand our social emotions, what we call *moral sentiments*, as derived from affect.[1]

Evolutionary accounts of human pro-sociality often rely on notions of cultural group selection and gene-culture coevolution; such an account readily includes the emergence of moral sentiments as a proximate evolved mechanism serving the ultimate "end" of cooperation and norm-compliance (Chudek and Henrich, 2011; Henrich, 2015). Thus, we understand these moral sentiments as evolved adaptations that motivate affective agents to live in groups and cooperate. Since, by construction, affective agents do not reason rationally, the only way that they can agree to work together is if they have *similar preferences* defined by the affective values attached to various features of the world. We suggest that evolution yielded agents whose moral sentiments depend on the similarity of their affective values because agents who like the same things are more likely to benefit from cooperation.

The simplest form of moral sentiment conceptualized in this way is the friendly attraction between people who are associated with similar features in some environment, for example between people who wear Spider-man T-shirts.[2] Why does such attraction exist? The reason is that since all agents maximize their affective values, they will tend to be surrounded be features that they like, which creates a correlation between perceivable features associated with someone and their unobservable affective values. Thus, being attracted to others with similar visible features is a strategy that helps to single out those with similar affective values, which in turn can be useful for working-together. Such friendly sentiments can lead to light forms of cooperation (herd behavior) that are characterized mostly by information sharing and some personal sacrifices (friends can help you to move a couch, but not too often). In more complex tasks (e.g., a stag hunt), such similar agents can agree on a joint course of action or working-together (Isoni and Sugden, 2018) simply because they like one another and thus resolve strategic uncertainty by believing that the other person will also hunt stag. At the same time, friendly feelings do not allow for cooperation in more demanding games (e.g., the Prisoner's Dilemma).

In our framework, these sentiments are represented by the affective values that agents attach to each other. For example, if agent i sees agent j with some visible features, agent i can compute the sum of his own affective values associated with these features and treat the resulting number

[1] Again, we ignore here the rational side of decision-making, which arguably plays an important role in regulating and directing the moral sentiments. By focusing on how an agent governed solely by affect would behave, we help formulate the "problems" which rationality may have adapted to help solve.

[2] This idea has a long history; e.g. Adam Smith refers to "the pleasure of mutual sympathy", or the pleasure that people get from knowing that they feel the same way about something as other people do (Smith, 1759)

as the affective value of agent j. This way, agents will like being around others with high affective values (which signals that their preferences are similar) and will try to avoid others with low affective values (which signals that their preferences are different). We argue that agents with similar *personal identities* (defined as the sets of affective values of all imaginable features) will choose to be around each other for the purpose of benefiting from cooperation. This mechanism demonstrates how the precursors for *in-/out-group sentiments* can emerge.

Generalizing a bit, a more potent version of working-together can arise if agents also have moral sentiments about their *social identities*. Social identity is defined via a set of features and corresponding affective values that anyone belonging to it must share (astronauts should be brave and own a space suit). Following the same logic as above, agents i and j who share a social identity know that they share similar affective values for the features associated with it. Social identities provide possibilities for stronger forms of cooperation because the consistency of affective values of their members is kept in check. If an astronaut exhibits cowardly behavior, others will judge him (i.e. reduce his affective value), and at some point, he will not be eligible to be considered an astronaut anymore. In this view, social identity prescribes a set of affective values against which any prospective or current group member can be evaluated.

Special manuals (e.g., the Bible) and rituals (e.g., Christmas) emerge for the purpose of making sure that the affective values of the members of a social identity (e.g., Christians) are in line. Behaving consistently with the values of a social identity and judging or punishing others who fail to uphold these values can be thought of as following *identity-based norms of individual behavior*. In addition, the desire to be associated with features that describe some social identity is a natural tendency given that identities help to work together (e.g., astronauts want to demonstrate that they are brave). This can be conceptualized as a moral sentiment related to the desire of *higher social status within an identity*.

The moral sentiments that surround social identity, namely those related to its maintenance through manuals and rituals, identity-based norms, and status help affective agents to support cooperation at higher rates within the identity, than without where only weaker form of cooperation can be established through friendly feelings. This property, that we demonstrate with a simple example, suggests why social identities emerge and are so ubiquitous in human societies. However, our framework also demonstrates the limits to cooperation that affective agents can enjoy. They, for example, cannot cooperate in the Prisoner's Dilemma without additional mechanisms that take into account the feelings and well-being of others, which suggests the role that rationality and moral reasoning (Kimbrough and Vostroknutov, 2020, 2021) could play in the development of pro-sociality in humans.

2 Individual Behavior

Reinforcement learning implies that an affective agent's state at a given moment is determined by the totality of his past experiences. Thus, we explicitly model how an agent's affective values reflect the cumulative effect of those experiences.

We begin by defining the *world* in which the agent lives and makes decisions. The world is defined as a large set of all possible *discernible features* $\mathcal{F}$ that the agent can perceive with his sensory systems as well as constructed, abstract, features that he can think about. For example, such features can be physical objects, other human beings, or concepts that the agent can imagine, e.g., democracy, bravery, sin, etc. Moreover, some entities can be perceived as sets of features: someone can be tall or short, black or white, blond or dark-haired, male or female, liberal or conservative, etc. Such characteristics are represented as collections of features associated with a person, for example someone can be perceived as $\{Tall, White, \textit{Dark-haired}, Male, Liberal\}$, which is a subset of $\mathcal{F}$.

The *life* of an affective agent is a countable sequence of experiences, each of which occurs while surrounded by some features from the set $\mathcal{F}$. These experiences can be the result of choice, but needn't be: the agent's "values" can change from simply experiencing things that were the result of exogenous events. We postulate that at any time t, the agent has *affective values* $v_t : \mathcal{F} \to \mathbb{R}$ associated with each feature $f \in \mathcal{F}$.[3] It is clear from the outset that the agent cannot ever experience all possible features that exist in the universe; for simplicity, we assume that for any feature f that the agent has never experienced up to period t, $v_k(f) = 0$ for all $k \leq t$.[4]

At the beginning of period t, the agent is *surrounded* by a finite subset of features F_t. The term "surrounded" means here both all the physical entities that the agent can perceive as well as all the concepts and thoughts that the agent has in his mind (also represented by features). The features in F_t define the *affective state* $V_t = G(F_t; v_t)$ of the agent, which is some function of the affective values of the currently present features. The nature of the function G can be debated, but for the sake of biological plausibility, we will assume that

$$V_t = G(F_t; v_t) = \sum_{f \in F_t} v_t(f),$$

or that the affective state of the agent is simply the sum of the current affective values of each feature $f \in F_t$ present in period t.[5] The affective state V_t can be thought of as the agent's current

[3] The tuple $\langle \mathcal{F}, v_t \rangle$ can be called affective agent's *heterophenomenology* (Dennett, 1991) or his *umwelt* (Uexküll, 1926). This is a description of how the agent sees and perceives the world around him, which can change with new experiences.

[4] In reality, our species' evolution probably endowed us with some innate aversions (e.g. snakes) or attractions (e.g. sweet things) to certain features, which nevertheless can be updated via experience.

[5] It is here that the affective value of zero for previously unexperienced features becomes important: such features, even if added to the set F_t, will not influence the affective state of the agent. In general, the value of zero assumed for the unexperienced features is related to the additive property of G. If G were different, for example

mood, reference point, or general emotional state at period t. We assume that the affective agent in period t feels only V_t and *does not have conscious access* to the individual affective values v_t nor does he have any understanding of the aggregation function G that produces V_t.[6]

The reason we divide the otherwise continuous life of the agent into separate periods is the idea that in each period *something happens* to agent's affective state V_t. If at the beginning of period t, the agent "has" the affective state V_t, that state is assumed to persist for some non-negligible interval of continuous physical time. Then, during period t, some physical interaction with the outside world can bring about a *change*, so that, instead of the current V_t, the agent feels an *affect* $V_t' \neq V_t$ (sometimes we will also call V_t' an *experience*).

To illustrate how such change can happen, imagine that the agent, who is surrounded by features F_t, *experiences some additional features* that can arise through the agent's choice, a choice made by some other agent(s), or an act of Nature. For example, the agent is in a restaurant choosing between two predetermined full-course dinners. In this case, the whole restaurant setting—the paintings on the walls, the view from the windows, the memories of past dinners with friends—constitutes the static (in period t) set of features F_t that create the affective state V_t. The full-course meals, which also consist of various features (starters, main courses, deserts, wine), are not part of F_t because the agent needs to choose between them and does not experience these features at the moment of choice. Since the agent's affective values v_t are also defined over the features within the full-course meals, the agent aggregates the affective values of the features in each meal using the function G; makes a choice between them; experiences the chosen meal, which produces the affect V_t'; and updates his affective values of all the features involved. The fact that the agent makes a choice is not important here, as the features (within meals) could have been chosen by someone else (e.g., the chef). What is important is that the agent experiences additional features that were not part of F_t (we return to the precise mathematical description of this later in the section).

Another way the agent can feel an affect V_t' is when *the affect is triggered by some events that are unrelated to any features*. This can happen when the agent experiences a consequence of his or someone else's action, a random event, or a combination of these factors. For example, in the restaurant the agent falls on a slippery floor and feels pain. In this case, V_t' will be defined by this event.

After the agent experiences V_t', the affective values of the current features get *updated*. Specifically, we follow a long tradition in the reinforcement learning literature (Sutton and Barto, 1998) and define value updates of each feature $f \in F_t$ (and also the meal-related features if they are

multiplicative, then the value of unexperienced features should change to 1 accordingly, so that adding them to F_t does not change the affective state. It is also important to notice that unexperienced features might be those that are present in the surroundings, but are not *discernible* by the agent, who might not have the sensory system refined enough or the computational capacity large enough to perceive them.

[6] It is possible that the agent can learn to access this information, which is arguably an important component of the interaction between rational and affective decision-making. We consider this possibility in Appendix A where we discuss emotional intelligence.

involved) as

$$v_{t+1}(f) = v_t(f) + \lambda(V'_t - V_t).$$

Here $\lambda \in [0,1]$ is a reinforcement learning parameter that determines the relative importance of the past versus current experiences. The difference $V'_t - V_t$ represents the *instantaneous emotion* that the agent feels when experiencing V'_t, which in the neuroscience literature is also known as *prediction error.* This is the fundamental element of mammalian brain architecture that is involved in most decision-making (Gllimcher et al., 2009). The features that were not present in period t are not updated, so for any $f \notin F_t$ we have $v_{t+1}(f) = v_t(f)$. We demonstrate how this works with an example.

Example 1. Catching a Cold in Paris. Suppose that you always wanted to visit Paris. You saw many photographs and heard nice stories about the city life, so your affective value of the feature *Paris* is high, for example, $v_0(Paris) = 5$. You have also heard good things about France in general, but not as specific as about Paris, so $v_0(France) = 3$, positive but not as high. In period 0 you travel to Paris. The set of current features is $F_0 = \{France, Paris\}$. Unfortunately, upon your arrival you catch a cold, which makes you feel bad for the duration of the whole trip (period 0). This sours your impression of Paris and your experience is $V'_0 = 1$, which is positive, but much less than what you have expected ($V_0 = 5 + 3 = 8$). You feel a negative emotion (disappointment) with valence $V'_0 - V_0 = -7$. Suppose that $\lambda = 0.9$, such that you put a high weight on your personal experiences relative to the opinions you heard from others in the past. So, you enter period 1 updating the affective value of the feature *Paris* to $v_1(Paris) = 5 - 0.9 \cdot 7 = -1.3$. The value of *France* is updated to $v_1(France) = 3 - 0.9 \cdot 7 = -3.3$. When you get back home in period 1, you have a negative impression about Paris (−1.3) and even worse one about France (−3.3). These negative impressions might prevent you from taking a new job in Paris because you feel that you do not like the place. You also might stop eating at a local French restaurant, because your perception of the food includes the feature *France* that you now like even less. □

This example shows how high expectations ($V_0 = 8$) can backfire, when an experience fails to live up to them ($V'_0 = 1$) and can lead to negative emotions that, in turn, influence future choices. In addition, two more worrying observations about the behavior of affective agents can be made. First, the affective values of both features are influenced by the presence of each other in the updating, as they are experienced *together*, which generates a sort of "exaggeration" of affective values. Second, all the negative feelings stem from a bad experience that is unrelated to physical properties of Paris or France, which also leads to wrong impressions about them. We refer to exaggerated affect based on mere associations as *irreality*, and we discuss it further below.

Next, we modify the example with an explicit choice of the affective agent instead of a feature-unrelated event "catching a cold." The setup is as above, only now suppose that instead of getting a cold you go to a restaurant in Paris and choose between ordering escargot (feature f_E) or quiche (feature f_Q). You heard good things about escargot, which you never tried before, so you believe that its affective value is higher than that of a quiche, or that $v_0(f_E) > v_0(f_Q)$. When you try escargot, you do not like it that much, which makes you experience $V_0' = 1$. After that the updating happens exactly as before with the same consequences for future behavior, and the value of escargot gets updated as well.

These examples demonstrate the universality of the reinforcement learning mechanism in dealing with many kinds of situations, regardless of the exact nature of the process that leads to the final experience V_0'. This is a good property, as it allows the agent to react to and learn from a wide range of different experiences. However, the "model-free" interdependence of the perception of the features can have its drawbacks and lead to irreality, that in its turn can result in bad decision-making in the future.[7]

In situations with certainty discussed up to this point, there are two types of experiences that the agent can have. We can describe them as follows:

1) Beginning of period t. The agent is surrounded by features F_t which give rise to his affective state $V_t = G(F_t; v_t)$;

2.a) Additional Features. The agent experiences an additional set of features F, which results in affect $V_t' = G(F; v_t)$; or

2.b) Feature-Unrelated Events. A feature-unrelated affective value $s \in \mathbb{R}$ is experienced by the agent. The affect is $V_t' = s$;

3) Updating. The agent updates all features in F_t (and in case 2.a also the additional features F) using the experienced value V_t' as described above: for each $f \in F_t$ (or $f \in F_t \cup F$ in 2.a) the updated affective value is $v_{t+1}(f) = v_t(f) + \lambda(V_t' - V_t)$. The affective values of all other features stay unchanged.[8]

2.1 Affective Decisions under Certainty

Given these definitions, we can formulate the general *decision problem under certainty* that brings them together. At the beginning of period t, the agent is surrounded by features F_t and has affective values v_t. Suppose that the agent is choosing one action from some set A and that for each potentially chosen $a \in A$ the agent would experience (or believes that he will experience) an

[7] In Appendix A we discuss how emotional intelligence can help with preventing irreality from taking hold.

[8] It is worth noting at this point that in this paper we only consider the experiences of the agent leading to affect V_t' that are caused by some *physical* events (having a meal, falling on the floor, etc.). There is an additional class of experiences that can change affective values related to the *arrival of new information*. This is an important source of change, but it is too complex to consider here. For example, it is important *from whom* this information is coming which determines whether the agent will believe it or not. We consider this and other related issues in a companion paper on social learning (Kimbrough et al., 2020).

additional set of features $F(a)$ and a feature-unrelated affective value $s(a)$. Agent's maximization problem is

$$\max_{a \in A} G(F(a); v_t) + s(a).$$

The two implicit assumptions here are: 1) if the agent does not experience any additional features ($F(a) = \varnothing$) then $G(\varnothing; v_t) = 0$; and 2) if the agent does not experience any feature-unrelated value after a, then $s(a) = 0$.

Suppose that action $m \in A$ maximizes the expression above and the agent chooses it. If the agent's beliefs about what will happen after m are correct, his affect will be $V'_t = G(F(m); v_t) + s(m)$, which is then used to update features $F_t \cup F(m)$ in v_t (all other affective values stay the same). If something unexpected happens after m is chosen, the affect experienced by the agent will be some other V'_t, which is instead used to update features $F_t \cup F'$ (F' here is potentially different from $F(m)$). There are no other consequences of inconsistent beliefs. For simplicity, we also assume that the affective values of the additional features $F(a)$ that would have appeared after the unchosen actions $a \in A \backslash \{m\}$ are not updated.

Note that the affective agent's approach to a decision problem is completely general as long as there is no uncertainty. This agent does not reason about the future and thus considers any choice as a one-stage decision-making problem as described here. In particular, the agent cannot "see through" to his own future choices either (Example 6 in Appendix A describes possible consequences of this), so there is no need to formulate more complex decision problems with a single agent acting in multiple periods. Rather, anything the agent cannot fit into the one-period decision problem with certainty is treated as "uncertainty." Such uncertainties include: future actions of the agent himself, the future or simultaneous actions of other agents, acts of Nature, etc. In the next sub-section we formulate the decision problem with uncertainty that covers all these cases and thus automatically describes how the agent will behave in any strategic environment of arbitrary complexity as long as the affective values of outcomes can be clearly spelled out.

