Theory of Minds

v1.0

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

May 2024

Abstract

This paper presents a theory of human mind built as a sequence of minds of increasing complexity from automatons to affect to cognition. In the associative model, we conceptualize mental processes as signals spreading over associative network of features (mental representations of objects, feelings, concepts, actions, etc.). These signals activate various affective and cognitive mind devices eventually connecting outside stimuli to behavior. This model can be helpful to neuroscientists, human evolutionary biologists, psychologists, and anyone else who is interested in the details of human mental processes.

In the context model, we present a reduced-form agent, based on the same principles, who works as a utility maximizer. In a parsimonious mathematical framework, we show how to model all aspects of human condition relevant for decision-making as a parametrized continuum of types (affective vs. cognitive preferences, bounded vs. full rationality, affective vs. cognitive morality) and how these aspects influence choice. We also suggest how to connect the model to applications with simple survey data, which can be helpful to economists in theory, practice, and policy.

Keywords: *affect, cognition, language, evolution, affective decision-making, bounded rationality, context-dependence, morality, social norms.*

[*]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. All mistakes are our own.

Contents

1 Introduction

In this paper, we present the theory of minds that describes main aspects of how human mind works. We discuss two models that use different mathematical approaches. The *associative model* uses features connected in the mind by associations as modeling primitives. Features can be any objects, feelings, concepts, actions, or anything that the mind can perceive or do. When a stimulus arrives from the environment, it activates the related features in the mind and the signal from them spreads over the network of associations. When some action features are activated in this way, the mind acts. Alternatively, the mind can act after it uses cognition to understand what is going on. With the associative model, we suggest that, in the process of evolution, minds of increasing complexity (eight of them) gradually evolved to form the affective and cognitive systems in humans.

The *reduced-form model* (aka *context model*) is built in economics style (see Section 7) and suggests an alternative modeling approach based on *contexts* as primitives. A context consists of all features that light up in the mind at one time and their intensities that record how relevant these features are. One context describes everything that agent feels, thinks, does, and imagines. We assume that the mind has affective and cognitive values defined over contexts and suggest how choice among actions is implemented in the world where mind moves from one context to another as a result of choice. The main idea of the context model is to develop a boundedly rational agent with context-dependent preferences that can be traced in time as agent learns. At the same time, agent's psychology becomes more cognitive, or rational, if agent keeps using cognition. The continuum of types of the resulting agent can exhibit all kinds of affective behavior, motivated beliefs, or any other psychological biases known to us (the authors).

It might be too difficult to describe the details of the models in the introduction given that we create two completely new mathematical frameworks with their own rules. That is why, we first briefly summarize what implications the models have for research in social sciences in general and then discuss what they can help researchers do in different fields of social sciences.

In our opinion, the main lesson from these types of models is that human mind consists of *several* mechanisms that produce decision-making. These mechanisms can be classified as *affective* or *cognitive*. None of these mechanisms is more or less important than others. They are all amalgamated together and all produce significant effect on minds and behavior. Thus, in order to understand *any* human behavior we cannot restrict ourselves to studying only cognition (economics) or only affect (psychology). We must study them together.

The reason we believe this is so important is the following. The conceptual difference between affective and cognitive mechanisms of decision-making is that the former are built on *model-free reinforcement learning* and the latter use *models of reality*. This has deep implications for how contexts are treated by the decision-maker. Model-free systems do not know how the world works (they do not have a model of it), they simply react to features in the contexts where

they need to make some choice. This implies that *all* features in a given context—no matter how insignificant from the perspective of cognition—will have the same influence on what the decision-maker does. In other words, affective decision-making mechanisms produce *context-dependent behavior* where context-dependence can be random, influenced by idiosyncratic past experiences, traditions, or anything at all. For example, some people love and some people hate spicy food. This is an idiosyncratic context-dependent preference that has no rational basis (today), but that nonetheless can have a very significant effect on behavior.

Cognition, on the other hand, produces behavior that is not context-dependent in the above sense. Of course, some features of the context are always important for making a good choice, but in case of cognition, these features serve as inputs into models of reality that produce behavior dependent on these inputs.

The point of this is that to perfectly predict the behavior of a purely cognitive agent, all we need to know is her models of reality. But to perfectly predict the behavior of an affective agent, we need to know all experiences of that agent in all past contexts, which is much harder. Given that human mind is a mixture of the two decision-making systems, its behavior will *always* be context-dependent in the affective sense mentioned above. It can be context-dependent to a larger or smaller degree depending on how cognitive the agent is, but for the vast majority of people it is safe to imagine that the degree of context-dependence will be rather high.

If random or culture-relevant features of contexts exert as much influence on behavior as our models suggest, then the only realistic way of studying such behavior is by *focusing on specific contexts* and trying to understand the idiosyncratic cultural phenomena that might drive the behavior of a specific group of people in them. In fact, this might not be as hard as it sounds. People live in communities that share common culture and have many common experiences. So, we can approximate the past experiences of individuals in a community from the past experiences of the community as a whole that we can find out in the local newspapers or other local sources of information.

This idea goes against the typical approach inherited from natural sciences where economists, psychologists, sociologists, etc. try to comprehend some "universal laws" of behavior that apply to all people in all contexts. Our models suggest that such universal laws might simply not exist—or be not as universal as some researchers desire—given that biology uses model-free decision-making mechanisms that can imbue any random features with significance.

Thus, we propose a *context-centered approach to research* where in a given, well-specified context and in a given population we study specific details of what might influence beliefs and behavior in these specific conditions. Practically, this does not even imply any changes in how we do research. Any field study or policy already focuses on a specific context and population. Thus, what we suggest is simply that instead of trying to apply some "universal law" models to this context, we need to understand how people see it, what ideas or concepts come to their minds in this context, which cultural attributes are important, etc.

This general argument applies mostly to social sciences like economics or sociology that study groups of people and their collective behavior. However, we believe that our models can be helpful to all social scientists. So next, we continue with some ideas along these lines.

Associative model describes how human mind works in small detail defining various devices that the mind uses for computations. These devices can be mapped to actual organs in human brain, which can help neuroscientists to study the whole interconnected system of brain organs and networks while having a good idea what they are for and which functions—from the information processing perspective—they exactly perform. This can potentially move the frontier of our understanding of human brain.

To human evolutionary biologists, the model can provide a language in which they can discuss theories of evolution of human mind, language, cognition, and cooperation (see e.g. Rusch and Vostroknutov, 2023). We propose how various characteristics and functions of the mind can be divided into the sequence of layers as they evolved in time. This can be helpful to clarify the steps of evolution of human cognition and to better understand how it functions.

To psychologists, the two models offer a mathematical framework where affective and cognitive processes, that lead to various psychological states or characteristics of behavior, can be rigorously modeled and tested. This can lead to the development of mathematical psychology that meets economics and studies human condition using mature and flexible mathematical models.

The models can help sociologists to study social identity, culture, and their influence on macro-level societal behavior and trends of development. It is possible to model how identity intersects with cognition and other mind functions and how it changes with new experiences (see also Kimbrough and Vostroknutov, 2022). Through our models, the study of social identity can be connected to neuroscience, psychology, and economics, and thus potentially enrich our understanding of human sociality.

To economists, the context model can serve as a replacement of their usual neoclassical framework. We believe that our models solve one crucial problem that stands on the way of proper behavioral modeling in economics. All behavioral models in existence (known to the authors) cover only one specific aspect of human behavior or cognition and never connect it to the rest. Thus, it is unclear how to use behavioral models in applications, since they are all inconsistent with each other; cover separate aspects of behavior or affect/cognition; and do not specify if there are some important features that were left out. The context model we present in Section 7 incorporates all relevant features and biases of human behavior, cognition, bounded rationality, morality, etc. Thus, it provides a *full* description of a human being where *nothing* was left out. Economists can use our models without worrying that there are some other effects that they did not consider. We believe that this provides a huge advantage for studying human behavior even if our models are not perfect. This can be fixed later.

The context model also provides ideas about which data should be collected to verify the model or use it for applications. In fact, the types of data that the model requires can be collected from public domain without even asking people for their individual preferences. The model suggests that we need to look for three separate types of preferences: affective value, familiarity value, and cognitive value. Affective value defines what people like to consume, what gives them pleasure. In a given population, this information can be obtained from commercials shown on TV, from the advertisements online or elsewhere. To understand what is familiar to people, all we need to do is to look where they spend their time, which is easy to learn. This can give us an idea of what people like because it is familiar to them. To uncover cognitive value, we need to learn which models of reality people use in their lives (physics, religion, etc.). This information is out there and is well-known to us: it is contained in school and college curricula, and in popular books, religions, and traditions that people follow. The context model suggests that people's preferences are defined by common culture in the society to a rather large degree. Thus, to study human behavior in groups we need to only obtain data on the group culture, common traditions, customs, and common information that the group under study receives.