2.2 Affective Decisions under Uncertainty

In decision problems with certainty, the behavior of the agent is identical to a standard utility maximizer in the sense that he simply chooses the highest affective value in the feasible set. However, this is where the similarity ends. Under uncertainty, the affective agent still only has access to the affective values of the features as coded in V, the current mood or reference point (in this section we drop the subscript t for convenience); thus, to model choice under uncertainty we have to make assumptions about how the affective agent imputes affective value to an uncertain outcome.

To begin, we offer a definition of an uncertain outcome, which cannot be based on probabilistic beliefs, since we assume these are not part of the affective agent's calculus. We assume that for the affective agent, each outcome with uncertainty (or a lottery) consists of two parts: a "certain" part, defined as above with some affective value s (possibly coming from additional features) and an "uncertain" part defined by a set of additional affective values U (see Appendix B). Only one of these uncertain values will be realized, but, at the moment of choice, the agent does not know which.[9] Overall, a *lottery* is defined as a tuple $L = \langle s, U \rangle$ with the idea that, if chosen, the agent gets the affective value s plus the affective value of the realized uncertain outcome from U. So, the set of *possible affective values* of lottery L is $U_L = \{s + u \mid u \in U\}$.[10]

To provide an example of choice between lotteries, consider a full insurance problem. The agent is choosing between 1) paying some amount of money s today and having no uncertainty about the future in case something bad happens (insurance fully covers the expenses); and 2) enjoy money s today and face uncertainty about the future where something bad can happen or not. The former lottery is given by $L_1 = \langle -s, \varnothing \rangle = -s$, which is a sure outcome. The latter lottery is $L_2 = \langle s, \{0, -100s\} \rangle$, where $-100s$ represents the affective value of something bad that can happen in the future (hundred times worse than $-s$). The set of possible affective values of L_2 is $U_{L_2} = \{s, -99s\}$.

How can the affective agent choose among lotteries thus defined? Since all the information that the agent possesses in any period is contained in V (the sum of the affective values of the current features), this value must be used to reason about uncertainty resolution. We assume that the agent treats mood V as informative about how the uncertainty will resolve. We propose a simple mechanism that we have dubbed *mood affiliation*. The agent reasons that since V contains all the information about the current environment, the uncertainty will resolve in a way consistent with it. Namely, the agent forms a *mood-dependent belief* by finding the possible affective value of L from U_L that is the closest to V. Let us define such element L_V as

$$L_V := \arg\min_{u \in U_L} |V - u|,$$

or the element of U_L closest in value to the current mood V. Given this mood-dependent belief, the agent simplifies the lottery $L = \langle s, U \rangle$ to its "expected" value $L_V - c(U_L)$, where $c(U_L)$ is the *cost of uncertainty*. In the absence of evidence, we remain agnostic about the exact nature of the function $c(U_L)$, with the caveat that it should be simple, given the other assumptions we make about the affective agent.[11]

[9] From the perspective of expected utility, the distinction between certain and uncertain parts of a lottery is irrelevant. However in affective decision-making, it does make a difference, which turns out to be consistent with "certainty effects" observed in many experiments (e.g., Kahneman et al., 1986). We discuss in more detail why lotteries might be represented this way in Appendix B.

[10] Notice that in this "light definition" we use only feature-unrelated affective values represented as numbers. In general, each outcome of a lottery should be described as a collection of additional features plus some feature-unrelated affective value, which we call an *occurrence*. For a general, notation-heavy definition of a lottery see Appendix D.1.

[11] One possibility is that $c(U_L)$ is inversely proportional to the worst possible outcome in U_L. For example, if U_L includes an affective value of, say, serious injury, this will make the cost of uncertainty very high. Alternatively, the

To see how this works in the example above, suppose that the agent is choosing between $L_1 = -s$ and $L_2 = \langle s, \{0, -100s\} \rangle$. Further, suppose that the agent is in a good mood, such that $V > 0$. Then, he will believe that in L_2 the uncertainty will resolve with the possible affective value s from the set $U_{L_2} = \{s, -99s\}$, which is the closest to his current mood V. In other words, the happy agent holds a mood-dependent belief that "nothing bad will happen in the future." With such a mood-dependent belief, the choice simplifies to having $-s$ if L_1 is chosen or having $s - c$, if L_2 is chosen. If the cost of uncertainty c is less than $2s$, then the agent will choose L_2, or no insurance. If, instead, the agent is in a bad mood, such that $V < -50s$, he will believe that the uncertainty in L_2 will resolve to $-99s$ (the agent feels that "something bad will definitely happen") and the affective value of L_2 consequently becomes $-99s - c$, which is much smaller than $-s$ that can be obtained from choosing L_1. As a result, the agent in a bad mood chooses L_1 to insure himself (Brighetti et al., 2014; Pauly and Kunreuther, 2019).

Now, we can define a *decision problem under uncertainty* (for the general definition see Appendix D.2). Suppose that agent's current mood is V and that he faces a choice among actions in some set A. Each action $a \in A$ leads to a lottery $L(a) = \langle s(a), U(a) \rangle$. The agent chooses the action that leads to the lottery with the highest *expected affective value*, or the one that solves

$$\max_{a \in A} L(a)_V - c(U(a)_{L(a)}).$$

After the choice is made, the uncertainty is resolved and the agent experiences the affect V' that is the sum of the certain part of the chosen lottery and the realization of its uncertainty. Then, the features are updated as before.

Example 2. Prospect Theory and Under-/Over-confidence. To illustrate this model of choice under uncertainty, we consider the main tenet of Prospect Theory (Kahneman and Tversky, 1984), namely that people are risk-averse in the gain domain and risk-loving in the loss domain.

Suppose the choice is between actions a_1 and a_2 that lead to a lottery $\langle 0, \{x, y\} \rangle$ and a sure outcome z with $x > z > y$ as shown in Figure 1. Suppose first that the agent is happy, so that the mood or reference point V is high (in red in the figure). In this case, all possible outcomes are in the loss domain (they feel worse than the current mood). For the lottery, the agent will believe that the outcome x will happen, since it is the closest to V (red circle). Thus, the expected affective value of the lottery is $x - c$. As long as c is small enough, we will have $x - c > z$ and the agent will choose the risk-loving action a_1. We can also say that the happy agent is *overconfident*, because he believes that the good outcome of the lottery will be realized. So, a happy and overconfident agent chooses the risky option in the loss domain. Now, suppose instead that the agent is sad and has a low V (in blue), which places all possible choice outcomes in the gain domain. Then, the agent will mood-dependently believe that the outcome y will be realized

cost of uncertainty may increase with the difference between the values of the best and the worst possible outcomes in U_L, a rough measure of variance, which is reminiscent of mean-variance utility in finance (Preuschoff et al., 2006).

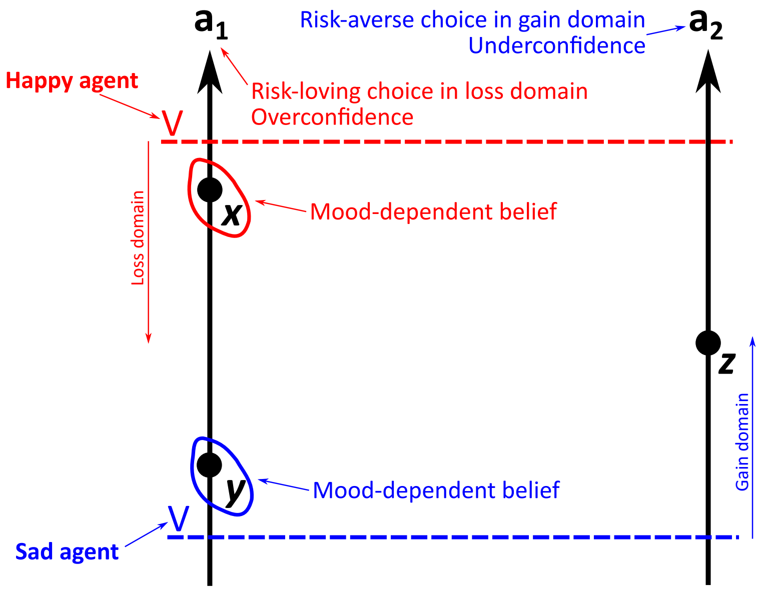


Figure 1: Choices of the affective agent are consistent with Prospect Theory.

(blue circle), in which case he will choose the risk-averse action a_2 (since $z > y - c$). This can also be seen as *underconfidence*, because the agent believes that the uncertainty will resolve in a bad way. Thus, a sad and underconfident agent chooses risk-averse actions in the gain domain. □

This example demonstrates that our framework of affective decision-making can be seen as a "psycho-economic" foundation of Prospect Theory and related phenomena of under- and overconfidence. Unlike its predecessors though, our theory explicitly specifies where the reference points come from (they are aggregated from the affective values of features), which makes it much more amenable to testing than other theories that employ reference points but are agnostic about what they are. Moreover, our model predicts how affective agents, who exhibit prospect-theory-like behavior, would choose in *any* situation.

For example, imagine that the options a_1 and a_2 in Figure 1 are not individual choices, but rather constitute a choice in the first node of a game. After a_2 is chosen, the game ends and the agent gets the affective value z. After a_1, the move goes to some other player, who chooses between actions leading to affective values x and y for the agent. Given that affective agents do not reason rationally, they will choose in this game in exactly same way as they would have in the individual choice problem (with the only difference that the other player becomes a feature that influences mood V, see Appendix E). Thus, our framework provides a model of boundedly-rational choice in any imaginable setting. All this is possible exactly because affective agents, unlike rational ones, are ignorant of any complexities emerging from modeling uncertainty (with probabilities for example) or from strategic interactions.

3 Moral Sentiments

In this section, we provide arguments about how affective agents can cooperate or work together. As noted in the introduction, our premise is that social emotions play an essential role as a proximate mechanism for achieving the ultimate goal of facilitating cooperation and coordination. Recall that affective agents operate in an informational environment defined only by the affective values of features and rely on mood affiliation to decide under uncertainty. This creates specific constraints on how affective agents can achieve cooperation, as we must root the social emotions in the same process of valuation that applies to any other feature. We argue that when affective agents "sync" their values of features, this allows them to work together because they value the same things. We then show that this gives rise to various recognizable (if not always rationally justifiable) moral sentiments, for example, in-group bias, out-group aversion, attraction to ritual, deference to higher status individuals, etc.

3.1 Herd Behavior

The simplest example of social coordination by affective agents is "herding behavior". In the individual decision problems described in Section 2, we assumed that agents have affective values defined over features and that they choose features that give them the highest affective values. If features are distributed unevenly in space, then isolated individual decision-makers will nevertheless converge on locations that contain features that they like. For example, at scientific conferences, attendees typically go to those presentations that are most appealing, given their interests; the consequence is that each session is populated with researchers who have similar affective values, simply because they are attracted by the same features of the presentations. Thus, agents can self-select into groups of similarly-minded individuals for purely "selfish" reasons. Such groups then become capable of herd behavior because they consist of individuals with similar preferences.

This example is almost trivial, but it illustrates an important property of affective agents' decision-making process that can result in them congregating in "like-minded" groups even without having specific spatial locations as attractors. The property is that when affective agents make choices over time, they will tend to be surrounded and thus associated with features that they like (and dissociated with features that they do not like). In other words, the features around the agents and their affective values naturally become *correlated*. This is crucial because affective agents do not observe each other's affective values directly, but can only observe features associated with various individuals. If features associated with others are correlated with their affective values, then these features can serve as *signals of similarity*.

For example, a child who likes Spider-man goes to a new school and meets other children in his class. Some of them wear T-shirts depicting Spider-man, which attracts the child to them because they apparently like the same things. As a result of a group dynamics based on this simple account, children in the class will assort into subgroups with shared interests (e.g., superhero fans, science lovers). This mechanism can be seen in action in the minimal group paradigm experiments (Tajfel and Turner, 1986; Chen and Li, 2009) where people are divided into groups depending on their preference for Kandinsky or Klee. Subjects are more pro-social towards the members of their own group (we discuss strategic interactions like this in Appendix E).

In our framework, then, the way the affective agent feels about others is defined by the affective value he assigns to the features that they possess or are associated with. Suppose that agent i with affective values v^i is introduced to agent j who has some visible features H_j (we again drop the time subscript t for convenience). For example, these can be $\{$*Red-haired, Spider-man T-shirt, Hat*$\}$. Agent i then forms an affective value $v^i(f_j) = \sum_{f \in H_j} v^i(f)$ associated with the feature f_j representing agent j. If agent i likes red hair, Spider-man, and hats, then agent j will have a high affective value, and vice versa. As time goes by, agent i might learn more things about agent j, for example that she has a cat and goes skiing. The features $\{$*Cat*, *Skiing*$\}$ get added to the set H_j that codes the features representing j with the corresponding re-computation of the aggregate feature f_j.

Despite its simplicity, the mechanism of assigning affective values to others based on the features that they possess has powerful implications. For example, skin color and sex are highly visible human attributes. Following the logic above, if an affective agent meets a stranger with a specific skin color and/or sex to which the agent assigns negative affective values (for whatever reason), then the agent will assign low affective value to this stranger and behave accordingly towards this person (see Appendix E). Such a mechanism can readily account for discrimination and many other bad behaviors that are completely unrelated to the actual qualities of the stranger per se. Moreover, the logic of this suggests that it does not matter *how many* features are associated with a particular individual. The judgement is based only on the information available at the moment, which in many cases boils down to visible physical attributes that end up being used to judge people in general.

At the same time, this mechanism also suggests that such negative affective biases are malleable. New information (e.g., awareness of new features to which an agent assigns a high affective value) can overcome initial negative bias. Similarly, if an agent identifies common interests with someone to whom he was initially negatively disposed, and as a result they have a positive experience when interacting, then all features get positively updated, including the features that were previously negatively biased. As a result, the agent's attitude towards others who share this feature becomes more positive. Thus, although our framework is simple, it makes important predictions about how to deal with discrimination, which are consistent with recent evidence on how positive intergroup interactions reduce prejudice (Mousa, 2020).

In sum, we reiterate that a key consequence of assigning affective values to others is that agents tend to become assortatively matched with those who value similar features. The effects of such a process needn't always be positive, but clustering does create opportunities for various

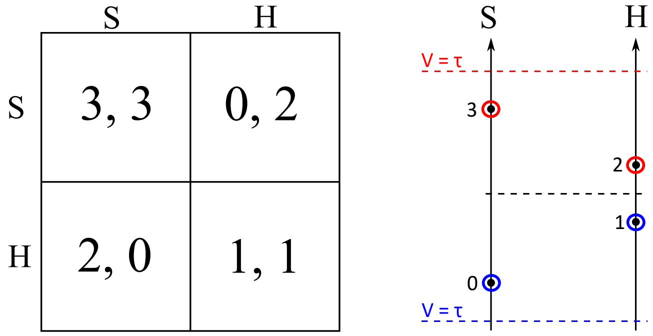


Figure 2: In the Stag Hunt game (Left Panel) agent i with mood $V = v^i(f_j) = \tau$ plays S when he has high affective value $\tau > 1.5$ attached to agent j and H otherwise (Right Panel). We ignore the cost of uncertainty assuming it equal for both actions.

simple forms of cooperative (herd) behavior. Consider how affective agents play the Stag Hunt game shown in Figure 2. The equilibrium selection problem is resolved by the agent's mood. In particular, the right panel of Figure 2 shows that the agent plays S when he "likes" the other player (high τ) and plays H when he doesn't (low τ). Thus, assorting on features encourages cooperation among affective agents, because meeting similar others puts them in a good mood. Notice as well that herd behavior cannot resolve more demanding cooperation problems. Figure 9 in Appendix C shows that good mood alone does not make affective agents cooperate in the Prisoner's Dilemma, which shows that herd behavior on its own is not enough for extended cooperation (though see Example 7 in Appendix A). Next, we suggest how this assortativity may bootstrap more complex forms of cooperation and coordination by generating notions of *identity*.

3.2 Identity, In-group, and Rituals

Social scientists increasingly understand agents' willingness to engage in the kinds of costly commitment and enforcement that undergird cooperation through the lens of *identity* (Akerlof and Kranton, 2000). If people think of themselves as such-and-such a person, then they will sometimes be willing to take costly actions to maintain that self-concept. We argue, using our terminology, that collections of affective values can be seen as defining who a person is or to which group he belongs. In particular, the affective values v^i_t of agent i in period t can be thought of as his *personal identity*, or "the totality of one's self-construal" (Weinreich and Saunderson, 2005). In this sense, we can talk about groups of friends or people with common interests discussed above as those who share some aspects of personal identity and cooperate based on that.

However, for more complex forms of cooperation the literature distinguishes *social identity* (Erikson, 1968), or that part of personal identity which defines individuals as belonging to a certain social group. Social identity can be conceptualized as a "meta-feature" that defines a group in terms of a collection of features by which one can evaluate its members (in good standing).[12]

For example, the social identity "astronaut" can be represented by the meta-feature $g = Astronaut$ that is defined by a set of features $F_g = \{g_1, ..., g_N\}$ and values $\tilde{v}_g(g_k)$ for $k = 1..N$ associated with g that define qualities, physical possessions, or opinions of people who properly belong to g ($\tilde{v}_g : F_g \rightarrow \mathbb{R}$ is different from the agent's affective values v^i, see below). For astronauts the features g_1 through g_5 might be $\{Resilient, Brave, Knowledgeable, Space\ Suit, Alcohol\}$ and corresponding values $\tilde{v}_g(g_k)$ that are high and positive for the first four features ($k = 1..4$) and very negative for the last one ($k = 5$). This means that astronauts should exhibit resilience, bravery, and knowledge, should have a space suit, and should (actively) avoid drinking alcohol. Those who satisfy these criteria and possess feature $g = Astronaut$ will be seen as astronauts, while those who do not, will not.

The implication is that social identities give rise to an *in-group* that consists of all agents who possess g. Similarly, an *out-group* is then defined as all agents who do not possess g, or perhaps those agents who no longer have g because they no longer possess certain features associated with it (an astronaut behaves as a coward). We discuss this in more detail in Section 3.3.

These definitions are very abstract and admittedly ad hoc. However, they can be useful to describe the ways social identities emerge and change with time. Some identities, such as ethnic and national identities, are assigned to anyone who was born and raised by particular people or in a particular place (e.g., Americans). Others, such as occupational or political/religious identities, can emerge from the process of assortative matching on affective values described in the previous section. A group of people, drawn together by common interests, can develop a social identity over time that is based on core features to which they share specific attitudes. This set is a union of features and values that all these agents have in common.[13] More to the point, this shared set of features and values can facilitate cooperation with the in-group.

The key to cooperation among affective agents with a shared social identity g is in the way they make decisions (Section 2). Suppose that two agents i and j are hunters (identity g) who share identical affective values $\tilde{v}_g$ of the features $\{g_1, ..., g_N\}$ which constitute identity g (so that for any g_k we have $\tilde{v}^i(g_k) = \tilde{v}^j(g_k) = \tilde{v}_g(g_k)$). Imagine that i and j go hunting together and that they need to cooperate in order to catch some prey (each agent individually cannot catch anything). Suppose that the agents face some choice that involves uncertainty. For example, they need to decide whether to go to site A or site B where two different types of prey may or

[12] The construction of such "meta-features" seems to be a natural by-product of how the brain organizes the sensory world. Connectionist theories in neuroscience suggest that our brain constructs our sensory experience out of layered abstractions. Perceptible "features" are constructed as the brain begins to classify recurring sensory inputs as elements of a "class"; when a set of such features co-occur with sufficient regularity, they become identifiable as a class of their own. This conceptual framework motivates the construction of neural networks (Buckner and Garson, 2019).

[13] Note that this also implies that the defining features of a social identity can evolve as the common affective values of people associated with them change.

may not be present. Given that the agents face the same environment (they hunt together) and thus the same set of current features $F_t \subseteq \{g_1, ..., g_N\}$, we know from Section 2 that they will agree on the choice between A and B since they have the same function $\tilde{v}_g$ defining their moods $V_t^i = V_t^j = G(F_t; \tilde{v}_g)$ and mood affiliation will thus lead them to reach the same conclusion.[14] Consequently, given same moods, they will work together successfully (choice under certainty goes without saying).

In general, as long as agents have the same affective values defined by $\tilde{v}_g$, they will agree on choices under uncertainty and will resolve uncertainty *even about the behavior of one another* in the same way. They will also agree on any future uncertain choices related to hunting that they may face. This is a much more powerful form of cooperation (as compared to herd behavior) that can get many things accomplished. Thus, we can conclude that if a group of affective agents possess a social identity g, they can efficiently work together in any common environment defined by the features in $\{g_1, ..., g_N\}$. Notice also that these agents will automatically like each other and strive to work together given the arguments in Section 3.1. Social identity arises from a strong form of common interest: all agents with shared social identity will have a deep appreciation of the values and features related to it.

This idealized example captures the mechanism through which affective agents with a perfectly shared social identity can work together. However, as we know from Section 2, things are not likely to be that simple. As long as agents' values are "close enough" they should be able to reap many of the benefits of cooperation, but agents who have different experiences will update their affective values differently. Therefore, it is reasonable to think that, as time goes by, the affective values $\tilde{v}_g$, even if perfectly attuned between agents i and j at some point in time, will start to diverge as a result of agents idiosyncratic experiences (so that now $\tilde{v}_g^i \neq \tilde{v}_g^j$). This suggests that social identities will not generally persist in a crystallized state once formed; identities need *maintenance*. Agents belonging to a social identity need to keep affective values $\tilde{v}_g$ of the identity-defining features $\{g_1, ..., g_N\}$ in tune with one another, to make sure that they are similar across everyone who has identity g and thus can reap the benefits of cooperation.

One way identities are maintained in the social world is through shared texts or manuals that codify and reinforce the identity group's salient features and associated values. For example, the Bible can be understood, in part, as a manual detailing the set of relevant features and the associated values $\tilde{v}_g$ for the social identity *Christian*. However, another important means of aligning the values of identity-related features is *ritual*. Rituals can be seen as "scheduled" common practices for a group of agents that ensure that the identity group's values are common knowledge and synchronize them by producing common experience (Chwe, 2013; Henrich, 2020). Indeed, people who pray together in a church, who perform a traditional dance, or attend a professional refresher course, all align their affective values through a *common experience*. In our framework,

[14] When agents i and j work within their social identity g (hunt), they use the values $\tilde{v}_g$ of features associated with g instead of their common affective values defined in v^i and v^j.

the importance of the fact that the experience is common is that people feel the same emotions together and thus update their affective values with the same affect V'. This necessarily brings their affective values closer to each other, as the updating is a convex combination of the old affective values and V'.

3.3 Social Status

Since membership in a social identity group facilitates mutually beneficial cooperation for affective agents along the lines described above, it follows naturally that agents might want to ensure that they are seen as exemplars of their identity group. Psychological evidence suggests that people usually want to have *some* social identity (Pickett et al., 2002) and suffer from not having one (Brewer, 1991). Children as young as 5 years of age already react to cues related to identity, in-group, and status (Nesdale and Flesser, 2001), which suggests that identities develop at early age, are learned from kin (Laland, 2018) and are hard to change (Wexler, 2006).

Thus we assume that early experiences that result in a high affective value being attached to a specific social identity g will also tend to yield a desire to possess the features associated with it. Such a desire can be understood as a *desire for social status within* g. In all human societies people with high social status (not necessarily rich ones) are those who are exemplars of their class. Famous shamans, scientists, hunters, artists, politicians, athletes, entrepreneurs, etc. are those who have earned respect by possessing the features that describe their chosen social identities.[15],[16]

We can define a measure of *status within social identity* g as the number or the aggregate affective value of features from the associated set $\{g_1, ..., g_N\}$ that a given affective agent possesses. Suppose you are learning to become an astronaut, a social identity described by the features $\{$*Resilient, Brave, Knowledgeable, Space Suit, Alcohol*$\}$. Then, if you are resilient, brave, and knowledgeable (which you have demonstrated through your behavior), but do not have a space suit and do not express an opinion that drinking alcohol is bad, then you are a worse astronaut than someone who does have a space suit. Yet another person who also does not drink is a better astronaut than both of you. The reflection of this idea can be seen everywhere in the artificial rankings that are created within each social identity (e.g., profession) to make the assignment of status easier. If you are an economist, you go through various stages: PhD student, Post-Doc, Assistant, Associate, Full Professor. Businessmen rank themselves by comparing their companies' market value or quarterly sales. These ranks reveal the criteria by which one achieves

[15] This obviously applies equally well to seemingly "unproductive" activities like conspicuous consumption among the leisure class as depicted by Veblen (1899). If you are born into Veblen's leisure class, then your social identity is defined by possessing features that project wealth (houses, land, lavish parties, etc.). Thus, the desire to consume conspicuously can be understood in our framework as a reflection of attempts to become a better representative of this specific social identity.

[16] Note that many moral and political arguments are often about precisely the criteria by which someone ought to be judged an exemplar.

status and also make for easier judgements by the general population or people who need to cooperate with someone from another profession. In the view of our framework, an important consequence of agreement on such a status hierarchy is that it facilitates cooperation among affective agents by the mechanisms described above, perhaps explaining why such hierarchies are sticky and widespread.

To understand how the motivation to acquire social identity and related concept of desire for social status fit into our framework we introduce an additional mechanism, *anxiety*, which pushes affective agents to acquire social identities (and status within them).[17] Translated into the framework of affective decision-making, this means that an agent, who currently (in period t) possesses certain features, experiences anxiety if there are some other features that the agent wants to possess but currently does not. This anxiety then affects the agent's mood.[18]

To provide intuition, suppose that an agent i wants to become an astronaut (identity g), which means that i wants to possess the set of features F_g that define the social identity *Astronaut* and has some affective values assigned to them. These values are defined by his current opinion $\tilde{v}^i_g$ about what $\tilde{v}_g$ is. His affective value of being an astronaut is then equal to $\tilde{V}_g = \sum_{f \in F_g} \tilde{v}^i_g(f)$. Suppose that at the beginning of period t, agent i possesses some features from F_g, say those in the set $F'_g \subseteq F_g$, with corresponding affective value $V_g = \sum_{f \in F'_g} \tilde{v}^i_g(f)$. The fact that i wants to acquire all features in F_g but does not yet have them creates anxiety that influences his mood V_t. This happens in the following way. Suppose that the agent is surrounded by some features F_t, then his mood at the beginning of period t is

$$V_t = \sum_{f \in F_t} v^i_t(f) - \phi_i(\tilde{V}_g - V_g).$$

Here V_t consists of two components: the aggregate affective value from the current features F_t and the anxiety that he feels due to not having the remaining features in F_g. This latter component is expressed as a negative of the difference between the value $\tilde{V}_g$ and V_g multiplied by some individual coefficient $\phi_i \geq 0$.

To illustrate, suppose that during period t the agent has a choice between *inviting friends for a party*, which does not help to obtain any features from F_g but brings a feature-unrelated affective value $s > 0$, or *doing homework*, which gives the agent the feature $g_1 \in F_g$ described as *Knowledge about International Space Station*. The affect from choosing to party with friends is equal to $V'_{party} = s$ and the affect from doing homework is given by $V'_{hw} = \tilde{v}^i_g(g_1)$. Thus, the agent will choose to do homework if $V'_{hw} > V'_{party}$ or when $\tilde{v}^i_g(g_1) > s$, and to party otherwise.

[17] We thus offer a first step toward addressing Agnes Callard's criticism that rational choice theory currently has no place for a notion of "aspiration" (Callard, 2018).

[18] There is evidence that individuals with (perceived) low status experience constant stress that is reflected in high levels of cortisol (Cummins, 2005). We suggest that stress can be understood as the anxiety of not having certain features related to these individuals' desired social identity.

There are several important things to note about this very simple example. First, if agent does not have high enough values associated with the social identity *Astronaut*, then he will choose to party *even though* he wants to be an astronaut, and *even though* partying means he will keep feeling the same anxiety from not being an astronaut in the future periods (which will keep the agent in a bad mood). Note that if the agent is presented with the same choice again in period $t+1$, he will again choose to party, *even though* his anxiety can be sizable. Such behavior can be called *procrastination* and is well-documented (e.g., O'Donoghue and Rabin, 2001, and also see Example 5 in Appendix A).

Second, the agent who chooses to do homework gets an additional boost to his mood, as in period $t+1$ not only does he acquire an additional feature g_1 to enjoy, but also his anxiety drops by $\tilde{v}^i_g(g_1)$. This reduction in anxiety can be related to the feeling of fulfillment that we have after accomplishing something that brings us closer to our long-term goal, or raises our status within our desired social identity (e.g., Pillemer et al., 2007).

Third, notice that the magnitude of anxiety per se does not enter the decision-making process of the affective agent, but simply persists in the background. Thus, without additional mechanisms that allow the agent to realize *why* he is anxious (see Appendix A on emotional intelligence), only particularly high values attached to an identity *Astronaut*, ones that are higher than the values of various distractions like partying, are sufficient to induce him to work towards his goal.

Regardless of the additional complications like procrastination, anxiety will motivate affective agents to acquire their chosen social identity through increasing their status within it. Many negative phenomena usually associated with status-seeking can occur as side effects (Cummins, 2005). To illustrate, notice that often there is no manual that describes what having a specific social identity exactly entails. This is the case, for example, in the famous tale of "keeping up with the Joneses" (Frank, 1985). The social identity g in question may be defined in purely relative terms; "success in life" might be relative to others whose wealth and possessions are also growing. Such an identity can create envy and positive feedback loops of status competition exactly because no one really knows what it means. When the Joneses buy a new car that is bigger than everyone else's in the neighborhood, they change the set of features related to g by adding the feature *New shiny car* to it. As a result, neighbors feel negative affect due to increased anxiety (envy) because now in order to become "successful" they also need to possess a new shiny car, which is now part of "successful" identity. This negative affect propagates through the updating mechanism to other surrounding features (e.g., the Joneses) and can have various bad consequences as was noted in many literatures (e.g., Easterlin, 2001; Santiago et al., 2011).

3.4 Identity-Based Norms of Individual Behavior

When an individual choice like above (e.g., partying instead of learning something related to becoming an astronaut) is observed by others, it can be judged on the basis of its consistency with the social identity *Astronaut*. When people see that agent i is aspiring to be an astronaut, but instead of trying to achieve this goal chooses to party, they will think that this person is doing something *inappropriate* (from the point of view of the identity *Astronaut*).[19] It is important that individual choices are judged on the basis of *some* identity and not as abstract choices, because without some benchmark (values $\tilde{v}_g$ corresponding to features within identity g) it is impossible to tell if the agent is doing something right or wrong. For example, if the agent is aspiring to become an accountant, then the choice to party instead of learning about the International Space Station is not necessarily inappropriate, since knowing about the Space Station is not seen as pre-requisite for being a good accountant.

Formally, suppose that agent j is observing that agent i has chosen to party instead of learning about the International Space Station and that i is known to aspire to become an astronaut, defined by identity g. What happens next is that agent j estimates the appropriateness of this choice using her own affective values by "simulating" how she would feel making this choice. This process is unconscious and invisible to agent j. What she does feel though is the outcome. Agent j (or her brain rather) has her own idea about what it means to be an astronaut, which is coded by values $\tilde{v}^j_g$ of features in F_g. Specifically, she has a value $\tilde{v}^j_g(g_1)$ for learning about the International Space Station as the part of preparation to become an astronaut. She also has an affective value of partying equal to s_j. Suppose first that $\tilde{v}^j_g(g_1) > s_j$, or that agent j believes that if *she* were learning to become an astronaut, then she would choose to do homework instead of partying. Now, the fact that agent i has chosen to party instead of doing homework means that *his* value of g_1 is no larger than s_j (a simple revealed preference argument). So, i's value is "off" by at least $\tilde{v}^j_g(g_1) - s_j$, which provides an (optimistic) measure of the inappropriateness of i's beliefs. As a result, agent j consciously feels *resentment towards* i of the size $\tilde{v}^j_g(g_1) - s_j$. Resentment is a negative affect $V'_t = -(\tilde{v}^j_g(g_1) - s_j)$ felt by agent j from observing the choice of agent i to party, which in her opinion violates the *identity-based norm of individual behavior*. As a result, j updates the feature f_i that codes agent i in her mind with this negative value, so that

$$v^j_{t+1}(f_i) = v^j_t(f_i) - \lambda(\tilde{v}^j_g(g_1) - s_j + v^j_t(f_i)) = (1-\lambda)v^j_t(f_i) - \lambda(\tilde{v}^j_g(g_1) - s_j) < v^j_t(f_i).$$

Thus, j's attitude towards i becomes worse than before. If we suppose to the contrary that $\tilde{v}^j_g(g_1) < s_j$, then agent j would agree with the choice of agent i, because she also thinks that partying is better than studying. In this case, no updating takes place since the action of i is

[19] This argument only holds for people who are not learning to become astronauts themselves. Those in the process of learning about what constitutes a particular social identity (young agents) might think that partying is what astronauts do and copy this behavior as part of their social learning process. We discuss this in more detail in another paper (Kimbrough et al., 2020).

consistent with what j believes it should be (see Appendix D.3 for a general specification of identity-based norms of individual behavior).