We believe that our models can also be helpful to philosophers. We touch upon several interesting issues in philosophy of mind, consciousness, and the connection between objective and subjective reality. We also develop a mixed model of morality (based on Kimbrough and Vostroknutov, 2023c) that combines affective and cognitive moralities (aka deontology and consequentialism) on a single continuum, which suggests a new approach to studying complex moral issues. Similarly to heterophenomenology approach (Dennett, 1991), we propose to study human behavior with first figuring out the inner world of the decision-maker, what concepts she has, how she sees reality, etc. We hope that our models can help to unify various philosophical approaches and to demonstrate how they can be reconciled in one model.

Finally, we mention how our models can shed light on the formation and evolution of institutions (see also Robinson et al., 2023). The new perspective that takes a lot of human nature into account can be very helpful for philosophy, economic history, macroeconomics, and policy. Overall, we believe that our models can create valuable interconnections between all fields of social sciences, unify their approaches, and help to solve pressing societal problems.

2 Associative Model

The theory we present in this paper is abstract in the sense that, like all other theories, it has some basic axiomatic elements on which everything is built. In this section, we start with describing these basic elements and what they can represent in the mind and in the real world.

2.1 Features

We start with the assumption that our minds perceive the world as a collection of *features*. We define features as any perceivable entities that a mind can feel. These can be sounds, tastes, colors, shapes, smells, movements, objects, concepts, etc. For example, the smell of mango is a feature as well as the color blue, raising your left hand, the word "democracy," and Superman. Features can be more elemental like basic senses (e.g., taste of sugar, feeling of pain) or composite, like a bear, consisting of many sub-features (head, paws, etc.). Whether something is being treated as a feature by a given mind can be a complicated question and we do not claim to actually answer it in this paper (research in neuroscience is needed to do that). But, what is important for our presentation here is the idea that *features are the basic elements that any mind perceives and operates with*.

Even though in our theory all perceivable entities are conceived as features, not all features are the same. We think of biological organisms as information processors, who act on the received information. Thus, such organisms should have *sensory features* that are activated when we perceive something with our senses (e.g., see a bright light, or hear an airplane) and *action features* the activation of which leads to the performance of the actions that they represent (we have goosebumps when the feature "switch on goosebumps" is activated). So, the activation of sensory features can be thought of as information input into the mind, and the activation of action features as the resulting commands to perform some actions after the information was processed. In Appendix A, we discuss in more detail our conceptualization of features, composite features, and how they can be used to think about different levels of abstraction in the mind.

2.2 Associations

Biological organisms receive information through sensory organs, then process it—which determines the action that needs to be performed—and finally the action is executed. In between the information input and the performance of an action lies "information processing" that is the main topic of this paper. We postulate that information is transmitted and processed in the mind by means of *associations*. An association is a link that connects two features and through which the activation of one feature activates another. For example, Pavlov's dog is known for salivating when hearing the bell ringing. This was happening because for quite some time the dog was always given food after the bell rang. As a result, the association was created in dog's mind between the feature "bell ring" and the feature "food," that in its turn was already associated with the action feature "salivate." Once the association between bell ring and food was established, the dog salivated upon hearing the bell because the bell associated with food and food associated with salivation.

This example shows how the information received from a sensory feature (bell ring) can be processed and turned into action (salivation) by means of associations via feature "food." Associations are the main mechanism that drives information processing in all levels of minds that we will construct below.

2.3 Associative Network

We assume that signals travel inside the mind on the *associative network* that has features as nodes and associations as links between them (so features and associations together constitute a graph). The specific shape and structure of this network will depend on the complexity of the mind. However, the way the signals are transmitted is common in all associative networks. We assume that sensory features get activated or "light up" when the appropriate stimulus is present in the environment (when a bear is present in your visual field, the feature "bear" lights up). This activation of the feature "bear" leads to a signal being transmitted through all links that are connected to this feature (e.g., the feature representing pain in case the bear attacks). When the signals reach other features on the network, they light up as well and send signals through all links that they have, etc. In this way, the information about the bear (the lighting up of the bear-feature) is processed in the mind by activating associations with other features including for example the action feature "run away." When this action feature is lit up (as a consequence of associative activations starting from the bear-feature), you start to run away.

Notice that we can easily observe how the associative network functions in our own minds. Whenever we perceive a new stimulus, like for example a beautiful flower, we automatically think of other features associated with it (e.g., mother, birthday party, favorite food) that "flash" in our imagination as pictures, smells, sounds, objects, abstract ideas, etc. This flashing corresponds to the activation of features on the associative network as the signal from the original stimulus gets propagated.

2.4 Values

The final basic element of our theory is *values* or utilities related to features. We assume that each feature has some value attached to it. This value is just a number that represents the desirability of this feature from the perspective of past experiences. Negative values represent features that are not good and should be avoided. Positive values represent features that are desirable and need to be obtained or experienced. As the organism has new experiences, the values of experienced features can change to better represent the actually felt outcome (through reinforcement learning). Values play an important role in allowing the organism to better feel itself by aggregating values across features and in providing guidance on the choice of action.

The values of features are perceived when features get activated on the associative network. This can happen when a feature lights up because of an outside stimulus (e.g., you see a bear

and perceive a negative value), or when a feature lights up through the associative network (e.g., you see a forest, which is associated with a bear, and you feel negative value when bear-feature lights up by association).

2.5 Automatism, Affect, and Cognition

The basic elements of the theory described above allow us to conceptualize what minds actually are and how they work. For example, we can reason in terms of the current *state of the mind* (or *mind state*) that is the full description of all information contained in it as well as some fixed parameters that define how information is processed. Given the definitions above, we can say that the information that the mind possesses in a current moment is given by all features present in it, all associative links between the features, and all current values related to them. In addition, a collection of various parameters (assumed fixed) determine how the mind processes new information. These parameters include the rate of reinforcement learning, the strength of the associative signals sent through the network, etc.

Taken together, the information and the parameters contained in the mind determine the current mind state. Using this concept, we can think now about minds of different complexity. For example, we can imagine the simplest possible mind, an *automatic mind,* built only with features and associations whose state is given "from birth" and does not change with the experience of new information. In automatic minds, the associative relationships between sensory features and action features are fixed and do not change. Thus for any collection of perceived sensory features, an automatic mind will always respond with the same set of actions. It is possible that some animals and plants have completely automatic minds.

The next level of mind complexity is what we call an *affective mind*. The main difference between an automatic mind and an affective mind is that the affective mind *changes its state when processing new information.* Specifically, it updates the values of the perceived features and changes associations between them. Nevertheless, affective minds are similar to automatic minds in the sense that information processing and action happen *right after* the new information was perceived. This also implies that affective minds *cannot act* unless a signal from the environment activates some action feature. We can observe the actions performed "affectively" when people exhibit an immediate emotional reaction to some information that they have learned (for example, running away in panic upon seeing a bear). We conjecture that most mammals have affective minds: their reactions to the environment change as they learn new things about it (for example, Pavlov's dog starts to salivate after learning that bell ring is associated with food). A lot of human behavior (for example, habits, immediate emotional reactions) can be classified as performed at the level of affective mind. Such behaviors are always triggered by new information that was just received from the environment and typically happen fast.

Finally, we define a *cognitive mind* that can exhibit more complex behaviors than an affective mind can. The main difference is that cognitive minds are capable of processing information and acting at any time, without the restriction that this should happen right after some information was received. In other words, cognitive minds *can change their state themselves when they want.* This ability, that includes *attention, concentration,* and *imagination,* in principle allows cognitive minds to access and change any information stored in their associative network and use it for action. It should be mentioned though that cognition needs training and such abilities are not given to us by default.

Following economics tradition, we can call the actions chosen by cognitive minds as resulting from *choice.* This means that cognitive minds can imagine various outcomes that can follow after available actions, compute the value that the potential outcomes can bring, and then choose the action that brings the highest "expected" reward (in a sense a bit different from economics though). Notice that affective minds also can perceive value and react to it. However, affective minds act habitually, without understanding that they might have a choice. So, the values that affective minds perceive are not used for choosing, but are rather affecting *mood*, which determines the actions that are performed without choice.

One may wonder why we spend so much time talking about the minds less complex than human, which is broadly assumed to possess cognition. The reason for this reflects the novelty of our approach. Specifically, we do not believe that the human mind is one single "computer" that performs all computations in one specific manner. Instead, we suggest that we have *many minds* all working at once. Our bodily functions, like for example goosebumps or regulation of breathing, operate on the level of automatic mind: we cannot control whether we breathe or not or when we get goosebumps. These actions are triggered directly by some sensory features beyond cognitive control and represent the automatic mind that is always active within us. We also know from psychology and personal experience that humans can act affectively in the sense of responding to information from the environment straight away without using much cognition. Usually, this looks like following habits or exhibiting immediate emotional reactions. Such behaviors, when performed without intervention of cognition, can be thought of as coming from our affective mind that is essentially shaped by the experiences we have and, in general, lives its own affective life. Finally, our cognitive mind operates when we actively think and try to understand what to do using the information available to us in our associative network. Cognitive behavior is usually goal oriented and thought-through in advance, which makes it different from affective behavior that happens fast and without thinking. Thus, our conceptual view of the human mind is that it is a mixture of all different minds that were built on top of one another in our evolutionary lineage. In this paper, we tell one story about how the sequence of minds that we consist of could have evolved.