At this point it is reasonable to ask why should j feel resentment at all. Why should j care if i becomes an astronaut or not? As above, we start from the premise that the function of the social emotions is, to a substantial degree, the facilitation of mutually beneficial cooperation. What j ultimately (but unconsciously) cares about is whether i is a good reliable person who is capable of working-together or not (be it within the identity *Astronaut* or within any other). The measure of resentment of agent j is the measure of how different the beliefs of i about becoming an astronaut are from what they should be according to j. Having wrong beliefs about a social identity is bad for everyone because people with wrong beliefs (values) will not choose to do the thing they need to do when necessary (as j has seen, agent i chose to party, when she thought he should be studying).

Notice that resentment in this model endogenously generates a sort of punishment via the resulting drop in j's affective value of i. When j thinks that i is bad (j has low affective value associated with i), this makes it more likely that j will behave in a selfish or even hostile way towards i in the future (see Appendix E). Indeed, in many cultures such drop in affective value corresponds to i's "losing face" or "being dishonorable," which leads to bad treatment in the future (Henrich, 2020). This mechanism is powerful since the result is that i is effectively punished in all future interactions with j and others who have seen or heard about his choice. If i realizes that he is judged by others in this manner, he will think twice before choosing to party (at least openly, so that he is observed by j), which provides more incentive to conform to identity-based norms and which ultimately facilitates more cooperation (see Appendix A on emotional intelligence for more details).

To see how social identity can facilitate cooperation, consider the game on Figure 3. Suppose that in outcomes DC and CD instead of receiving feature-unrelated affective values, one of the players gets a feature $f \in F_g$, for which players belonging to identity g have a low affective value $\tilde{v}_g(f) = v_1 < 3$ (for example, alcohol for some religious groups). Outside g half of the agents also have the same low affective value for f, but another half has high value $v_2 > 3$ (e.g. religious groups that do not prohibit alcohol).

When both players are from g, they will always cooperate because 1) they play a Stag Hunt game (see also Figure 2) and 2) their mood is good in the presence of another member of g (high affective value τ attached to the opponent). Moreover, whenever a member of g plays D against another member of g, he may be excluded from g, because he violated the identity-based norm of individual behavior (his value of f is not low enough). As a result, group g will maintain cooperation when playing with each other. When a member of g plays with an outsider, his mood is (relatively) bad, because the outsider is not a member of g and has a low affective value (let us assume). Thus, the member of g will defect, precluding exploitation by anyone whom people from g "do not know." When two outsiders play this game, they will achieve outcome

	C	D
C	3, 3	0, $v_{1/2}$
D	v_1, 0	1, 1

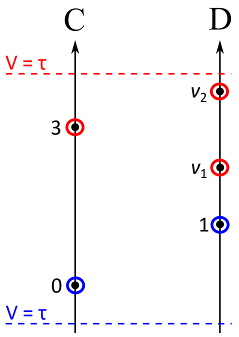


Figure 3: A game involving feature $f \in F_g$ that gives affective value v_1 to the members of g and v_1 or v_2 (shortened to $v_{1/2}$) to the outsiders.

CC in at most 50% of cases. CC only occurs if both players value the feature at v_1 and if they like each other enough (high τ), which might not be the case because outsiders have less in common with each other than members of g.

This argument suggests why social identity can amplify cooperation. This boost in cooperation rates is exactly due to the fact that members of a social identity keep track of the affective values associated with it (rituals), react negatively to norm-violations (identity-based norms of individual behavior), and (potentially) do not trust strangers.

3.5 Affective Decision-Making in Markets

Understanding market outcomes is the bread and butter of economics; thus it is worthwhile to consider how affective decision-makers behave in markets. We consider two simple examples of market exchange. Our first example highlights that affective agents find it more difficult than rational agents to exploit the benefits of simple bilateral exchange. Our second example shows how bidding behavior in first price auctions will be influenced by mood, potentially yielding cycles of bid escalation. Thus we illustrate how affective decision-making can also account for some market pathologies.

Example 3. A Goods Exchange Game. Suppose that Player 1 possesses some quantity of good A that he does not derive any affective value from and that Player 2 possesses some quantity of good B that she also does not care about. However, both players derive affective value of $x > 0$ from consuming the good of another player. In this case, the players can exchange their goods or they can choose to not do anything. The left panel of Figure 4 illustrates.

If players are selfish and rational, both should play Enter as long as they put non-zero probability on the other playing Enter, which is not an unreasonable thing to believe. Thus, the standard economic intuition holds.

Now let us consider affective players playing this game. Suppose that Player 1's mood is $V = \tau$, where τ is the affective value that he attaches to Player 2. Then action Not leads to a sure outcome 0. The action Enter leads to a lottery $\langle 0, \{x - c, -c\}\rangle$ with two possible outcomes $x - c$

	Enter	Not
Enter	x, x	0, 0
Not	0, 0	0, 0

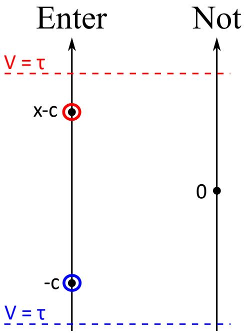


Figure 4: **Left Panel.** A Goods Exchange Game. **Right Panel.** The choices in the game transformed by an affective player to the decision problem under uncertainty.

and $-c$ that take into account some cost of uncertainty $c > 0$. To choose between these actions, Player 1 will use mood affiliation and will decide that outcome $-c$ of the lottery will realize if V is closer to $-c$ than to $x - c$. This happens when $V = \tau < \frac{x-2c}{2}$. Thus, under this condition, Player 1 will choose to do nothing. The condition $\tau < \frac{x-2c}{2}$ can hold for positive τ given some large enough x and small enough c. To put it differently, in order for Player 1 to choose Enter, the inequality above should *not* hold, which happens if τ is high enough. The same logic can be applied to Player 2. □

This example is remarkable for at least three reasons:

First, there is a stark difference between the behavior of (selfish) rational agents and affective agents. The former always "cooperate" in this game by choosing Enter. The latter enter *only if* they "trust" the other player enough, which as noted above, will tend to depend on their similarity to one another. This similarity requirement may have far-reaching consequences, since the prospect for mutually beneficial exchange often arises because two agents are *different*. It is true that affective agents have a built-in tendency to bridge a trust gap in order to gain from exchange; if I have some feature that you want, then that will make you positively disposed to me (and vice versa). However, if in the course of acquiring that feature, your affective values have come to diverge widely from mine, then we will find it hard to overcome the lack of similarity in order to benefit from exchange. The proverbial distrust between, say, rustic farmers and city-slicker bankers provides an example of the kind of friction we have in mind.

Second, this simple case ignores any additional outside influences on the mood of affective agents. If such additional influences exist (for example, Player 1 has a stomach ache), then cooperation will be less likely.

Third, the mood-dependence of affective agents' choice creates a layer of strategic uncertainty that anyone who wants to deal with them will necessarily face, which makes cooperation with affective agents difficult. In order to understand whether an affective agent will enter a mutually profitable exchange of goods, one needs to consider all factors that can put the agent in a bad mood, which can be anything. This may be one reason that business relationships are often established and solidified with entertainment.

In sum, this example demonstrates that affective agents' willingness to trade doesn't just depend on the potential gains; they also need to be in the mood to cooperate even in the simplest forms of obviously mutually profitable interactions. Our assumption that the resolution of strategic uncertainty about an agent depends on one's affective value of that agent, which in turn depends on one's similarity to the agent, means that affective agents will tend to cooperate mostly with others they trust a lot (high τ). Those with high similarity and high trust are usually kin or friends, and thus affective agents may find it hard to get along with strangers. This tendency can create nepotistic networks and corresponding market inefficiencies in places where normally economists would never suspect that anything can go wrong (see e.g., Perez-Alvarez and Strulik, 2021). The further implication is thus that well-functioning impersonal markets must be built on more than just affective decision-making. One important role of the rational component of decision-making is thus in moral reasoning that counsels treating others with the same respect, regardless of their (dis)similarity. Such principles are necessary to mitigate the parochial tendencies of affective decision-making.

Example 4. First-Price Auctions. Consider a simultaneous-move sealed-bid first-price auction played by two affective players 1 and 2, who have positive affective values v_1 and v_2 of the item being sold. Consider Player i and let his affective value of the other player be τ_{-i}. Suppose that Player i's mood is $V_i = M_i + \tau_{-i}$, which is the sum of some outside influences M_i and the influence of the other player's affective value.

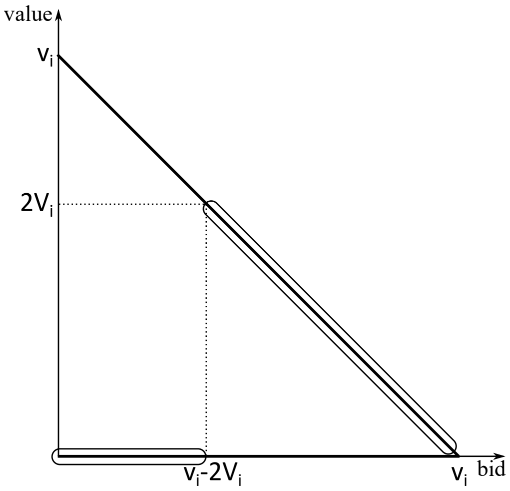


Figure 5: Choice of a bid by Player i in the first-price sealed-bid auction.

From the perspective of Player i, the choice of any bid $b_i \geq 0$ leads to a lottery with two outcomes because i can either win the object or not. Given that any choice leads to such a lottery let us assume for simplicity that the cost of uncertainty c is constant across all choices and thus does not influence the decision-making process (assume $c = 0$). Then, a bid b_i gives either affective value $v_i - b_i$, if i wins the object, or 0 if he does not. Figure 5 shows possible bids on the x-axis and for each bid, the values $v_i - b_i$ and 0 as thick lines.

Now, for the most interesting case, suppose that $v_i > 2V_i$. Then, mood affiliation dictates that Player i will believe that the uncertainty for low bids $b_i \in [0, v_i - 2V_i)$ will realize in losing the auction, because for these bids V_i is closer to 0 than to $v_i - b_i$ (Player i feels that low bids are not going to win the auction). For higher bids $b_i > v_i - 2V_i$, Player i will mood-dependently believe that he will win the auction. These beliefs are shown with ovals in Figure 5. Given this, Player i will choose the bid that in his opinion gives the highest affective value, namely $b_i^* = v_i - 2V_i$. This says that the better the mood, the lower the bid and suggests that, in case players have similar values of the object, the player with the worst mood will win the auction. □

What is most interesting about this result is that social attitudes towards other players influence the optimal bids through mood affiliation. If the two players are friends and like each other (high τ_{-i}, good mood), they will bid low, which happens when people auction off some items among neighbors in their community (Smith, 1990). Conversely, if players do not like each other and are, for example, competitive (low τ_{-i}, bad mood), then they will bid high, thus driving the prices up. This logic also suggests that any outside events that change M_i, for example, some bad news that make people unhappy, will drive the prices up and vice versa.

Here we have considered a static game and assumed selfishness. This setup, however, is easy to generalize to dynamic auctions where people bid at any time during a bidding session. In this case, when a player makes a bid, but does not win because someone else places a higher bid (while the auctioneer says "one, two, three") he feels disappointment that lowers his mood. Lower mood, in its turn, makes him bid more. This mechanism can produce sequences of higher and higher bids, reminiscent of "auction fever" (e.g. Ku et al., 2005). Notice as well that disappointment from not winning during the auction will lower the affective value τ_{-i} of the other player through updating. If this weight becomes negative, then players will start bidding above their valuation v_i in order to make the other player worse off if they care about dissatisfactions of others (see Appendix E), thus creating a winner's curse. These simple examples highlight that our framework can be fruitfully employed to analyze a wide range of economic interactions.

4 Conclusion

In this paper, we outline a model of boundedly-rational affective agents rooted in biological constraints. We start from the premise that human decision-making is the product of both reason and emotion, and we develop a simple model of the emotional mind. We refer to agents in our model as affective agents, since their choices are driven by affect. Affective agents do not reason rationally, but rather make choices based on their affect towards the perceptible features in their present environment. Via simple reinforcement learning, the affective values of these features update over time in response to past experience. Under certainty, such agents simply choose the outcome with the highest affective value (all features, "relevant" or not, considered). Under uncertainty, we provide a model in which affective agents choose based on a kind of motivated "reasoning" that we refer to as mood affiliation. Since simple reinforcement learning precludes explicit reasoning about probabilities, we suggest that affective agents transform uncertainty into certainty by assuming that uncertainty will resolve in a manner most consistent with their present "mood".

This simple model reproduces a striking number of behavioral phenomena documented in the laboratory and field (e.g. prospect theory, procrastination, addiction, etc.), and yet, assuming that agents' moral sentiments are tuned to identify and prefer "like-minded" others, the model also predicts affective agents are also strikingly capable of coordinating and cooperating – since they will tend to assortatively match over time. Such assortativity naturally results in herd-behaviors and in a sort of in-group bias, and it yields an intuitive notion of identity, defined as shared affective values. With an added bit of abstraction, identity-based cooperation and identity-based norms of behavior can be rooted in having (and preserving) shared affective values.

Nevertheless, we also show that there are limits to the capacity of such agents to cooperate, and thus we argue that a primary value of the rational component of the mind is its power to overcome the narrowness of affective decision-making. In practice, moral sentiments are paired with moral reasoning, which counsels us to consider not only how an action will make us feel but also how it will make others feel. See (Kimbrough and Vostroknutov, 2020, 2021) for models of moral reasoning that complement this account of affective decision-making.

References

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *Quarterly Journal of Economics*, 115(3):715–753.

Bénabou, R. (2015). The economics of motivated beliefs. *Revue d'économie politique*, 125(5):665–685.

Bogacz, R. (2007). Optimal decision-making theories: linking neurobiology with behaviour. *Trends in Cognitive Sciences*, 11(3):118–125.

Brewer, M. B. (1991). The social self: On being the same and different at the same time. *Personality and social psychology bulletin*, 17(5):475–482.

Brighetti, G., Lucarelli, C., and Marinelli, N. (2014). Do emotions affect insurance demand? *Review of Behavioral Finance*.

Buckner, C. and Garson, J. (2019). Connectionism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2019 edition.

Callard, A. (2018). *Aspiration: The agency of becoming*. Oxford University Press.

Chen, Y. and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–57.

Chudek, M. and Henrich, J. (2011). Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in cognitive sciences*, 15(5):218–226.

Chwe, M. S.-Y. (2013). *Rational Ritual*. Princeton University Press.

Cummins, D. (2005). Dominance, status, and social hierarchies. In Buss, D. M., editor, *The Handbook of Evolutionary Psychology*, chapter 20, pages 676–697. John Wiley & Sons, Inc.

Dayan, P. and Abbott, L. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press.

De Waal, F., Macedo, S., and Ober, J. (2006). *Primates and philosophers: How morality evolved*. Princeton University Press.

Dennett, D. C. (1991). *Consciousness explained*. Boston: Little, Brown and Co.

Easterlin, R. A. (2001). Income and happiness: Towards a unified theory. *The economic journal*, 111(473):465–484.

Erikson, E. H. (1968). *Identity: Youth and crisis*. Number 7. WW Norton & company.

Frank, R. H. (1985). *Choosing the Right Pond: Human Behavior and the Quest for Status*. New York: Oxford University Press.

Fudenberg, D. and Levine, D. K. (2006). A dual-self model of impulse control. *American economic review*, 96(5):1449–1476.

Gllimcher, P. W., Camerer, C. F., Fehr, E., and Poldrack, R. A., editors (2009). *Neuroeconomics: decision making and the brain*. Elsevier Inc.

Henrich, J. (2015). *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Henrich, J. (2020). *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*. Penguin UK.

Isoni, A. and Sugden, R. (2018). Reciprocity and the Paradox of Trust in psychological game theory. *Journal of Economic Behavior & Organization*.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1986). Fairness and the assumptions of economics. *Journal of business*, pages S285–S300.

Kahneman, D. and Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, 39(4):341.

Kimbrough, E., Tremewan, J., and Vostroknutov, A. (2020). A theory of descriptive norms. Mimeo, Chapman University, Higher School of Economics, Moscow and Maastricht University.

Kimbrough, E. and Vostroknutov, A. (2020). A theory of injunctive norms. mimeo, Chapman University and Maastricht University.

Kimbrough, E. and Vostroknutov, A. (2021). Axiomatic models of injunctive norms and moral rules. mimeo, Chapman University and Maastricht University.

Ku, G., Malhotra, D., and Murnighan, J. K. (2005). Towards a competitive arousal model of decision-making: A study of auction fever in live and internet auctions. *Organizational Behavior and Human decision processes*, 96(2):89–103.

Laland, K. N. (2018). *Darwin's unfinished symphony: how culture made the human mind*. Princeton University Press.

Mousa, S. (2020). Building social cohesion between christians and muslims through soccer in post-isis iraq. *Science*, 369(6505):866–870.

Nesdale, D. and Flesser, D. (2001). Social identity and the development of children's group attitudes. *Child development*, 72(2):506–517.

O'Donoghue, T. and Rabin, M. (2001). Choice and procrastination. *The Quarterly Journal of Economics*, 116(1):121–160.

Pauly, M. V. and Kunreuther, H. (2019). Responses to losses in high-deductible health insurance: persistence, emotions, and rationality. *Behavioural Public Policy*, 3(1):72–86.

Perez-Alvarez, M. and Strulik, H. (2021). Nepotism, human capital and economic development. *Journal of Economic Behavior & Organization*, 181:211–240.

Pickett, C. L., Bonner, B. L., and Coleman, J. M. (2002). Motivated self-stereotyping: heightened assimilation and differentiation needs result in increased levels of positive and negative self-stereotyping. *Journal of personality and social psychology*, 82(4):543.

Pillemer, D. B., Ivcevic, Z., Gooze, R. A., and Collins, K. A. (2007). Self-esteem memories: Feeling good about achievement success, feeling bad about relationship distress. *Personality and Social Psychology Bulletin*, 33(9):1292–1305.

Preuschoff, K., Bossaerts, P., and Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron*, 51(3):381–390.

Robson, A. J. (2002). Evolution and human nature. *Journal of Economic Perspectives*, 16(2):89–106.

Santiago, C. D., Wadsworth, M. E., and Stump, J. (2011). Socioeconomic status, neighborhood disadvantage, and poverty-related stress: Prospective effects on psychological syndromes among diverse low-income families. *Journal of Economic Psychology*, 32(2):218–230.

Smith, A. (1759). *The Theory of Moral Sentiments*. Liberty Fund: Indianapolis (1982).

Smith, C. W. (1990). *Auctions: The social construction of value*. Univ of California Press.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, Mass.: MIT Press.

Tajfel, H. and Turner, J. (1986). The social identity theory of intergroup behavior. In Worchel, S. and Austin, W., editors, *The psychology of intergroup relations*, pages 7–24. Chicago: Nelson-Hall.

Uexküll, J. v. (1926). *Theoretical biology.* Harcourt, Brace & Co.

Veblen, T. (1899). *The Theory of Leisure Class: An Economic Study of Institutions*. London: Allan and Unwin.

Weinreich, P. and Saunderson, W. (2005). *Analysing identity: Cross-cultural, societal and clinical contexts*. Routledge.

Wexler, B. E. (2006). *Brain and Culture: Neurobiology, Ideology, and Social Change*. A Bradford Book.

Appendix (for online publication)

A Emotional Intelligence

If we take the behavior of affective agents at face value, it becomes apparent that the degree of "boundedness" of their rationality is rather extreme. Specifically, affective agents do not understand why they feel what they feel: they only have conscious access to their mood V_t and not to the affective values $v_t(f)$ of the individual features that are aggregated in it, nor can they grasp the aggregation of values (through G) or value updating process, which can lead to irreality and "exaggerated" affective values. Affective agents do not understand why they believe what they believe (through mood affiliation), which can lead to over/underconfidence and "biases" in choices under uncertainty. These characteristics also spill over to social identity, which results in 1) inability of affective agents to realize why they feel bad when their choices do not lead to achieving their goals (e.g., procrastination); 2) inability to realize why they feel bad about someone who has different features from them or who did something "wrong" (e.g., discrimination); and 3) inability to understand that others might judge them for or have moral sentiments about their own behavior (e.g., getting into conflicts).

From a rational economic perspective, it may seem that assumptions like this are unnecessary since everyone is (seemingly) capable of understanding the sources of their feelings, can understand why they believe something, know what they want to achieve, can predict that others will judge them, etc. Nevertheless, common sense, folk intuition, and vast strands of literatures in psychology and behavioral economics suggest countless examples of such "irrationalities" observed in non-pathological, adult human beings (e.g., Loewenstein, 2000; Salovey and Grewal, 2005). At the same time, we are also well aware that people are capable of understanding their own emotions, controlling them, and making better decisions as a result (e.g., fighting procrastination, addiction, or own stereotypical beliefs). For biological reasons, we believe that the affective system mapped out in this paper indeed functions in the way it is described (extremely boundedly rational) and that all these additional capabilities are built *on top* of it in the form of *emotional intelligence*, or the attempts by the (presumably rational) brain to control its own emotional urges. Therefore, in this appendix we describe how emotional intelligence can be conceptualized within our framework, since it is an integral part of human nature and behavior.

We decided to abstain from providing a full mathematical description of an emotionally intelligent affective agent, which is a task for the future research. Instead, we will give several examples of common "irrational" behaviors that affective agents exhibit and will suggest how they can overcome these irrationalities by learning something about themselves or using rational reasoning (Kimbrough and Vostroknutov, 2020, 2021, further KV).

We start by recalling Example 1 where an affective agent, who caught cold in Paris, had negative feelings about the city and France in general due to the updating of the corresponding affective values. This can lead to irreality and unwanted consequences in the future (e.g., not taking a job in Paris). We argued that such updating was "unreasonable" since negative emotions related to having a cold had nothing to do with actual (affective) properties of Paris or France. An emotionally intelligent agent could prevent irreality from taking place if his brain could consciously *interfere in the process of updating of affective values*. In other words, such agent can realize that feeling negative affect due to cold has nothing to do with Paris or France, prevent the update from happening, and thus retain positive impression of the city, which he might actually like a lot. Physiologically, this can be done by sorting out own emotional experiences, making a connection between the cold and the bad affect that it causes, and correcting the update of the values of other features. Such act requires certain level of access to own internal emotional processes, which (we presume) can be gained by practicing self-reflection and introspection (e.g., Herwig et al., 2018).

This example of emotional intelligence relates to a single moment in time when the agent consciously corrects his feelings, which already can have a positive effect on his well-being (e.g., Schutte et al., 2002).

Another form of emotional intelligence involves corrections of "compulsive behaviors" that affective decision-making system can generate in some specific contexts. We illustrate with an example.

Example 5. Depression and Developing an Addiction. Suppose an affective agent is in a fixed environment determined by some features with aggregated value 0 (any other value would work as well). He is aspiring to become an astronaut (social identity g) and is looking for a job related to his profession. His mood in period t is given by $V_t = -\phi V_g$, which is negative because he feels dissatisfaction $V_g - 0$ from not possessing any features from the set F_g associated with being an astronaut (see Section 3.3).

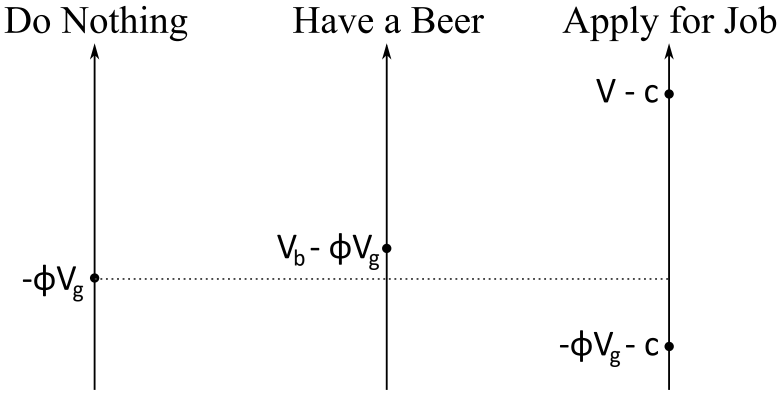


Figure 6: A choice faced by an affective agent looking for job.

Figure 6 shows three choices that the agent can make. He can do nothing, in which case his affect V_t' is equal to his current mood $-\phi V_g$; he can have a beer, in which case his affect is $V_b - \phi V_g$ equal to his current mood plus a feature-unrelated value of beer $V_b > 0$; or he can apply for a job, which is an action with uncertain outcome: if he gets a job he will get some high value $V \gg V_b$, if he does not get a job, then he receives no additional affective value. There is also some cost of uncertainty $c > 0$ associated with applying for a job (taking an uncertain action). This specifies the decision problem under uncertainty (Section 2.2). With mood affiliation, the agent will believe that applying for a job is going to result in him not getting it, because this consequence is closer in value to his currently bad mood than the consequence of getting a job. So, out of the three available actions, the agent chooses the one with the highest affective value, which is having a beer.

The next day, in period $t + 1$, the pleasurable effects of beer vanish and the agent is back to the exact same choice as in the previous period. Following the same logic, he will again choose to have a beer, and this will continue ad infinitum. As a result, the agent is depressed (always in a bad mood) and becomes addicted to alcohol from constantly drinking beer. □

This example demonstrates a typical *dynamic choice* problem faced by an affective agent that cannot be resolved in an optimal way (we assume that, in the end, the agent prefers to become an astronaut rather than an unemployed alcoholic). In this class of problems, the agent who is originally in a bad mood (precursor of depression) can either attempt to do something useful for his future, which is associated with some uncertainty, or he can choose to enjoy something instantaneous and certain, but useless (or damaging) in the long run. Due to mood affiliation, the choice in such cases will generally go in the direction of small but useless rewards (e.g., beer), which will typically be associated with negative long-term consequences (e.g., addiction). When left unchecked, affective decision-making in such contexts can result in: procrastination (leading to depression), overeating (leading to obesity), the phenomenon of *hikikomori* (Kato et al., 2019), hoarding, and various other compulsive behaviors.

Emotional intelligence can break this behavioral pattern in at least two ways. First, the agent can realize that his mood-dependent beliefs about the chances of getting a job are not reflective of reality (we assume that the chances of getting a job are related to agent's characteristics, like skill, rather than his mood). So, his brain could *interfere in the process of constructing beliefs about the resolution of uncertainty*. If the moody agent changes his beliefs to more positive ones, he can choose applying for a job instead of having a beer, thus breaking the vicious cycle. Second, the agent can realize the dangers of constantly drinking beer and correct the corresponding affective value. This can be done by *focusing on the long-term consequences instead of instantaneous ones*, which is a skill his brain can develop through constructing future value (done in ventromedial Prefrontal Cortex, Benoit et al., 2014).

Both of these emotionally intelligent strategies can be thought of possessing elements of rational reasoning (KV), which in its essence is different from affective decision-making exactly because it is focused on the future rather than the present. A class of important decision-making phenomena are related to this as the next example illustrates.

Example 6. Future Consequences: Health, Education, and Environment. One property of the affective system that makes it very different from the rational (cognitive) one is that affective values are experiences felt directly through the senses (smell, taste, etc.) or their immediate interpretations by the brain (e.g., fear, joy). This implies that affective decision-making is not oriented towards computing or evaluating the future consequences of experiences, but is rather focused on simply feeling something right now. Given this biological constraint it is not surprising (from the rational perspective) that affective decision-making can be "myopic." The examples of such behaviors abound: smokers keep smoking even when they understand that this behavior seriously damages their health in the future; the same holds for myriad of other unhealthy choices that many people make (food, drugs, etc.). The reason is of course that the choice to do something unhealthy is made "affectively" taking into account only the immediate pleasure of the experience and not its long-term consequences. In other words, affective system cannot take into account the long-term consequences of these decisions because it *cannot feel them*. There is simply no bodily sense that tells us that we are doing something that will have bad consequences in the future. Thus, from the affective point of view such consequences do not exist.

The same logic applies to decisions when, instead of having an instantaneous pleasure today with negative consequences later (e.g., unhealthy eating), the decision is to pay a little affective cost today to have a large reward later. This relates for example to the decisions to get educated. Education is difficult and involves either direct unpleasant experiences (e.g., learning mathematics) or forgoing pleasant experiences (studying instead of partying). Affective system will be reluctant to make such sacrifices when it cannot directly feel the future reward (a degree and a good employment). From this follows that affective system will rarely choose to invest today into some abstract future benefits. In fact, in Section 3.3 we discussed the possibility that a special mechanism has evolved that makes affective agents feel anxiety when they are aspiring to acquire some social identity, which is exactly a solution to this problem in a specific context (of acquiring a skill).

However, even though the affective system might possess such a specially designed contraption to push affective agents to acquire skills, this mechanism is context-specific and does not work for other situations. This relates to problems with the environment. In the times of global warming, everyone needs to make little sacrifices today for the sake of future generations. The reluctance with which people make such choices can be explained by the idea that affective system does not recognize such sacrifices as good decisions because it does not feel alarmed about the dangers of not protecting the environment. As a result, environmental policies are not being implemented no matter how hard some individuals or even countries try to push them. □

It seems that emotionally intelligent strategy that can help with resolving the problems with future-oriented decision-making is *suppression* of affective decision-making all together in favor of rational reasoning. It is not impossible actually that rational reasoning has evolved as a specific device to solve future-oriented problems. After all, it can hardly be denied that such ability would give our ancestors a serious survival advantage (Suddendorf and Corballis, 2007). The way this can be implemented in the brain is the *construction of value* that presumably takes place in the prefrontal cortex (O'Doherty et al., 2021). Rational reasoning might work by substituting the affective values with "abstract values" or utilities obtained through logical implications involving the same affective values only felt later. For example, if I drink beer all the time, I will become unhealthy, which will then prevent me from getting a good job, which will make me (affectively) unhappy due to the absence of money. If the brain can use this logic and correspondingly change the affective value of beer today, it can solve an otherwise impossible future-oriented problem.

The mechanisms of emotional intelligence working for individual decisions can equally be used to resolve problems with social interactions. For example, as was mentioned in Section 3.1, affective agents form affective values of strangers (their social weights) by taking into account only their observable characteristics (e.g., skin color, sex, clothes). While in some specific circumstances this can be a viable strategy, in many others it can lead to discrimination, exclusion, and bad societal outcomes. Emotional intelligence can help with overcoming the urge to judge others by their looks alone. This is achieved, as mentioned above, by conscious interfering with the process of aggregation of affective values of the current features. Interestingly, evidence exists (e.g., Gamberini et al., 2015) that under pressure or stress people become more discriminatory towards unfamiliar strangers, which is suggestive of such mechanism. Similar argument can be made about judging or forming beliefs about actions of others in games, which we illustrate with an example (check Appendix E for relevant definitions).

Example 7. Prisoner's Dilemma. Consider a Prisoner's Dilemma game shown in the left panel of Figure 7. Similarly to Example 3, the middle panel of Figure 7 depicts the representation of the choice between Cooperate (C) and Defect (D) by any of the two affective players in this game. We assume that player's own social weight is 1 and the other player's is $\tau \in [0, 1]$ and that his propensity to follow norms is $\phi \geq 0$. Given that both actions now have uncertain outcomes, we set the cost of uncertainty c to zero, under a simplifying assumption that it is the same for both actions C and D and thus cancels out and does not influence the choice (similarly to Example 4).