3 Automatism

In the short overview above, we mentioned that an automatic mind, the simplest there is, does not change with experience. However, this is not exactly accurate. What we should say instead is that automatic mind does not change its state during the lifetime of the organism, but it does change across different generations of organisms. This change happens due to the evolutionary pressure that keeps only the most successful minds alive and thus gradually rewires them to better react to their environment. This evolutionary process suggests that automatic minds are not wired randomly, in the sense that they do not have some random associations between sensory features and action features, but rather that the wiring has some structure useful for the organism's survival. It can be said that the information coded in an automatic mind reflects some knowledge about the environment in which the organism lives.

In this paper, we will not explicitly model how automatic minds came about and what evolutionary pressures forged them. Instead, we want to argue that this evolutionary idea provides the basic blueprint of the organism's "instincts" conceived here as reactions to the activation of some sensory features. For example, humans are instinctively afraid of snakes, spiders, heights, etc. (we automatically react with fear when sensory feature "spider" is activated); we are also instinctively attracted to sugar, fat, sex, having a nice private home and other things that we inherited from our evolutionary predecessors. These instincts—like in the famous Maslow's pyramid of needs—form the basis of what we value, desire, and strive for. The important thing here is that the "preferences" that emerge in automatic minds also get utilized as guidelines for action in minds of higher complexity that build on automatic minds. This is why, we believe that it is very important to describe how such automatic minds might work, so that we could better understand the role they play in human behavior.

3.1 Automatic Action: Spot

We decided to give the minds in our theory individual names, so that it is easier to refer to them in the text and also to remember which one is which. Spot is how we call the simple automatic mind presented in this section (following by more complex Tommy, Freddie, etc. in later sections).

To describe Spot we use the basic elements of our theory mentioned in the previous section, namely features and associations. Remember that we defined sensory features as those that activate when some specific stimulus is present in the environment (for example, light or smell) and action features as those the activation of which triggers the performance of an action (for example, wiggling the tail). Associations connect sensory features—that are designed to detect specific types of stimuli—with the appropriate action features. When a sensory feature gets activated by the outside stimulus, the action feature associated with it gets activated as well by a signal sent through the associative link.

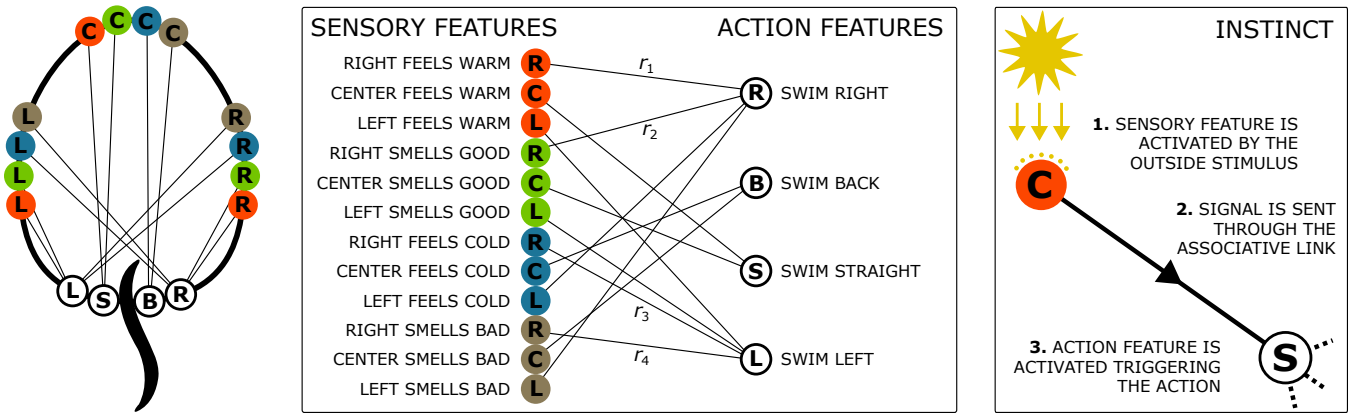

Figure 1: **Left Panel.** A representation of Spot as a swimming organism. **Middle Panel.** Spot's associations between sensory and action features. **Right Panel.** The steps of instinctive information processing by Spot.

To imagine how Spot's mind operates it is easier to think of simpler organisms where the automatic system is more pronounced than in humans, who rely mostly on more complex systems described later (though, "basic instincts" cannot be fully discounted in humans either). The left panel of Figure 1 shows Spot as a simple swimming organism. He has a body and a tail used to swim and navigate around. Spot has four types of sensory features: those that detect warmth (red circles), cold (blue), good smell (green), and bad smell (brown). There are three sensory features of each type located at the center and on the two sides of Spot and marked on the figure with letters *L* (left), *C* (center), and *R* (right). To move around, Spot has four action features shown as the white circles next to the tail. The activation of these features makes Spot swim left (*L*), straight (*S*), back (*B*), or right (*R*).

To process information from the environment and act on it, Spot has associations between sensory and action features built in from birth. We will call this specific type of associations *instincts* to distinguish them from general associations discussed later. We assume that in the process of evolution warmth and good smell have become the signals of something important for Spot's survival, so Spot is attracted to them. At the same time, cold temperature and bad smell have become the signals of something detrimental for Spot, so he tries to avoid them. Thus, whenever Spot feels warmth or good smell (the appropriate features get activated by outside stimuli) he goes in the direction of the stimulus and, when Spot feels bad smell or cold, he goes in the opposite direction.

The associations between sensory and action features, shown in the left panel of Figure 1 and also laid out in more intuitive way in the middle panel, reflect this logic. For example when the central warmth feature (red circle with *C*) gets activated ("center feels warm"), this leads to the activation of the action feature "swim straight" by means of a signal that propagates through the associative link. This process is shown graphically on the right panel of Figure 1. When Spot

feels bad smell or cold on the right, he swims to the left, but when he feels good smell or warmth on the right, he swims right, etc.

Finally, we would like to capture the idea that signals propagating through the associative links can have different "strength" or *relevance*, as we call it. This is important since the signals detected in the environment are continuous like temperature or smell and can be present in different amounts or intensities. Spot can react to the intensity of the stimulus by performing more or less action depending on the strength of activation of the action feature. When action feature "swim left" is activated a little bit, Spot's tail starts to perform a bit of the movements needed to turn left. When "swim left" is activated a lot, Spot starts to swim left as fast as possible. Such situations are not hard to conceive. Suppose Spot feels warmth from the left and bad smell from the right. Both stimuli push him to swim left, which he will probably do faster since both stimuli active the action feature "swim left," and their effect adds up.

We capture this idea by assuming that the associative link can send signals of different relevance measured by some positive number $p > 0$. So, when a sensory feature sends the signal with relevance p (which is proportional to the strength of the outside stimulus), the connected action feature activates to degree p. As a result, the action is performed with "effort" p. If two sensory features send signals p_1 and p_2 simultaneously to the same action feature, then the resulting activation of the action feature becomes stronger and equal to $p_1 + p_2$.

However, for this system to work well, we need to assume that there should be some restrictions to the strength of the communicated signal. Specifically, the signal should not get too strong for otherwise Spot might pull a muscle or try to do something dangerous, outside his physical capabilities. We assume that each associative link has the highest *capacity* shown for some links in the middle panel of Figure 1 as the numbers r_1 to r_4. For example, the link between the sensory feature "right feels warm" and the action feature "swim right" has capacity r_1. This means that whenever a signal of relevance $p < r_1$ is sent, the action feature gets activated to degree p. But if the signal is too strong, namely $p \geq r_1$, then the action feature is always activated to the capacity r_1. Given that each associative link in Spot is fine-tuned for a particular sensor and a particular action, we believe it is reasonable to assume that such restrictions exist. Moreover, they will play the crucial role in more complex minds discussed below.

This example shows how Spot, an organism with automatic mind, can in principle live and survive on his own by means of appropriate reactions to his environment. He can detect food and other resources and swim towards them as well as avoid harmful surroundings. Spot's body functions (like heart beat, breathing, or digestion in humans) can probably also be maintained with automatic, instinctive connections like those shown above.

In general, we can say that automatic minds like Spot's consist of some sensory features (probably a very large amount of them) each of which is connected to *one* action feature. This is the simplest possible representation of an automatic mind. It is also possible to imagine that there might be more complex connections between sensory and action features, or that action

features are connected to each other to form "action programs" like walking or gripping. We discuss these and other possibilities in Appendix B, but for the main exposition it is enough to grasp the idea that automatic minds simply take in outside stimuli and transform them into actions by following a pre-determined routine that we call instinct, which makes them unchangeable by experience.