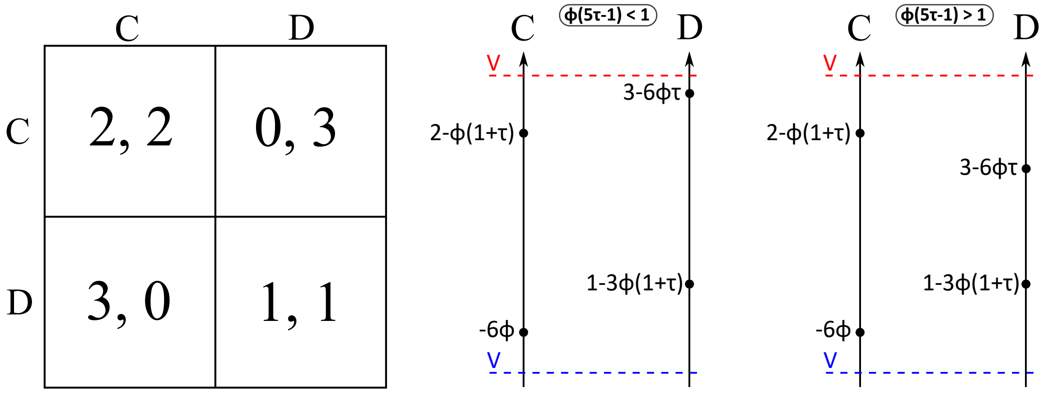


Figure 7: **Left Panel.** Affective values in a Prisoner's Dilemma game. **Middle Panel.** The representation of the game as the decision problem under uncertainty when $\phi(5\tau - 1) < 1$. **Right Panel.** Same as the middle panel only with $\phi(5\tau - 1) > 1$.

The values in the middle panel of Figure 7 are transformed with normative terms that incorporate dissatisfactions of both players given all possible outcomes in the game (KV). Under the condition $\phi(5\tau - 1) < 1$, which holds when both ϕ and τ are low (the player is selfish and/or does not care about the other player), the values from both actions are ordered in exactly same way as in the standard Prisoner's Dilemma on the left.[1] Namely, given any action of the other player, defecting gives more affective value than cooperating. Notice that in this case if the player is in a good mood V (shown as a red dashed line in the middle panel), he will think that the uncertainty will resolve in value $2 - \phi(1 + \tau)$ from playing C and in value $3 - 6\phi\tau$ from playing D, which is equivalent to believing that the opponent will cooperate. Since

[1] The condition $\phi(5\tau - 1) < 1$ is a rearrangement of the condition $3 - 6\phi\tau > 2 - \phi(1 + \tau)$, or when the norm-dependent affective value from defecting while other player cooperates is higher than the value from cooperating while the other player cooperates (as it is in the standard Prisoner's Dilemma).

the latter value is higher than the former, the player will defect. Defection is similarly the best choice when the mood V is bad (the blue dashed line), which is equivalent to believing that the opponent will defect. To summarize: when the player is in a good mood, he will believe that the opponent will cooperate and when the player is in a bad mood, he will believe that the opponent will defect. In either case, the best choice is defection given that ϕ and τ are low.

This logic leads to an important additional conclusion. Notice that the mood V includes τ as one of the terms since the opponent is one of the features surrounding the player. Therefore, it is possible that for low enough τ—for example, because she is a stranger with some undesirable features—the player will always be in a bad mood and thus *will believe that the opponent will defect*. This demonstrates how low affective value τ associated with a stranger can lead to a belief that she will defect in a Prisoner's Dilemma because her mere presence puts the player in a bad mood. This conclusion about the behavior of the opponent can be based on absolutely no relevant information about what she actually is planning to do.

Despite this grim prospect, cooperation is still possible in this game as shown in the right panel of Figure 7. Under the condition $\phi(5\tau - 1) > 1$, or when ϕ and τ are high enough (the player is norm-following enough and has a high enough regard τ for the opponent), the order of the norm-dependent affective values $2 - \phi(1 + \tau)$ and $3 - 6\phi\tau$ switch order. Now, if the player is in a good mood (the red dashed line in the right panel), he will choose to cooperate because he will believe that the opponent will also cooperate and because now the normative value of mutual cooperation is high. Given that player's mood V depends on the value of τ, which is now high, the player will be in a good mood upon meeting the opponent and cooperate with her. This is the mirror opposite of the situation describes above. □

What this example demonstrates is that affective players can choose to cooperate or defect in Prisoner's Dilemma simply because of the feelings they have towards the opponent (good or bad). These feelings, in their turn, can be based on scant and irrelevant information (for example the way someone looks), which can eventually lead to bad decisions, be they cooperative or not. Emotional intelligence can again help to resolve these problems if the affective player overcomes his instinctive urge to act in the way described in the example above and considers more relevant characteristics of the opponent or more relevant information about her.

Similar to this, involving emotional intelligence to understand that *others* might base their decisions to cooperate or defect on irrelevant features can improve decision maker's welfare if he takes this information into account. For example, an emotionally intelligent agent might realize that not wearing a formal suit on a job interview is a bad idea even if he personally hates wearing suits and even if not wearing suits is a part of his social identity. This also refers to our previous discussion of partying while aspiring to become an astronaut (Section 3.4). Here, an affective agent might realize that partying can make a bad impression on others and choose to study instead.

B Certain and Uncertain Parts of Lotteries

In Section 2.2 we have introduced the notion of a lottery $L = \langle s, U \rangle$ that consists of a *certain part*, represented by some affective value s, and an *uncertain part*, given by a collection of affective values U. The idea behind this construction is that if the lottery is chosen then the certain part is experienced for sure as well as one of the elements $s' \in U$. The affective value of the lottery then is the sum of the affective values $s + s'$. From the expected utility perspective, such division is unnecessary, since rational agent experiences some utility $u(s + s')$, so the certain part just gets added to the uncertain part. However, for the affective agent such distinction might be important. In this appendix we provide some arguments supporting this modeling choice.

Our first argument why certain and uncertain parts should be distinguished relates to the physical nature of features that people encounter during their lives and attach affective values to. The features that surround an affective agent in any environment can be divided into two broad classes: those that are "constantly present" in this environment and those that are "changing." For example, when you choose to walk to the beach expecting to meet some of your friends there, you choose a lottery that includes a certain outcome "beach," which consists of fixed and certain features (sand, sea), and some uncertain outcomes including features related to your friends who may or may not show up. You know that even if no one except you comes to the beach, you will still be able to enjoy the "constant" features out there (e.g., swim in the sea). Another class of situations where the division into certain and uncertain parts makes a difference is choice with intertemporal components. As in the insurance example in the main text (Section 2.2), sometimes you need to pay for something today, with benefits accruing in the future (or to not pay today with no benefits in the future). Such situations clearly demarcate what are the certain consequences of this lottery (paying or not paying today) and the uncertain ones (the future events). Given that most physical environments can be easily divided into certain and uncertain components like those described above and given that affective system needs to devote a significant amount of energy (neural calculations) to figure out how uncertainty will resolve, it is biologically sensible to assume that certain and uncertain pieces of the environment are treated somewhat separately. Such division helps the affective system to choose better in uncertain conditions by singling out the certain parts and thus diminishing the costs of uncertainty. Therefore, we believe that perceiving lotteries as having certain and uncertain parts is reasonable given physical and biological constraints of the affective system.

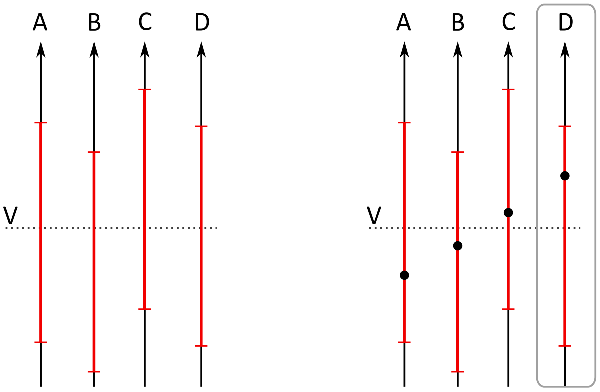


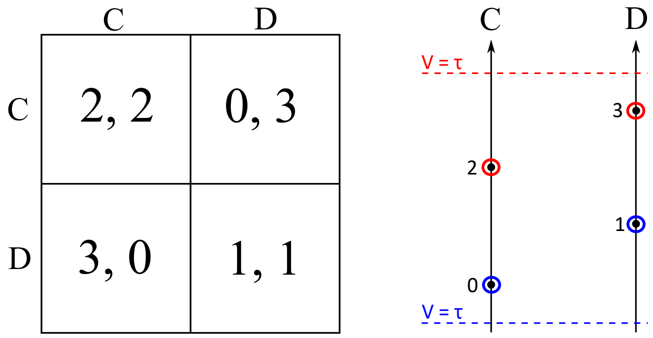
Figure 8: **Left Panel.** Choice among lotteries with certain parts defined. **Right Panel.** Choice among lotteries without certain parts defined.

Our second argument relates to the behavior of the affective agents in very uncertain environments. Imagine that the agent is choosing among lotteries A, B, C, and D shown in the left panel of Figure 8. The red intervals represent all possible affective values of the lotteries (certain and uncertain parts summed up) on a continuum. This is an example of a "large" uncertainty for the affective agent because mood affiliation is not of a great help with figuring out which lottery is better. Each lottery has an outcome that is exactly consistent with the current mood for a wide range of its possible values. Thus, if the affective agent tries to choose among these lotteries as described in Section 2.2 and Appendix D.2, he will not be able to choose, or will choose randomly because mood affiliation would generate the same belief for each lottery. This is not a good prospect from the evolutionary perspective as such indecisiveness is not survival-enhancing and can cost the agent dearly.

However, in this special case of large uncertainty for each lottery (which we do not discuss in the main text), the affective agent can simplify his decision-making by considering only the certain parts of the lotteries. In the right panel of Figure 8, the certain parts of lotteries *A*, *B*, *C*, and *D* are marked with black circles. Given this information and the fact that mood affiliation is useless to help the agent make a choice in these specific circumstances, it is plausible that affective agent might just ignore the uncertainty all together and choose among certain parts of the lotteries as if uncertainty is not there. After all, when uncertainty is large, the agent might "predict" that it always will resolve same way everywhere and simply ignore it. In this case, the agent will choose the lottery *D*, which has the highest certain part. Even though this alternative mechanism might not be particularly attentive to the specifics of the uncertainty, it is nonetheless better than random choice, since it at least maximizes among the certain outcomes that the agent can get.

The arguments above provide some biological and evolutionary reasons why lotteries can be perceived by affective agents as divided into certain and uncertain outcomes. There exists some circumstantial evidence of uncertainty-ignoring behavior implied by this idea (e.g., Callen et al., 2014). However, more specific experiments are needed to properly test the arguments presented here. We leave it for the future research.

C Additional Graphs



	C	D
C	2, 2	0, 3
D	3, 0	1, 1

Figure 9: In the Prisoner's Dilemma game (Left Panel) agent i with mood $V = v^i(f_j) = \tau$ plays D for any V except in the interval $(1, 2)$. We ignore the cost of uncertainty assuming it equal for both actions.

D General Definitions

D.1 Occurrences and Lotteries

In this appendix we lay down the basic formalisms used in the subsequent appendices to construct general definitions related to our framework. We start with the definition of *occurrence*. Occurrence is a collection of features together with a feature-unrelated affective value that an affective agent might believe he (or someone else) will obtain after choosing some action or after something else will have happened (a move of Nature, etc.). Formally, an occurrence is an element $r \in \mathcal{O} := 2^{\mathcal{F}} \times \mathbb{R}$. We can write $r = (F_r, s_r)$, where F_r is a collection of features from the set $\mathcal{F}$ and $s_r \in \mathbb{R}$ is an affective value. For example, the agent can expect that if he goes to the beach he will experience the set of features $\{Sea, Sun, Sand\}$ and the feature-unrelated value -1 from sand-burnt feet. We can say then that the agent considers an occurrence $r = (\{Sea, Sun, Sand\}, -1)$, where $F_r = \{Sea, Sun, Sand\}$ and $s_r = -1$.

One characteristic of occurrences is that they can occur together with other occurrences. We can define an operation $\cup$ defined on the set $\mathcal{O}$ of occurrences as follows. For any two occurrences $r, w \in \mathcal{O}$ let $r \cup w = (F_r \cup F_w, s_r + s_w)$. In words, the co-occurring occurrences simply present the agent with the union of their features and the sum of their feature-unrelated affective values. For example, the occurrence "beach" described above can co-occur with the occurrence "friend," who may also come to the beach (or not).

When reasoning about his choices that lead to some occurrences, the agent has to transform occurrences into affective values. This is done by means of the affective values of features v_t defined in period t and the aggregation function G described in Section 2. Specifically, the aggregate affective value of occurrence $r \in \mathcal{O}$ is

$$G_t(r) := G(F_r; v_t) + s_r = \sum_{f \in F_r} v_t(f) + s_r.$$

Here, the new short notation $G_t(r)$ presumes that the current affective values are v_t. Given this notation we can define a *decision problem under certainty* (also defined in Section 2.1) as a choice among actions in some set A that lead to some occurrences $r(a)$. Thus, the agent is solving $\max_{a \in A} G_t(r(a))$.

Next we provide a general definition of a *lottery* as follows. A lottery $L = \langle r, R \rangle$ is a tuple consisting of a *certain occurrence* $r \in \mathcal{O}$ and a collection of *uncertain occurrences* $R \in 2^{\mathcal{O}}$. If R is an empty set, then lottery L becomes equivalent to the sure occurrence $L = \langle r, \varnothing \rangle := r$. The idea is that if L is chosen or otherwise obtained then the agent will experience the compound occurrence $r \cup r'$, where $r' \in R$ is one of the uncertain occurrences that actually takes place when uncertainty is resolved. For the future use, we denote by $F_L := F_r \cup (\cup_{w \in R} F_w)$ the set of all features that can occur in lottery L.

To give an example, suppose that the agent (Zak) is heading to the beach where he plans to meet his friends Ann and Bob. This means that he expects to experience a certain occurrence "beach" described by $r = (\{Sea, Sun, Sand\}, -1)$ and five uncertain occurrences a_1, a_2, b, c_1, c_2. Occurrences a_1 and a_2 correspond to Ann with different feature-unrelated values: $a_1 = (\{Ann, Hat, Skateboard\}, 0)$ and $a_2 = (\{Ann, Hat, Skateboard\}, -5)$. In the former case Zak does not feel shy in the presence of Ann (value 0) and in the latter he does (values -5). The presence of features *Hat* and *Skateboard* emphasizes that Ann is planning to bring them with her. Occurrence b corresponds to Bob who is an astronaut: $b = (\{Bob, Astronaut\}, 0)$. Finally, $c_1 = a_1 \cup b$ and $c_2 = a_2 \cup b$ are compound occurrences which represent both Ann and Bob coming, appended with the possibility of getting shy (c_2) or not (c_1). Thus, there are two dimensions to uncertainty: 1) who will show up – only Ann (a_1 or a_2); only Bob (b); or both (c_1 or c_2); and 2) whether Zak will be shy (a_2, c_2) or not (a_1, b, c_1). The lottery is then described as $L = \langle r, \{a_1, a_2, b, c_1, c_2\} \rangle$.

D.2 Affective Decisions under Uncertainty

In this appendix, we provide the general definition of a *decision problem under uncertainty* that expands on the more succinct version given in the main text (Section 2.2). Suppose that in period t an affective agent is surrounded by features F_t with affective values determined by the function v_t. This determines agent's current mood $V_t = G(F_t; v_t)$.

Suppose that the agent has a choice among actions in some set A that lead to lotteries $L(a)$ as defined in Appendix D.1. Then, the mood-dependent beliefs about the resolution of each lottery $L(a)$ are established by the agent. Specifically, for each $a \in A$ and $L(a) = \langle r(a), R(a) \rangle$, the *possible affective values* of lottery $L(a)$ are computed. These are given by $R_{L(a)} = \{G_t(r(a) \cup w) \mid w \in R(a)\}$. The elements of this set are compared to the current mood V_t in order to establish which one is the closest to it (mood affiliation). The mood-dependent belief is then a number

$$L(a)_{V_t} := \arg\min_{w \in R_{L(a)}} |V_t - w|.$$

Now that the mood-dependent beliefs $L(a)_{V_t}$, which are given by some affective values, are computed, the agent can choose between them. However, at the time of choice these occurrences are still uncertain and the agent takes into account the cost of this uncertainty. Thus, we define the *expected affective value* of the lottery $L(a)$ as

$$E[L(a)] = L(a)_{V_t} - c(R_{L(a)}).$$

Here $c(R_{L(a)})$ is the *cost of uncertainty* that depends on the set of possible affective values $R_{L(a)}$.

After all these calculations, the agent solves the maximization problem

$$\max_{a \in A} E[L(a)]$$

and chooses the action $m \in A$ that maximizes this expression. After the choice, some affect V_t' is experienced, which can be consistent with agent's mood-dependent beliefs, in which case $V_t' = R_{L(m)}$, or not (so then V_t' is something different). Regardless, the values of all features directly involved in choice get updated following the logic discussed in Section 2.