4 Affect

Spot is a wonderfully simple mind that can live on its own and compete for survival. However, he has some serious drawbacks that do not allow Spot to live in any environment, but only in some of them. For instance, Spot does not know how to deal with stimuli that trigger opposite actions. Suppose that Spot feels warmth on the right that also happens to smell bad. In this case, two features will light up: Right Feels Warm and Right Smells Bad. The former activates action feature Swim Right, whereas the latter activates Swim Left. So, what will Spot do? Given his architecture, Spot will start trying to go both left and right simultaneously, and this might not be very good for survival.

Realistically, Spot should face such problems all the time. Imagine that there is a place that smells both good and bad from all directions. Here, Spot would get completely confused and not be able to follow any coherent course of action. In fact, given Spot's design it becomes relatively clear that the only environment where Spot can be successful should not be too "overcrowded" with stimuli. It should be an environment where warmth, cold, and various smells are all concentrated one by one in separate locations without intersection. It is a world where there are separate patches of warmth, cold, good smell, and bad smell. In this case, Spot will be able to swim between the patches to satisfy his needs. However, once the world becomes too complicated—for example, there are many places now that are warm but smell bad, and places that are cold but smell good—Spot will fail to obtain valuable resources simply because he will try to move in all directions at once.

The problem of too much stimuli and the resulting confusion with action choices will go like a sliver thread through most minds that we construct in this paper, suggesting that environment plays an important role in evolution of minds. Some organisms, who are not pressured by overcrowded stimuli, might not need to evolve any additional thinking skills and can live well at the level of Spot. But some others, to whom our ancestors obviously belonged, were under pressure to live in a world full of contradicting stimuli, and in such a world new ways to perceive reality should have emerged that would help to deal with the confusion.

4.1 Values: Tommy

Economics and neuroscience suggest that people have values or "utility" attached to features because they need to make choices. And indeed, choice is something that involves value comparison and maximization. However, the fact that choice uses values does not immediately imply that values evolved for choice. It might well be that values evolved for something else entirely and then got later incorporated into the choice system.

The problems that Spot faces in his life (trying to move in many directions at once) might suggest that values can play another role, namely that of helping Spot to figure out what to do in situations with confusing stimuli. The main idea how this works is the following. Suppose that features have values and that the values of features that are currently active can be aggregated, or summed up. Then, if the sum of values is positive, this means that everything is more or less fine. Positive sum of values means that most of them are positive, though there might be some negative ones in the mix as well. But regardless, we can imagine that an organism who can perceive the sum of values can resolve the confusion problem mentioned above. The organism can choose to keep searching for food when the sum of values is positive (because the environment suggests on average that there are good things around and the body does not send too many negative health signals) and start avoiding everything, for example, when the sum is negative (the environment contains aversive stimuli and the body might be sending some signals that something is wrong with it). Thus, by having *mood features* representing good mood and bad mood (positive and negative sums of values), the organism can act more purposefully than without them. Additionally, we propose that the *changes* in values are also represented as features that we call *derivative features*, so that the organism can understand that the change is positive or negative and how large it is. Derivative features can allow the organism to navigate quickly in the direction of positive change or away from negative change. In Appendix C we discuss in more detail our assumptions on mood and derivative features and propose some other modeling possibilities.

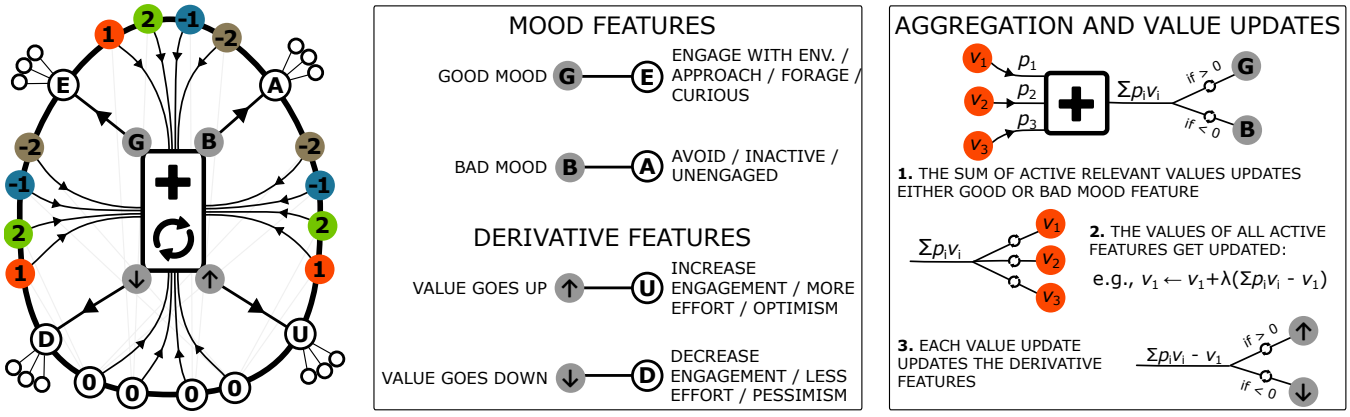

Figure 2: **Left Panel.** A representation of Tommy. **Middle Panel.** New features that come with the valuation system. **Right Panel.** The aggregation and updating process.

In this section, we describe the mind of Tommy that incorporates these ideas in the simplest possible way. Tommy is a much more advanced mind than Spot. He has the whole body rewired so that now each feature has a value that can be updated in real time. Biologically, it is most likely not a trivial task: the valuation system not only involves additional structures that register the values, but also a new brain organ that does the aggregation and updating. The left panel of Figure 2 shows Tommy's valuation system. We kept Spot's instincts in grey on the picture to emphasize that Tommy has a Spot inside him. The instincts inherited from Spot might not be as strong in Tommy, because he can now rely on more advanced valuation system. But even though instincts can be overridden, they are not completely gone, as we all know from our own human experience.

Let us now describe how the valuation system works. The left panel of Figure 2 shows the same four types of sensory features as in Spot in Figure 1, but now we put values inside the circles. Notice that warmth and good smell have positive values of 1 and 2 correspondingly, while bad smell and cold temperature have the values with the opposite sign. We should also attach values to actions, since actions are also features. For simplicity, we set the values of the four actions coming from Spot to 0 (they can be easily changed to anything else).

In addition to the features inherited from Spot, Tommy has special features that reflect the functions of the valuation system. Specifically, the left and middle panels of Figure 2 show four new sensory features marked with G, B, and two arrows in grey circles. These features activate when Tommy is in good/bad mood, and when he feels increase or decrease in values of various features. The activation of mood or derivative features leads to the activation of four new action features (E, A, D, and U) as shown with arrows in the left panel of the figure. These action features represent new behaviors that Tommy can exhibit and that are missing in Spot. Specifically, these actions define the mode of engagement with the environment (mood features) and immediate actions that need to be performed (derivative features).

When in good mood (feature G and then E light up), Tommy actively engages with the environment by, for example, searching for resources (e.g., foraging) or by becoming eager to perform some tasks. Notice that E is also connected to other action features inherited from Spot and shown as three small circles connected to it. These action-action associations are fixed as in Spot (see Appendix B) and represent activities that Spot usually performs when the discovered resources should be utilized or some other tasks should be undertaken. All this means that when Tommy is in a good mood, he will search for things to do and will perform different actions from the repertoire connected to E (the choice of action will depend on capacities and relevances of their associations with E). When Tommy is in a bad mood (feature B and then D activate), he will avoid engagement with the environment, hide, try to find a safe place, or do something else represented by action features connected to A that is necessary to survive in a dangerous situation.

The derivative features provide an additional, more fine-tuned and immediate way to react to the environment. For example, when Tommy, being in a good mood, tries some food known to him that is spoilt and is not good to eat, he will feel a sudden negative change (the down arrow feature and then action feature D light up) that will make him spit the food out. This is one of the action features connected to D that helps Tommy to avoid bad outcomes without changing his mood (the negative change is assumed to be not very large, so that Tommy's mood does not turn negative). This constitutes an important difference from Spot, who might not be able to continue searching for food after a negative stimulus. The same holds in the opposite direction. When in a bad mood, Tommy might be scared and not be willing to do anything including eating. However, when given some tasty snack, he will feel a sudden positive change (the up arrow feature and then action feature U light up) that will make him eat it despite his negative mood. This can help Tommy survive in situations where Spot would be confused by contradictory signals.

On the left panel, we can see that Tommy also has a new organ, the *value aggregator*, that sums up values and updates them (shown as a rectangle with a plus and an update sign in it). Notice that all features inherited from Spot are connected to the aggregator, which is shown with arrows on links that go towards it. In fact, we assume that all features are connected to the aggregator (except probably those mentioned above that are part of it and are updated differently). When some features are active, they send signals to the aggregator, which then sums up their *relevant values* (expressions of the form $p_i v_i$), activates the mood and derivative features, and updates their values as well as the values of all active features.

To see how it works exactly consider the right panel of Figure 2. At any moment in time, aggregator does several things. First, all active features (suppose there are three with values v_1, v_2, and v_3) send signals with relevances p_1, p_2, and p_3 to the aggregator. Second, the aggregator sums up the relevant values $p_i v_i$ to $M = \sum_{i=1,2,3} p_i v_i$. The reason we define relevant values in this way—specifically that relevances of features multiply their values—is that the mind needs to care about both of them at once. Suppose you see a bear, but the bear is far away and does not see you. This situation does not present too much danger even though it is clear that the bear has very negative value. This happens because the bear has low relevance (it is far away). Thus, the mind that multiplies relevance and value estimates the relevant value of the bear as some unimportant negative number close to zero. Even though the bear is dangerous, it does not present any direct threat at the time. The same logic can be applied to all features.

After the aggregator computes the sum of relevant values, the updating starts. The aggregator updates the value of one of the mood features depending on the sign of M. If M is positive, then the value of the good mood feature G is updated. If M is negative, then the value of the bad mood feature B is updated. At the same time, the values of all active features are also updated with the same value M. The updates happen in accordance with the standard reinforcement learning. Specifically, we assume that the current value v of any feature that is being updated

with value M gets new value v that is computed as follows:

$$v \leftarrow v + \lambda(M - v).$$

This formula means that v after the update becomes $v + \lambda(M - v)$, where $0 < \lambda < 1$ is some parameter. Finally, the value of one of the derivative features gets updated with the value $M - v$ whenever a feature with value v is being updated (see Appendix G for other versions of the updater that make more sense in our framework). If for some feature with value v, we have $M - v < 0$, then the value of the negative derivative feature gets updated. If $M - v > 0$, then the value of the positive derivative feature gets updated.

If we compare the behavior of Spot and Tommy upon perceiving the same features, we can conclude that Tommy does a better job. When he is in a good mood, he will be engaged with the environment that sends many positive signals. This increases Tommy's survival probability since his actions—activated by the good mood feature and directed at searching for useful resources—will be successful in such environment. Some experiences of negative values that can happen from time to time will not change his mood (on average the mood is good even with some negative values), so Tommy will on average benefit from the favorable environment. When Tommy is in a bad mood, he will try to abstain from normal activities, heal, or hide, which is necessitated by the environment that sends many negative signals. Some rare positive experiences will again not change his mood, which is consistent with the overall alarming signals coming from the environment.

In comparison, Spot is not doing too well. Whenever contradictory signals are received (e.g., a rare negative experience in an overall positive environment), he will try to move in two opposite directions and will not achieve anything, thus losing evolutionary competition to Tommy. This simple example demonstrates that values can be very useful even without an organism trying to make choices. Values help to assess the environment better because they can be *added* into some aggregate (similar to the average) and allow to have a more realistic view of the world. Thus, the reason behind the existence of the valuation system might be its ability to aggregate information. We discuss details and extensions in Appendix C.

4.2 Associative Memory: Freddie

Tommy has a formidable mind that is capable of aggregating information from the environment and use it to act more purposefully than Spot. In fact, Tommy's mind is so formidable that the next three levels of more advanced affective minds are the enhancements of Tommy rather than some conceptually new developments.

Despite this, Tommy has drawbacks. One of them is the inability to "predict the future." By this we mean that Tommy can react only to features that are currently active, and he cannot react to anything else, like some inactive features that could provide valuable information about what

is going on in the environment. To illustrate, suppose Tommy walks in the forest and sees the fresh footprint of a bear. It is likely that footprint as such has zero value, because it does not present immediate danger or reward. Thus, Tommy will ignore the footprint and keep walking as if nothing happened. However, this is potentially not a survival enhancing behavior since the footprint obviously suggests that there is a bear close by. Tommy will react to a bear only when he sees it, but at this point it might be too late to escape. As a result, Tommy's inability to associate footprints with bears can lead to a bad outcome.

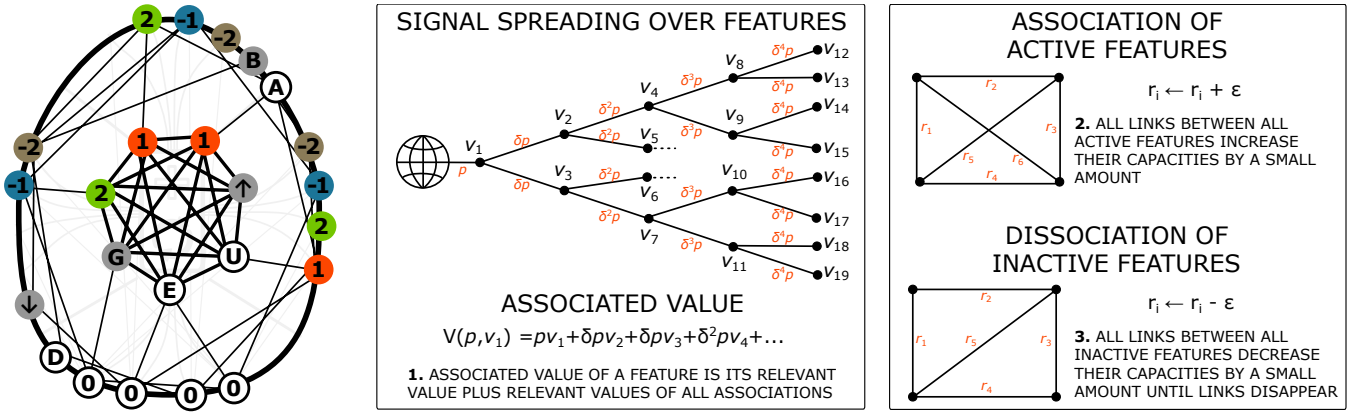

Figure 3: **Left Panel.** A representation of Freddie. **Middle Panel.** The spreading of the signal from active feature v_1 on the associative network. **Right Panel.** Increase in capacities of active features' links and decrease in capacities of inactive features' links.

This problem can be resolved if the mind could *associate* features with each other, thus forming *associative memories*. For example, if at some point a mind sees a bear and a footprint, it could create an association between the two sensory features that represent them: an associative link. Such link could help to retrieve information about the bear when seeing the footprint and react to it appropriately.

In this section, we describe Freddie, whose mind is capable of creating associations between any features, be they sensory or action-related. This is essentially the only additional characteristic of Freddie that nevertheless gives him an advantage over Tommy because Freddie can understand from secondary cues that important features are present in the environment (like bears, food, or something else). The left panel of Figure 3 shows a simple representation of Freddie. As before, we greyed out the graphical representations of Spot and Tommy on the picture since they are both parts of Freddie. The existing associations between features are shown with links between them. We put some features in the middle of the picture to emphasize that the associative network thus formed can have any kind of topology. The seven features in the middle, for example, are highly associated with each other, which will play an important role for Freddie as well as more complex minds presented below. But it is also possible that some features are loosely associated (with links that carry low capacity r), or associated with only few other fea-

tures. In Appendix A, we provide more details about the possible topologies of the associative network and what it implies for behavior and the ability to represent abstract concepts.

The middle panel of Figure 3 shows how Freddie computes values of features. Unlike Tommy, who perceives only relevant values of the active features, Freddie can also perceive features *associated* with the active ones. The example in the middle panel shows one active feature with value v_1 activated from the environment (represented by the Earth symbol) and with relevance p. Once this feature is activated, it sends signals through all links that are attached to it. We assume that the relevance of the "retransmitted" signal is lower than the original relevance coming from the environment. Specifically, we assume that whenever any feature is activated by a signal with some relevance p, it is retransmitted further with relevance δp, where $0 < \delta < 1$ is some constant. This assures that the signal will die out with time. We believe this is important because if the signal does not decay, it will eventually light up the whole associative network and Freddie would have an epileptic seizure. Plus, lighting up the whole network is not very useful anyway, because then Freddie would feel the same about all possible features.

As the features associated with the original feature v_1 get activated, Tommy inside Freddie aggregates them into an *associated value*, which is defined as

$$V(p, v_1) = pv_1 + \delta pv_2 + \delta pv_3 + \delta^2 pv_4 + ...$$

This is the sum of the relevant value pv_1 of the original active feature v_1 and the relevant values of all associated features. We do not specify here how many levels of associations Freddie can perform: this is an empirical question. We leave it for now and propose specific additional assumptions about this and alternative modeling choices in Appendix D. Notice as well that on the tree in the middle panel, not all features have to be different. When the signals spread through a highly associated features like the seven features in the left panel, they will be lit up multiple times by signals coming from other features in the cluster. Thus, the values on the tree can repeat themselves. This suggests that highly associated clusters will be more prominent within the associated value since the features inside them will be counted multiple times. This leads us to the idea of *concepts*, or collections of highly associated features, that we discuss in Appendix A.

What is important to understand about Freddie is that his mood, computed from associated values, depends not only on the features currently present in the environment, but also on the features that are not present but "imagined." This helps Freddie to extract more information about the current situation from his past experiences reflected in associated features.