D.3 Identity-Based Norms of Individual Behavior

In this appendix we describe in more detail the reactions, or *moral sentiments*, that affective agents have when they observe someone making individual choice that does not have any direct, "utilitarian" impact on others (for social choice see Section E.2 and Appendix E.2.1). As was mentioned in Section 3.4, this happens when the observed agent i has or aspires to obtain a social identity g and makes some choice that involves features F_g that are associated with g. It is assumed that moral sentiments of *resentment* or *admiration* arise in this case because this choice reveals something about i's affective values $\tilde{v}^i_g$ or his attitude towards his chosen identity. He can value g not enough in the opinion of the observer j (resentment) or he can value it more than the observer (which evokes admiration). In either case, this signals something about the qualities of i as a potential partner for future cooperation.

We start with a situation when j observes i making a choice in some decision problem under certainty. Consider such problem defined by a set of actions A and the corresponding occurrences r_a for $a \in A$. Suppose as well that some features present in r_a for some $a \in A$ are part of the set F_g associated with identity g that i is known to belong to. In other words, $F_{r_a} \cap F_g \neq \varnothing$ for some $a \in A$. Depending on the presence or absence of these identity-features, j will be able to infer how good or bad i is at "being g."

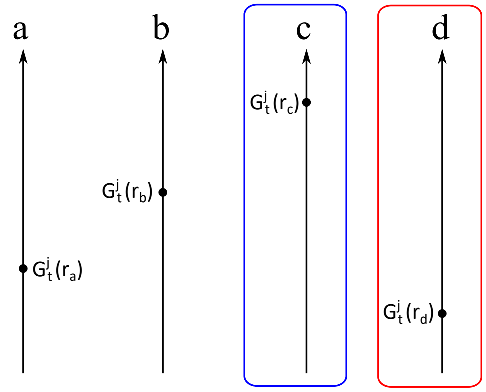


Figure 10: The choice among four occurrences made by the observed agent i (red rectangle) and the choice that the observer j would have made (blue rectangle). The values of occurrences $G^j_t(r_a)$, etc. are from the perspective of agent j.

Suppose that agent i chooses action $d \in A$ that leads to occurrence r_d as depicted in Figure 10 (in the red rectangle). At the same time, the current affective values of the observer j, including her idea $\tilde{v}^j_g$ of what it means to belong to g, are such that from her perspective action $c \in A$ is the best choice (blue rectangle in Figure 10) because $G^j_t(r_c) > G^j_t(r_d)$ with affective values $\tilde{v}^j_g$ used for all features in F_g and v^j_t used for the rest (the notation is defined in Appendix D.1). Assume as well that $F_{r_c} \cap F_g \neq \varnothing$, or that occurrence r_c contains some identity-features from F_g *that i did not choose.*

The fact that i did not go for action $c \in A$ that leads to some identity-features suggests to j that something is "wrong" with i's idea about his identity. After all, why is it (j deliberates) belonging to g implies choice c when in reality i chose d? We assume that j attributes this behavior to something being wrong exclusively with i's ideas about his identity coded in $\tilde{v}^i_g$ and nothing else (see the end of this section for discussion). Given this, j can make the following steps of reasoning. She knows that

$$G^j_t(r_c) = \sum_{f \in F_{r_c} \setminus F_g} v^j_t(f) + \sum_{f \in F_{r_c} \cap F_g} \tilde{v}^j_g(f) + s_{r_c} > G^j_t(r_d).$$

This can be rearranged as

$$K_j = \sum_{f \in F_{r_c} \cap F_g} \tilde{v}^j_g(f) > G^j_t(r_d) - \sum_{f \in F_{r_c} \setminus F_g} v^j_t(f) - s_{r_c} = X.$$

In other words, j's aggregate affective value of identity-features within r_c, that we call K_j, is higher than some constant X on the right side of this expression, or $K_j > X$. In the mind of j, the fact that i did not choose c then implies that for i

$$K_i = \sum_{f \in F_{r_c} \cap F_g} \tilde{v}^i_g(f) \leq X,$$

or that the aggregate value of identity-features in c, as i perceives them, are at best equal to X (or lower).

Given this reasoning, j comes to the conclusion that i *undervalues* the identity-features in $F_{r_c} \cap F_g$ by at least $K_j - X$. It can, of course be worse than that, but j does not have additional information to make that judgement, so j feels resentment (negative affect) towards i equal to $-(K_j - X) < 0$, which is the most optimistic estimate of how much less i values his identity than he should have. This is important for j, because it means that i is not taking his identity seriously enough (low values imply wrong choices, like this one being observed), which means that i is not going to cooperate to full degree within identity g and possibly within other identities that j might care about. As a result, j updates the affective value of the feature f_i that code agent i with negative affect $-(K_j - X)$ that she just felt:

$$v^j_{t+1}(f_i) = v^j_t(f_i) + \lambda(-(K_j - X) - v^j_t(f_i)) = (1 - \lambda)v^j_t(f_i) - \lambda(K_j - X) < v^j_t(f_i).$$

This constitutes the end-result of observing individual behavior of agent i.

Several things are worth noting about this procedure of feeling resentment about individual behavior of others. First, the observer j uses her own affective values to make all these calculations and feels resentment based on how *she* evaluates everything and not i himself. Specifically, j does not think about what affective values agent i might have that drove his choice. This goes very much against standard rational approach where each agent usually tries to imagine what preferences are behind the behavior of others. This difference is important. In the affective framework, agents perceive their moods and other reflections of their affective values as currently best available knowledge about the world in general (from this comes mood affiliation). For affective agents, their affective values is their *umwelt*, or the way they *see the world*, which is by definition the only correct one. Therefore, they never think about that others might have some other views. Of course, people can imagine that others see the world differently, but this is a part of emotional intelligence, which builds on top of the affective system (discussed in Appendix A).

Second, there are slight complications with the computation of resentment presented above. Given that the observer j tries to figure out by how much i's values related to identity g are "off," it is possible to imagine that occurrence r_d might also have some features from the set F_g, or that $F_{r_d} \cap F_g \neq \varnothing$. In this case, it is possible that i overvalues those identity-features instead of undervaluing features in $F_{r_c} \cap F_g$. This is indeed a valid concern and we believe that in such cases resentment might be not so unambiguous as in the case when $F_{r_d} \cap F_g = \varnothing$.[2] To give an example, suppose that i is an astronaut and that he sells his space suit (r_c) to buy recreational drugs (r_d). This choice obviously deserves resentment because i forgoes something crucial to his identity to have fun. But now imagine that i sells his space suit to help a poor person in need. Here the situation is not so clear, because by helping a person in need i upholds another quality of being an astronaut (he is not selfish). So, people might disagree on whether this act deserves resentment or not. Some might say that space suit is a sacred object for an astronaut that he should never sell for anything, some others would think that it is reasonable to do that when someone else is suffering. So, our framework predicts that resentment will be much more clear and pronounced when $F_{r_d} \cap F_g = \varnothing$ rather when the opposite holds.

This brings us to another interesting case that can also take place, namely that of admiration. Imagine that the setup is as depicted in Figure 10, but that now $F_{r_c} \cap F_g = \varnothing$ and $F_{r_d} \cap F_g \neq \varnothing$. In words, agent i

[2] Though, the calculations go through in the same manner as presented above anyway, because by *not choosing* c agent i shows that he undervalues features in r_c, which is more important for detecting future failures of cooperation than when he overvalues some other features in F_g. This idea predicts for example that people should care more about others' undervaluing features than their overvaluing them, because the latter does not threaten future cooperation, whereas the former does.

chooses action d that leads to the features contained in g, even though from the perspective of j he should have chosen a much higher value following action c that does not have any identity-features related to it. This would hold, for example, when astronaut refuses to sell his space suit (r_d) even when he personally is in a very bad situation and needs money to eat (r_c). In this case, all calculations as with resentment go in the same way only with the inequality signs being switched everywhere. This leads to *admiration*, or positive affect that j has for i upon observing such act, because it signals that i is willing to stick to his identity even when anyone else would have sold his space suit. This increases the affective value of i in the eyes of j. This mechanism also lies behind the admiration of *heroes*, who make personal sacrifices for the sake of a social identity which go beyond what others would have done.

Now that we looked at how individual behavior is judged in decision problems under certainty, we can do similar exercise for decision problems under uncertainty. The only difference in this case is that before making the judgement, the observer j uses mood affiliation to determine how uncertainty should be resolving in various available lotteries. Thus, her moral judgement of agent i will become *mood-dependent*.

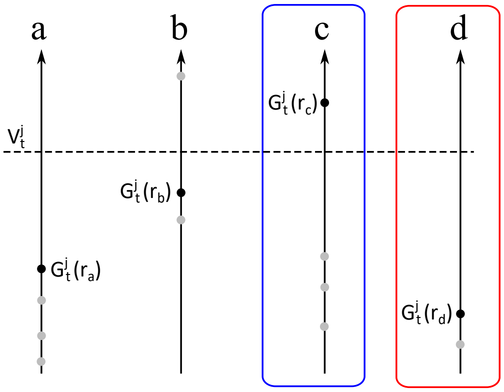


Figure 11: The choice among four lotteries made by the observed agent i (red rectangle) and the choice that the observer j would have made (blue rectangle). The values of occurrences $G_t^j(r_a)$, etc. are the mood-dependent beliefs of agent j given mood V_t^j.

Suppose that agent i was facing a choice among actions in some set A that lead to lotteries $L(a)$ and that he chose action d shown in the red rectangle in Figure 11. However, observer j, given some current features F_t and her subsequent mood V_t^j, would have chosen action c (in blue rectangle, grey circles represent the values of other possible occurrences in the lotteries). In order to judge the behavior of i, observer j considers only the occurrences from the lotteries that she believes will happen. Specifically, on Figure 11 she would consider r_c and r_d, which are compound occurrences picked up from the lotteries $L(c)$ and $L(d)$ with mood affiliation. After that she uses the same procedure as for the decision problem under certainty to calculate her resentment or admiration for i depending on the presence of identity-features from F_g in either of these occurrences.

The most important conclusion that this analysis suggests is that when affective agents judge the individual behavior of others in the presence of uncertainty, their judgements will become mood-dependent. So, the same action of agent i might be judged differently by happy and sad observers even if they share the exact same affective values (they might be surrounded by different features at the time of observation). This shows that uncertainty can have detrimental effects on the social cohesion of the community and consequently cooperation as people might judge the same observation differently and disagree. It would be interesting to test this prediction experimentally.

On a final note, we would like to mention that, in principle, it is possible to judge individual behavior of others even without singling out any social identity g. Indeed, in Figure 10 for example, observer j could simply feel resentment for i equal to the distance between the values of occurrences r_c and r_d when agent i chooses differently from how j would have chosen. This would be analogous to a situation when someone resents the behavior of others in any situation as long as it is different from how they would have chosen. Such behavior probably even occurs sometime. However, we believe that moral sentiments like resentment and admiration evolve for specific evolutionary purposes, namely to support social identities as engines of cooperation. Cooperation would not be enhanced, and actually would be rather diminished if everyone felt resentment whenever they observed someone doing something different from how they would have done it. Therefore, we do not consider this possibility in this paper.

E Social Behavior

In this section we describe how affective agents behave in strategic environments (games), which represents the ultimate level of sophistication of their ability to work-together. The main addition to what we proposed above is that affective agents are also capable of caring about how appropriate their behavior is in situations when their actions directly influence the payoffs (or feelings) of other affective agents. Here we assume that, like rational agents (see KV), affective agents can follow norms that guide their choices in strategic situations that take into account dissatisfactions of others. Unlike in KV though, here we explicitly specify how the *social weights*—that regulate the intensity with which different agents' dissatisfactions should be considered—are formed.[3]

In KV, we assumed that in forming opinions about the appropriateness of certain outcomes in a game, rational agents compute the dissatisfactions of all players at some outcome and then aggregate these dissatisfactions weighting them with some social weights that determine how much they care about each other player relative to themselves. We did not explain where these weights come from. Here we propose that social weights are determined by the affective values attached to the other players. Essentially, the social weights that agent i uses in the aggregation of dissatisfactions of other players *are* his affective values attached to these players. So, if agent i has high affective value associated with some agent j, then he will believe that it is appropriate to favor this agent at the expense of some other agent k, who has low affective value in the eyes of i. The affective values or social weights of others can be formed through simple observation of features associated with them (Section 3.1), through observation of their behavior within some social identity (Section 3.4), or through assignment of a measure of status within an identity (Section 3.3).

E.1 Games with Affective Players

As was mentioned in Section 2, affective agents are not rational. Therefore, they do not reason rationally in strategic situations. Rather, they see any move in any game as a choice under uncertainty as described in Section 2.2. Let us take any normal or extensive form game with affective players $N = \{1, ..., N\}$ and a finite set of outcomes. Take one node from anywhere in the game, in which player i moves, and consider the set of outcomes C that are reachable after i's possible moves. Suppose that each $c \in C$ is an allocation of feature-unrelated affective values to N players (the general specification can be found in Appendix E.1.1).

We assume, as in KV, that player i computes dissatisfactions in each outcome $c \in C$ for himself and other players using the set of allocations C. Moreover, when it is time to move, player i has affective values $\tau_{ik} = v^i(f_k)$ of features f_k, $k = 1..N$, that correspond to himself and the other players in N.[4] The values τ_{ik} are used by player i as social weights to compute the norm function $\eta_i : C \to \mathbb{R}$ as in KV. This gives player i the *norm-dependent affective value* in outcome c given by

$$u(c) = s^i(c) + \phi_i \eta_i(c),$$

where $s^i(c)$ is the individual affective value that player i gets in outcome c and $\phi_i \geq 0$ is his propensity to follow norms (as in KV and also the same parameter as in Section 3.3). The norm function η_i has a subscript i, because, unlike in KV, player i uses his own social weights τ_{ik} to compute the norm function that can be different from the weights of other players.

[3] It is worth noting that the assumption that affective agents can use dissatisfaction-based norms is debatable. On the one hand, empathy that allows to estimate the dissatisfaction of others is an affective phenomenon. However, on the other hand, the computations needed to construct the norm function as in KV are pretty intense and this suggests that dissatisfaction-based norms might be a cognitive phenomenon.

[4] We assume that player i has also a *self-feature* f_i with affective value $\tau_{ii} = v^i(f_i)$ that can be any number. This allows for situations when player i thinks that he is, for example, "not worthy" (very low τ_{ii}) or that he is "better than everyone else" (very high τ_{ii}).

Now suppose that player i has the set of actions A available in the node of the game under consideration. Each action $a \in A$ corresponds to the set of outcomes $C(a)$ that are reachable after it is chosen. Let us define a lottery

$$L(a) = \langle 0, U(a) \rangle = \langle 0, \{u(c) \mid c \in C(a)\} \rangle.$$

Here $U(a) = \{u(c) \mid c \in C(a)\}$ is the set of norm-dependent affective values reachable after action $a \in A$ that constitute the lotteries among which player i can choose.

To complete the formulation of a decision problem under uncertainty (as in Section 2.2), we need to define the current features surrounding player i. These can be some features F plus the features corresponding to the other players $F_P = \{f_1, ..., f_N\}$. Thus, the mood of player i can be computed as

$$V_i = \sum_{f \in F \cup F_P} v^i(f).$$

The mood V_i and the lotteries $\{L(a) \mid a \in A\}$ define the decision problem under uncertainty. Player i makes a choice using mood affiliation with V_i as described in Section 2.2.

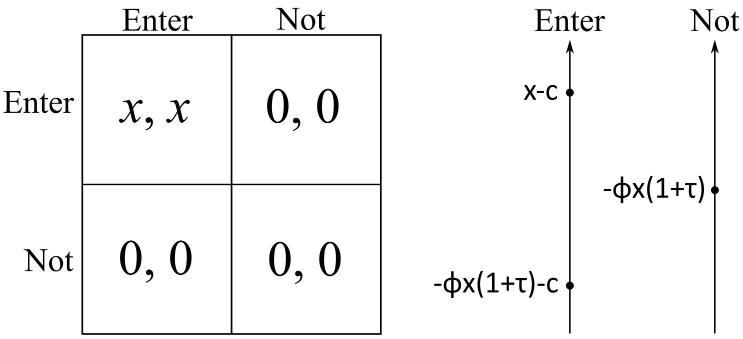
At this point it is important to discuss the feature-updating that player i still needs to perform after he made a move. If we treat the game that he is playing as a classic game-theoretic construction where payoffs are realized only at the very end of the game, when some end node in C has been reached, then, technically, player i can only update the features at that end node. However, most dynamic games that are realistically considered in economics are repeated games or games with observable actions, which are some sequences of normal forms *where payoffs are realized after each move*. From the standard, rational perspective, whether the (partial) payoffs are known after each move or only at the end of the game is irrelevant: rational players do not care about that because they strategize about the whole game. But for affective players this makes a big difference. If affective players play a sequence of normal forms, then we should treat each normal form in the sequence as a separate game with the updating happening after each move, whereas if payoffs are only realized at the end, then affective players cannot perform intermediate updates, which changes the nature of their interaction.