This brings us to the discussion of how features get associated in the first place. The right panel of Figure 3 shows four active features that have associative links with capacities r_1 through r_6 (when features are not associated the capacity of the link between them is zero). We assume the simplest possible way of association. For any set of features active in the environment at

some moment in time, all links between them increase their capacity by some small positive number $\varepsilon > 0$. This means that when some of these features get activated in the future, they will potentially send signals of higher relevance to other features that were experienced together with them in the past. This in its turn will create associated value that can more realistically reflect the current situation. For example, if Freddie has footprints associated with a bear, then, upon seeing a footprint, Freddie will feel the negative value associated with the bear. This can put Freddie in a bad mood and he will turn around or avoid the place he was intending to go.

The opposite process also takes place. When associative links are not "used" and no signals pass through them for a long time, they deteriorate and their capacity decreases. Presumably, this is why we forget things that happened a long time ago: old links with low capacity do not activate features that we did not experience recently. We assume for simplicity that, at each moment in time, unused links decrease capacity by the same small number ε as shown on the right panel of Figure 3. In Appendix D we discuss more general cases.

In Freddie, we can already see characteristics pertaining to mammalian and human behavior. For example, Pavlov's dog, who salivates when hearing the bell, has been experiencing the sound and food together for many experimental trials. This has increased the capacity of the bell-food associative link to the degree when the bell can trigger the instinctual link (at the level of Spot) between food and salivation. Thus, we can conjecture that Pavlov's dog is a Freddie.

Human behavior is also often works at the level of Freddie. For example, living at one place for a long time develops *habits*. When we are exposed to the same features and activities at home many times over, the capacities of associative links increase and the actions, triggered by some features or other actions, are performed without deliberation. For example, many people have habitual "rituals" that they perform before going to sleep, like brushing their teeth, etc. These can be thought of as actions strongly associated with other actions in chain, so that each one triggers the next.

4.3 Episodic Memory: Molly

Freddie's mind is already complex enough to produce elements of human behavior like habits and associative predictions about what will happen. However, Freddie has a problem, which is strangely the opposite to Spot's, who was trying to execute incompatible actions. As Freddie's associative network gets more and more interconnected with time—which would be the case if Freddie lives in the same area, always sees the same features, and never travels—he stops being able to react to stimuli in different ways. This is simply because any features that such Freddie perceives will light up the whole large cluster of associated features, thus putting Freddie in the same mood no matter the stimulus. Overly interconnected Freddie will act in roughly the same way in all situations, which is not good for survival in *changing environments*. By trying

to solve Spot's problems with incompatible actions, evolution has eventually created organisms who perceive the world in the same way all the time.

We believe that *episodic memory* can resolve these issues. By episodic memory we mean the ability to recall *contexts*, or the collections of features that were present together in the past. One may say that Freddie already was solving this problem by strengthening associations between currently active features in a context. However, what Freddie does is not enough for the following reason. Contexts can often be interconnected in the sense that they share common features, but at the same time different enough so that different actions are required. For example, suppose that bears are dangerous and can attack you in early spring, when they are hungry, but are not dangerous in summer, when there is enough food around. So, bear in summer is different from bear in spring. However, Freddie whose associative network is overly interconnected will not be able to react differently to these two conditions, because when seeing a bear a large cluster of associated features lights up that covers both spring and summer experiences. As a result, Freddie will be in the same mood when seeing a bear in spring and summer and may not react correctly. Episodic memory can resolve this by creating stronger associations with a specific context experienced in the past that is similar to the current situation and thus evoking more specific associated values. In this case, bear in the spring can have different associated value than bear in summer.

In this section we discuss Molly, whose mind is able to store episodic memories. Molly has a new ability that was absent in the previous minds. Specifically, she can create *memory features*. A memory feature is connected to all features that were present in the environment at some moment in time. Memory features are also connected with each other in a chain that gets appended with each new memory. Apart from this, memory features function in the exactly same way as all other features. Specifically, when signals spread over the associative network, they spread also through memory features, thus evoking episodes from the past.

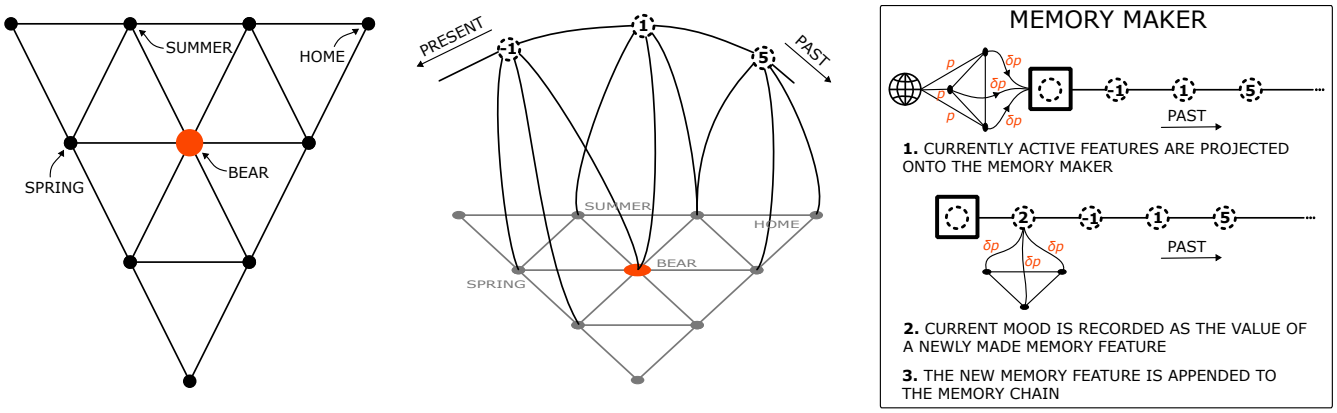

Figure 4: **Left Panel.** A highly-interconnected associative network. **Middle Panel.** Same network with memory features. **Right Panel.** The process of making a new episodic memory.

Consider a highly-connected associative network shown in the left panel of Figure 4. The feature Bear in the middle is connected to many other features including Summer and Spring because bears are present in both seasons. Notice that Freddie who has such a network will have very similar associated value for Bear in both spring and summer (for example, when features Summer and Bear are active versus when Spring and Bear are active). This is simply because all these features are connected to each other and light each other up multiple times when the signal from feature Bear is spreading. Thus, Freddie will most likely react to a bear similarly in both seasons. Moreover, when Freddie is at Home and experiences three features in the top right corner (there is no bear), his associated value of home will also heavily depend on the value of Bear because the features are highly interconnected. So, Freddie might be scared of bears and suffer unnecessarily even when he is in a safe environment.

Now let us consider Molly's associative network in the middle panel where there are three memory features with values -1, 1, and 5 that are linked to all features that were active at that time (shown with curved lines). The first memory feature with negative value was made in spring when Molly experienced an angry bear. The second with positive value was made in summer when Molly saw a bear, but it was not dangerous. Finally, the third memory was made when Molly is safely at home where she feels really good (value 5).

These memories can help Molly to distinguish the three different conditions better than Freddie. Suppose that Molly experiences features Bear and Spring. Then she will feel the same associated value as Freddie plus additional relevant values related to the memory feature with value -1 and the third feature connected to the memory. Similarly when Bear and Summer are active, there are additional terms added to the associated values: the memory feature with value 1 and the third feature in the memory. As a result, Molly can be in good mood in summer and in bad mood in spring (when seeing a bear), which will make her avoid bears in spring and not avoid them in summer. In addition when Molly is at home, she will not be scared of bears because she will feel an additional high relevant value coming from the memory feature. Thus, episodic memories can help to focus better on the features that relate to the current context.

We can also notice that, given a context, Molly is capable of remembering not only the contexts similar to the current one, but also contexts that happened immediately before or after the memory that got remembered. So, Molly can associate current context with possible future consequences that happened in the past in similar contexts, or something that preceded similar contexts. This gives Molly the ability to perceive time, something Freddie is incapable of.

To see in more detail how Molly's mind works, let us look at the right panel of Figure 4. We assume that Molly possesses an additional organ, the *memory maker* that produces new features and attaches to them the links to the current context. On the picture we see three active features coming from the environment (the globe symbol) that all have relevances p (for simplicity; they can be anything). Following assumptions about signal spreading on the network, we assume that these relevances get weaker (multiplied by δ) before getting to the memory maker. Once

memory maker receives the signals, it produces a new feature with value equal to the current mood, or the associated value of the current context, and stores that as the value of the newly produced memory feature. After that, the new memory feature gets appended to the *memory chain*, a sequence of memory features from the past that are associated with each other in a linear fashion.

All these steps take place automatically whenever a new memory should be made. We leave it for the future research to determine what conditions should be satisfied for new memories to form, but discuss additional assumptions on the memory maker in Appendix E.

4.4 Affective Behavior

At this point in our presentation, many readers might have questions about how the *affective devices* described above—namely values, associative memory, and episodic memory—are related to behavior and concepts that they know from various social sciences. In this section, we will try to provide some connections to the existing ideas in economics and psychology to make the exposition easier.

An economist might wonder how, when, and if at all Molly maximizes something. In economics tradition, all behavior is conceived as a result of some utility maximization, and it is a legitimate question why we never consider this idea. The difference between the general economics approach and ours lies in the framework that is used and how actions are conceptualized. Economic agents are modeled as decision-makers, who face some choice among given, fixed alternatives, and the question economists typically ask is how a decision-maker chooses one of the available options. In such framework, where agents are stuck with an abstract choice that they must make, it is indeed logical and intuitive to think of maximization as the way of choosing. However, if we consider a broader setup that we define above, where actions are not given and fixed, but rather arise *endogenously* from the affective mind and the environment, organisms do not have to maximize utility to get what they want. An entirely different mechanism is in place that produces some form of "optimal" behavior that does not require maximization.

Starting with Tommy, we talked about general classes of behaviors that can be termed roughly as *approaching* and *avoiding*. Tommy avoids features that have negative values and approaches features that have positive values (the sign of the value comes from Spot and deeper evolutionary levels). So, we assume that Tommy is free to do what he wants and move around in his environment as he pleases without being forced to make any specific choice among a preset collection of alternative at all.

These general tendencies, when considered on the levels of Freddie and Molly, can generate behaviors that might seem eeriely similar to maximization but are actually not it at all. To understand why, we should remember that Molly (or Freddie) strengthens her associations among features whenever she sees them and records new memories that associate them even more.

Given that Molly in general approaches features that have positive values and avoids features that have negative values, she will experience features that she likes *more often* than features that she does not like. Thus, Molly will have more associations in her mind with features that have positive values and less associations with features that have negative values. This is simply because avoiding certain features is the same as not creating associations with them (since they are avoided).

So, Molly has many associations with pleasant features and few associations with unpleasant features. This implies that whenever Molly faces some context or situation, she will be reminded of good features more often and will move in the direction where good features are, thus strengthening associations with good features even more while any associations with bad features that are being avoided will gradually decay. As a result, Molly will spend most of her time "hanging out" in places where features provide a lot of positive values and will not hang out in places that give her negative values.

From an economics perspective this can be interpreted as if Molly has positive utility of some places and negative utility of other places and that she "chooses" to hang out where the utility is higher. However, the process that generates her behavior has nothing to do with maximization whatsoever. Molly automatically moves where her associations take her, and since she mostly has associations in places that she approaches, she will automatically hang out there more than in places that she avoids. In other words, the very process of associating and creating memories dictates Molly's behavior without any choice.

A psychologist may wonder how our affective devices relate to psychological concepts like emotions and feelings. The answer to this not very hard to provide. In our framework, emotions can be related to the valuation system. For example, we can say that Molly has *instantaneous emotion* whenever the derivative features activate, or when Molly perceives a difference in values and updates them. In relation to human behavior, instantaneous emotions thus defined can be called anger, joy, sudden pain, elation or something else depending on the context. In our theory, these terms are just names for the same activation of derivative features in different contexts. So, from this perspective all negative and all positive instantaneous emotions have the exact same mechanism, it is just that culturally we learned to call them differently depending on the context in which these emotions are perceived.

We can also say, for example, that Molly has *feelings* when she recalls some episode from her memory triggered by some feature in the environment. For example, Molly looks at the photograph of her grandchildren and has warm feelings, because the picture reminds her of the times when she visited them. The memory episode with positive value is associated with the photograph and is registered in the valuation system. So, we can for example say that feelings arise from associated features, whereas emotions from the features coming from the environment.

Finally, Molly has good and bad mood features one of which is constantly active. In this case, we can say that Molly has either positive or negative *emotional state*. The state is positive when the good mood feature is active and negative when the bad mood feature is.

We do not believe that these definitions should be taken too seriously by psychologists. Moreover, we think that future research should clarify the connections between the existing psychological ideas and our model and hope that the framework we propose is rich enough to encapsulate different psychological concepts thus synchronizing various lines of research.

5 Language

All minds that we considered up to now live and act in their environment as solitary organisms learning to react to outside stimuli. This is of course not very realistic as all organisms need to procreate, which means that they need to get involved with their conspecifics one way or another. We discuss this extension in Appendix C. On top of that, many organisms including humans are also engaging with conspecifics for other reasons. They, for example, care for their young, warn others about coming danger, work together on common tasks (e.g., hunting), etc.

We propose that, for pursuing all these activities, organisms use *language*, which is defined in our framework as the special *ability to express perceived features with action*. This ability creates the flow of information that goes in the opposite direction to what we considered above. Usually, organisms receive information *from* the environment and update their associative network (values, associations), which may also lead to some action. In case of language however, organisms perceive some features *within their minds*, which in principle does not have to be related to outside stimuli, and act in order to pass what they feel *to* the environment. It is important that organisms *act* in some way to pass the information for otherwise there is no way for others to figure out that something is being signalled to them. For example, when we see written text, we perceive information sent to us by someone who acted before (by writing), it is just that, with writing, information can be perceived any time after someone chose to act and not only at the exact moment of action.

Not all species use language for things not related to procreation. Some animals lead solitary lives and only meet conspecifics for the purpose of mating. But since humans are a social species, heavily dependent on language, from now on we will focus on organisms that evolved to "talk" to each other for whatever reason that may be. From the section on cognition below it will also become clear that language plays an important role in making cognitive processes much more efficient than they would have been otherwise.

To imagine how language ability could have evolved, we should first recognize that the purpose of language is to send information to and receive it from other individuals of the same species. Language is built from the set of actions available to an organism. Thus, there should be a mechanism that would allow organisms to distinguish between actions intended to send

information and other actions that are intended for something else. This refers not only to what organism wants to say, but also to understanding others. We believe that the first prerequisite to the emergence of language defined in this way is the ability to *single out conspecifics* from other features present in the environment. This ability determines *who* can talk or be talked to. The second prerequisite is the ability to *interpret* what others are saying. This ability allows organisms to understand information sent to them. If an organism can do these two things, then language can pick up and evolve.

It is not the purpose of this paper to investigate how these two abilities came about since our main goal is to build a theory of human mind, and humans already have these abilities. Thus, we will assume within the model that these abilities are already in place. What interests us though is what *kinds* of languages can emerge from these assumptions and what it implies for human social behavior.

As with Spot, where we assumed that evolution has already built certain associative links between features and actions, we start from the most basic, innate, but very important form of communication that happens between parents and babies. Communication of this sort presumably exists in all species where parents spend effort on raising their young (like humans do). Thus, we assume that 1) babies have a "built-in" ability to recognize parents and send information about their condition to them and 2) parents have a built-in ability to understand this information and act on it. For example, (human) babies cry when they feel pain or need something, and laugh when they are happy and content. Thus, babies deliberately act to send information about their feelings to the environment, which means that they use language. At the same time, parents react to cries and laughter by attending to babies' needs or by continuing to provide stimuli that make babies happy. Parents understand this language.

This example is important, because it shows that we have a built-in ability to express what is on our minds and to read the expressions of others. Moreover, we *want* to express our feelings as if something inside tells us to do it. For example, when people are upset, they might go and publicly protest or they might strive to express their negative emotions in some other way (e.g., by slamming doors or using expletives). When people are happy, they gather for a celebration and try to signal others that they feel good by singing, dancing, or laughing. According to our definition above, all these expressions are part of a language that delivers certain information about perceived features and their values to others.

We suggest that babies and adults use the same *affective language* that their minds are equipped with from birth. As babies grow up and gain experience about the world, their repertoire of words, or actions used for communication, increases and their language becomes more expressive. As babies turn into adults, their affective language serves as the foundation for more complex *cognitive language* that uses words according to some rules (grammar) to deliver high-fidelity information. In this section, we describe affective language. Cognitive language is discussed in Appendix F.

5.1 Language Handler

To introduce language into our theory, we assume that the mind is equipped with a special module, the *language handler,* that allows organisms to recognize conspecifics, whom we will call *agents,* and to use certain actions to send and receive information that we will call *words.* The language handler has three main functions: 1) to recognize, keep track, and update values of sensory features that count as agents; 2) to recognize, keep track, and update values of action features that count as words; and 3) to suggest words that should be used in the presence of some agents. This description already suggests that the language handler is very similar to the value aggregator introduced in Tommy with a difference that it deals with a subset of features that have special social meaning (agents and words). And indeed we will model language handler in exactly same way as the value aggregator with the idea that it simply *is* a value aggregator only used for a specific purpose.