Note as well a conceptual difference in norm computation between games with affective players and games with rational players in KV. Rational players when computing the norm function use the *full set of outcomes that can be reached in the whole game*. This makes sense, because rational players consider the whole game when thinking about who will do what and when, which is important for the formulation of their own strategy. Affective players are not rational, thus at every move *they only consider the outcomes that can be reached after that move* and do not consider counterfactual outcomes that cannot be reached anymore. This happens because affective players do not make plans for the future, do not consider how others will behave after different contingencies, and do not strategize in any way. Therefore, this difference in reasoning style can change how the two types of agents see what is morally right or wrong.

We finish this section with the extended version of Example 3 in the main text. Here we assume that players care about dissatisfactions of others.

Example 8. A Goods Exchange Game II. Suppose that Player 1 possesses some quantity of good A that he does not derive any affective value from and that Player 2 possesses some quantity of good B that she also does not care about. However, both players derive affective value of $x > 0$ from consuming the good of another player. In this case, the players can enter an exchange where they switch their goods or they can choose to not do anything. The left panel of Figure 4 illustrates.

If players are selfish and rational, both should play Enter as long as they put non-zero probability on the other playing Enter, which is not an unreasonable thing to believe. Now, suppose that players are rational and follow norms as in KV. They attach social weight equal to $\tau_{11} = \tau_{22} = 1$ to themselves and some non-negative social weights to each other ($\tau_{12} = \tau_{21} = \tau$, common knowledge) and have some non-negative propensities to follow norms ($\phi_1 = \phi_2 = \phi \geq 0$, also common knowledge). Then, the game with norm-dependent utility can be normalized to have the exactly same payoffs as in Figure 4. Thus, again, players are very likely to enter the exchange. The standard economic intuition holds in both cases.



	Enter	Not
Enter	x, x	0, 0
Not	0, 0	0, 0

Figure 12: **Left Panel.** A Goods Exchange Game. **Right Panel.** The choices in the game transformed by an affective player to the decision problem under uncertainty.

Now let us consider affective players 1 and 2 playing this game with the same setup (including social weights and propensities to follow norms). Suppose that Player 1 is surrounded by two features corresponding to himself and Player 2. In other words, his current mood is $V_1 = 1 + \tau$ (τ serves also as an affective value for Player 2). Then action Not leads to a sure outcome $0 - \phi(1 \cdot (x - 0) + \tau(x - 0)) = -\phi x(1 + \tau)$, which takes into account the dissatisfactions from not having x himself (times 1, social weight on self) and from Player 2 not having x (times τ, social weight on the other), see right panel of Figure 4. The action Enter leads to a lottery $\langle 0, \{x - c, -\phi x(1 + \tau) - c\}\rangle$ with two possible outcomes $x - c$ and $-\phi x(1 + \tau) - c$ that, in addition to dissatisfaction, also take into account some cost of uncertainty $c > 0$. To choose between these actions, Player 1 will use mood affiliation and will decide that outcome $-\phi x(1 + \tau) - c$ of the lottery will realize if V_1 is closer to $-\phi x(1 + \tau) - c$ than to $x - c$. This happens when $V_1 = 1 + \tau < \frac{x-2c}{2+\phi x}$. Thus, under this condition, Player 1 will choose to do nothing because $-\phi x(1 + \tau) - c < -\phi x(1 + \tau)$, the value he gets from choosing Not. The condition $1 + \tau < \frac{x-2c}{2+\phi x}$ can hold for positive τ given some large enough x and small enough ϕ and c. To put it differently, in order for Player 1 to choose Enter, the inequality above should *not* hold (with positive right-hand side), which happens if τ is high enough. Same logic can be applied to Player 2. □

E.1.1 General Definitions for Games with Affective Players

In this appendix we present the general treatment of behavior of affective agents in games. Given that, as was mentioned in Section E.1, affective agents see any choice as a decision problem under uncertainty and do not strategize as rational players do, we are not going to develop the notation specifically for games. Instead, we will rather look at a single choice of a single affective player in the decision problem under uncertainty where outcomes specify what all players in a game receive. Any move by any player in any game can be easily transformed into such problem (see Section E.1). Despite this simplification of the strategic aspects of a game, the fact that affective players operate with occurrences instead of payoffs creates additional structure that enhances standard game-theoretic payoffs (see Appendix D.1 for the definitions and notation used in the rest of this appendix).

Consider a game with $N = \{1, ..., N\}$ players and a move of affective player i in it. Suppose that i can choose from a set A of actions such that the choice of action $a \in A$ leads (eventually) to some consequences $C(a) \subseteq C$, where C is the set of all consequences of the game reachable after all actions in A.[5] Suppose that upon reaching a consequence $c \in C$ each affective player $j \in N$ experiences an occurrence $r_c^j \in \mathcal{O}$, which are known to i. Player i computes the affective values of these occurrences given his own current affective values v_t^i. For any r_c^j this is given by $u_c^j := G_t^i(r_c^j)$ (we do not mention i in the notation u_c^j for simplicity, but it is assumed that these values are from the perspective of i).

[5] We follow KV and call the set C consequences rather than outcomes.

This procedure generates a set of affective values (real numbers) for each player for each consequence, which can now be used to compute the norm function (KV). To do that, player i uses the set of social weights $\tau_{ij} := v^i_t(f_j)$ defined by the affective values of features f_j that code players $j \in N$. To compute the norm function $\eta_i : C \to \mathbb{R}$, player i uses the affective values u^j_c and the weights τ_{ij} for all $c \in C$ and for all $j \in N$ (see KV).[6] Finally, the *norm-dependent affective value* of player i in consequence $c \in C$ is

$$u(c) := u^i_c + \phi_i \eta_i(c),$$

where $\phi_i \geq 0$ is i's propensity to follow norms as in KV. Now, i's decision problem under uncertainty is defined by the sets of *possible affective values*

$$R_a := \{u(c) \mid c \in C(a)\}.$$

Notice that we omit the definition of lotteries as in Appendix D.2 since they are transformed into these sets anyway for i to choose among them. Suppose that i is surrounded by the set of features $F^i := F^i_t \cup F_N$, where F^i_t are some features specific to i and F_N are the features representing all players in the game. This generates the mood $V^i_t = G(F^i; v^i_t)$ of player i. Finally, player i chooses optimal action in A using $\{R_a \mid a \in A\}$ and mood V^i_t as described in Appendix D.2.

In any normal-form game this procedure specifies what each affective player $i \in N$ chooses. This, in its turn, determines the consequence c^* that takes place. The occurrences $r^i_{c^*}$ are experienced by each player $i \in N$ and the updating of affective values takes place. This finalizes the description of how the normal-form game is played.

For the extensive forms the situation is somewhat different as players can judge others for their social behavior as the game unfolds (analogous to the resentment in KV). We discuss this in Appendix E.2.1, where we also finalize the description of optimal behavior in extensive-form games.

E.2 Identity-Based Norms of Social Behavior

In Section 3.4, we described how others judge individual behavior of agent i who aspires to have some social identity, but signals that he is not up to the task through his behavior. Behavior inconsistent with such aspirations results in decreased affective values that others attach to i. Since, as we have just seen, affective values attached to players are also the social weights used in the computation of norms in games, low social weights of misbehaving agent i will automatically lead to others' treating him badly and moreover believing that this is the right thing to do. This serves as a "built-in" punishment for violating identity-based norms of individual behavior.

Such treatment constitutes the core of what we can call identity-based norms of social behavior. Suppose that a community of agents are tied together by a social identity g. For example, they are Christians. This implies that they should go to church on Sundays (high affective value of the feature *Going to Church*), should not commit sin (low affective value of the feature *Sin*), etc. These values are part of the identity-specific set of values $\tilde{v}_g$ associated with Christianity as social identity g. When we discussed *social status within identity* g, we postulated that higher status corresponds to agents possessing certain features associated with g and behaving in a way consistent with the values $\tilde{v}_g$ attached to them. Thus, as agents in a community live together and observe each other's individual behavior they work out the affective values attached to everyone depending on how well each individual followed the prescriptions coming from g. The agents whose behavior strictly follows the prescriptions coded in $\tilde{v}_g$ (always go to church, never sin)

[6] It is important to note that in this paper, unlike in KV, we do not use the normalization of the norm function to interval $[-1, 1]$. Here the norm function is just the negative of the sum of dissatisfactions. Thus, $\eta_i(c) \leq 0$ for all $c \in C$. In KV, the normalization was used to compare norm functions across games, but affective agents do not do that, this is why we do not use it here. It is open for debate whether $\eta_i(c)$ should be non-positive or it can get positive. The difference for affective players is that in the latter case they would feel positive affect from following the norm, which might play some role in some specific examples, but not in general.

gain high status, whereas individuals who often violate these prescriptions (choose to party instead of going to church) gain low status.

To illustrate, suppose that agent i never goes to church on Sundays. Everyone in the community sees that, because they all are in the church each Sunday and they never see agent i there. Therefore, a consensus emerges that agent i is a bad Christian, which results in his low affective value (low status). Given that this is common knowledge, an *identity-based norm of social behavior* emerges and prescribes that in *any* interactions with agent i he should be treated as a bad person (low social weight). This implies that if someone in the community shares food with agent i, for example, then *this act* will be seen as a violation of this norm since agent i has low status, and anyone who is a Christian (belongs to g) should behave accordingly and never give agent i anything. Similarly, not sharing food with someone who has high status (high social weight) is also considered a violation of identity-based norms of social behavior, because individuals with high status (with high affective value attached to them) deserve being treated reverently.

To summarize, social identity g and individual behaviors of everyone in the community allow affective agents to determine publicly known social statuses of each individual by observing their individual behavior. These statuses are the affective values attached to each agent, which also serve as the social weights in the computations of η_g, which we define here as the norm function in some game that uses these social weights. Anyone, who acts in violation of the norm function η_g, breaks identity-based norms of social behavior that prescribe the "social attitudes" and the corresponding treatment of everyone consistent with their status. Such violations can be detected when someone acts in a game in a way that is inconsistent with maximization of η_g with publicly known social weights. After a violation, the affective values (social weights, status) of the person breaking the norm go down accordingly. In Appendix E.2.1 we describe mathematically how such judgement of social behavior takes place.

E.2.1 General Definitions for Identity-Based Norms of Social Behavior

In this appendix we present how affective agents judge behavior of others when they play games, or in other words, when their choices have consequences for the affective values of others. This same mechanism is also used by affective players in extensive-form games when the judge the behavior of other players who moved before them (see Appendix E.1.1).

We use the same setup as in Appendix E.1.1. Suppose that player i chooses an action from some set A with reachable consequences C. Each consequence $c \in C$ leads to occurrences $r_c^j \in \mathcal{O}$ for each player $j \in N$ (who is playing the game). Suppose as well that agent k, who might be playing the game ($k \in N$) or not ($k \notin N$), observes the choice of i. Agent k can judge how socially appropriate this choice was from the perspective of her norm function $\eta_k : C \to \mathbb{R}$ that she can compute using her affective values v_t^k that also include the social weights τ_{kj} of the players $j \in N$ in the game (same as in Appendix E.1.1). To do that, agent k forms *normative occurrences* $(\varnothing, \eta_k(c))$ that represent the affective values expressed by η_k for each $c \in C$ and uses them to judge the behavior of i similarly to how it was done in Appendix D.3. Specifically, suppose that k's mood V_t^k is determined by the features surrounding her (some features F_t^k and features F_N corresponding to the players in the game). She uses mood affiliation to determine which normative occurrences will occur after each action $a \in A$ and then compares what she would have chosen with what i did. Suppose that i chose some action $d \in A$ leading to the normative occurrence $\eta_k(c_d)$ (abusing notation) and k would have chosen the action $e \in A$ leading to normative occurrence $\eta_k(c_e)$. Since k is maximizing, it will always be true that $\eta_k(c_e) \geq \eta_k(c_d)$. So, k will feel resentment towards i of the size $\eta_k(c_e) - \eta_k(c_d)$ and experience non-positive affect $V_t' = -(\eta_k(c_e) - \eta_k(c_d))$ with the consequent updating of the feature f_i coding player i. This finalizes the description of k's judgement of the social behavior by player i.[7]

[7] Above we did not make any difference between observer k who is part of the game ($k \in N$) and that who is not ($k \notin N$). However, it is not inconceivable that the difference exists between what they use to judge actions of i. While it is relatively intuitive that outside observer $k \notin N$, who does not play the game, uses $\eta_k(c)$, since she

Notice that this procedure is somewhat different from that in Appendix D.3 where social identity of i played the main role. Here, we do not focus directly on social identity, though it possibly operates in the background determining the social weights τ_{kj}, but rather on the norm function that k forms. Also, the difference is that with social behavior *k can never feel admiration*, because the best thing that i can do is to choose the most socially appropriate action from k's perspective, in which case k does not update i's social weight. This is also the same mechanism that is described in KV (resentment), only there rational agents use this kind of resentment to compute punishment that i deserves, whereas here the social weight of i is updated, which represents a second, parallel mechanism of punishment.

Finally, if $k \in N$ and k is a player in an extensive-form game (with payoffs being revealed only at the end of the game, see discussion in Section E.1), she will update social weight of i (as would all other players) after any move that i makes. In extensive-form games all players update social weights of each player who moves following the procedure described above. Therefore, in order to compute the outcome that will happen in an extensive-form game, we need to start from the first node of the game, determine what move each player who is active in this node will make and then update the social weights of all these players within the affective values of all players in the game. After that we can determine what the next move in the game with updated social weights will be, etc. This "forward induction" procedure will lead eventually to some endnode that we can consider as the predicted outcome of the game.

does not have any stakes in it, observer $k \in N$, who does have a stake, might use the norm-dependent values of consequences $u_c^k + \phi_k \eta_k(c)$ instead of simply $\eta_k(c)$ to make judgements. Or maybe not with ϕ_k, but some other coefficient ϕ'_k. This is possible for two reasons: 1) being involved in the game gives k a reason to blame i for her misfortunes as well as for the appropriateness of actions and 2) when playing the game k evaluates $u_c^k + \phi_k \eta_k(c)$ anyway, so neurophysiologically it might be difficult to construct another value $\eta_k(c)$ and separate it from the said norm-dependent value. We believe this is possible as intuitively the behavior when people blame others for their personal losses (unrelated to norms) are very common.

Additional References in Appendices

Benoit, R. G., Szpunar, K. K., and Schacter, D. L. (2014). Ventromedial prefrontal cortex supports affective future simulation by integrating distributed knowledge. *Proceedings of the National Academy of Sciences*, 111(46):16550–16555.

Callen, M., Isaqzadeh, M., Long, J. D., and Sprenger, C. (2014). Violence and risk preference: Experimental evidence from Afghanistan. *American Economic Review*, 104(1):123–48.

Gamberini, L., Chittaro, L., Spagnolli, A., and Carlesso, C. (2015). Psychological response to an emergency in virtual reality: Effects of victim ethnicity and emergency type on helping behavior and navigation. *Computers in Human Behavior*, 48:104–113.

Herwig, U., Opialla, S., Cattapan, K., Wetter, T. C., Jäncke, L., and Brühl, A. B. (2018). Emotion introspection and regulation in depression. *Psychiatry Research: Neuroimaging*, 277:7–13.

Kato, T. A., Kanba, S., and Teo, A. R. (2019). Hikikomori: multidimensional understanding, assessment, and future international perspectives. *Psychiatry and clinical neurosciences*, 73(8):427–440.

Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *American economic review*, 90(2):426–432.

O'Doherty, J. P., Rutishauser, U., and Iigaya, K. (2021). The hierarchical construction of value. *Current Opinion in Behavioral Sciences*, 41:71–77.

Salovey, P. and Grewal, D. (2005). The science of emotional intelligence. *Current directions in psychological science*, 14(6):281–285.

Schutte, N. S., Malouff, J. M., Simunek, M., McKenley, J., and Hollander, S. (2002). Characteristic emotional intelligence and emotional well-being. *Cognition & Emotion*, 16(6):769–785.

Suddendorf, T. and Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and brain sciences*, 30(3):299–313.