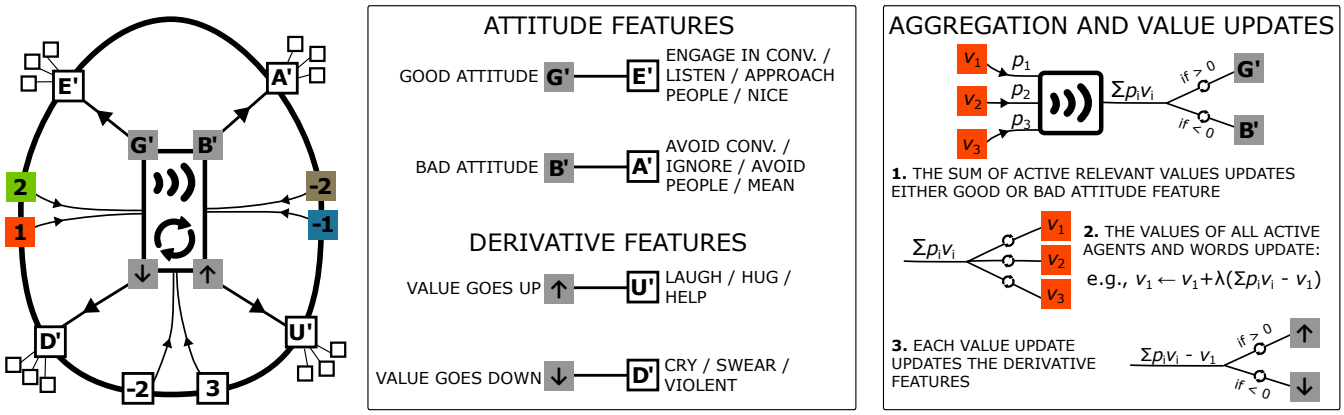

Figure 5: **Left Panel.** Language handler. **Middle Panel.** Explanations of new action features. **Right Panel.** The process of aggregation and value updates.

The left panel of Figure 5 shows how language handler works. It is the same idea as in value aggregator described above, only now we use colored squares to represent agents and white squares to represent words. Values of agents can be positive or negative. For example, a baby might attach positive values 1 and 2 to her parents and values -1 and -2 to strangers. Similarly with words: expletives, rude gestures, and harmful social actions might carry a negative value, like for example -2, while praises, hugs, and kind words might bring positive value, like 3. Notice that we do not describe the process by which features become agents and actions get marked as words. This is very culturally dependent and needs future research.

The language handler comes with new *attitude features* G' and B'. We use the word "attitude" to represent social mood, or feeling of friendly or hostile social environment. The new action features E', A', D', and U' play the same role as similar actions in the value aggregator. They serve as features to which words are attached that should be used in different social situations (represented by triplets of small squares). When we have good attitude (we are surrounded by nice people that we like) we might approach them, start a conversation, be nice, etc. When we

are surrounded by people we do not like, we have bad attitude and try to avoid or ignore them, or if we cannot, we might be mean to them. The middle panel of Figure 5 shows this graphically.

The derivative features marked by upwards and downwards arrows activate with the changes in social environment and when updates happen to the values of agents and words (see the right panel of the figure). For example, when some agent you like makes a rude gesture at you, this is unexpected, and you might lower the value of this agent while feeling betrayed, crying, and swearing back at him (downwards arrow feature activates). Or when a stranger, who might originally have a negative value, helps you, you might feel thankful, laugh, hug him, and update his value up (the upwards arrow feature activates).

In summary, we propose that social environment is treated by the mind in the same way as non-social, but with a designated language handler that keeps track of social interactions separately from everything else (that value aggregator deals with). One important issue that we cannot resolve without further research though is the *connection between the value aggregator and the language handler.* Are values of agents and words influenced by individual experiences and vice versa? For example, when we go to a meeting with a stomach ache, do we project the perceived individual pain on others and then treat them badly? It seems likely that this can be the case in some situations. However, when we go to a yoga class and the instructor tells us to do something that causes muscle pain, we might actually respect him more because he helps us become healthier. This connection does not play much role in the rest of the paper, so we leave this very interesting question for future research.

5.2 Affective Language: Talking Molly

In this section, we use the idea of language handler to describe affective language that Molly is capable of. To distinguish Molly with language from mute Molly introduced above, we call her Talking Molly. We will also focus on verbal communication and words in the common sense of the word, or words that are pronounced and written, and leave gestures and other social actions aside. This will become important in the next section where we talk about cognition that uses written or spoken words to represent and operate with concepts. The main questions we want to address in this section are how Talking Molly (and Talking Freddie before her) could learn new words, what they mean to them, and how they use them in speech thus forming affective language.

We start with Talking Freddie. Consider the left panel in Figure 6 that shows a piece of Freddie's associative network that involves feature Bear in different conditions like Spring and Summer. To understand what happens when Freddie tries to learn the word "bear" suppose that someone says "bear" whenever Freddie sees one. This, in a sense, is similar to the Pavlov's experiment with his dog who was learning to associate bell ring with food. Since Freddie treats all features in the same way, the pronunciation of the word "bear," that is represented by a new

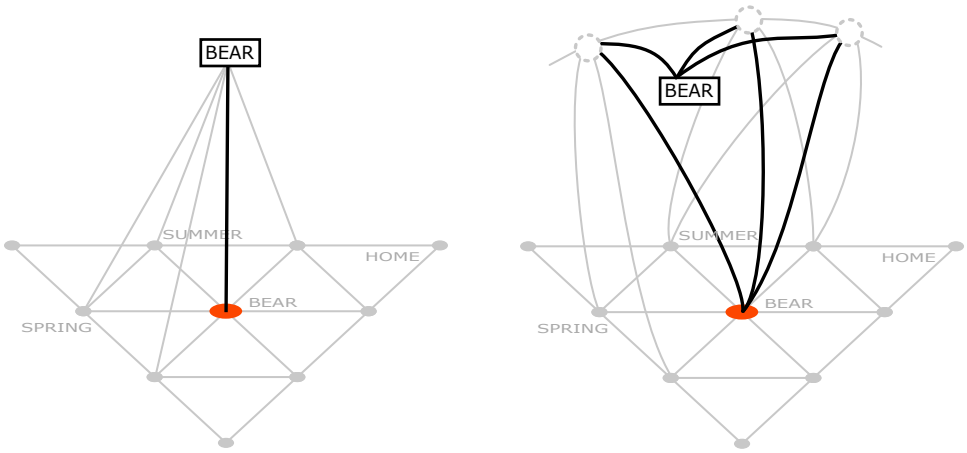

Figure 6: **Left Panel.** The word "bear" associated with features on Freddie's network. **Right Panel.** Additional links with the word "bear" introduced in Molly.

auditory feature shown in the rectangle on the figure, will associate with all features currently active in the environment. So, if Freddie always sees bears only in one fixed condition, for example in Summer, he will not be able to understand that this word even refers to a bear since the associative links between word "bear" and all features of the environment will strengthen to the same degree (see the right panel of Figure 3). Freddie will associate the word with any feature present in the Summer context. Now, if Freddie also hears the word when he sees a bear in Spring, he will also strengthen his associations between the word and the features in the Spring context that also includes feature Bear. Thus, all features in both Summer and Spring contexts will associate with the word to an equal degree except for the feature Bear that will associate with the word more, given that Bear is present in both types of contexts. This is emphasized on the left panel of Figure 6 with a thick black link.

From this, we can deduce that Freddie is not going to be very good at associating the word "bear" with feature Bear unless it is repeated over and over again in many different contexts. In other words, Freddie will need extensive experience in very different contexts to strongly associate the word with the sensory feature it is supposed to represent. It is unlikely that Freddie can achieve that since most organisms, including humans, do not normally experience many different contexts, but rather live in a more or less fixed environment and see bears in only few contexts that share many other features in common.

So, what kind of language should we expect Freddie to have? Interestingly, there are features that are always present in most contexts and that can get associated with words in this way. These are features G, B, G', B' from the value aggregator and the language handler. Indeed, by construction, some of these features are always active since Freddie always perceives something and might often be around other agents in very different circumstances. Thus, it is possible that a group of Freddies might develop words that are strongly associated with these features, words that mean good environment, bad environment, good social environment, and bad social

environment. The same can be said about the four derivative features that can also get strongly associated with words.

For example, expletives are words that people say automatically when they suddenly get upset (negative derivative features). So, we suggest that Freddies can have a limited language that attaches words to mood, attitude, and changes in them. And indeed, we are all familiar with this language. It is the language of facial expressions. People have very specific ways of showing their mood (e.g., happy, sad), social attitude (e.g., angry, empathetic), and the changes in them (e.g., laughter, crying).

The situation with Talking Molly is different. As shown in the right panel of Figure 6, Molly has additional links that connect to experienced features through the memory chain (the links between Bear and "bear" are emphasized as thick black curves). Imagine that Talking Molly hears the word "bear" whenever a bear is present in the same way as Talking Freddie. Each time this happens, Molly will form a new episodic memory that has Bear and "bear" present together. Multiple memory episodes of this kind would allow Molly to associate Bear with the word easier because whenever the word is pronounced, the associative signals go through the memory chain and hit the feature Bear multiple times which will make it light up much more than in case of Freddie. So, the more memories of Bear and "bear" Molly has, the more she will be able to separate these two features from the rest, simply because episodic memories can reinforce the links between co-occurring features much more than Freddie's associative memory can. Thus, we can conclude that Talking Molly should be able to learn more words than Talking Freddie given the advantage that episodic memory provides. However, Talking Molly will still need to experience many different contexts that contain Bear and "bear" in order to learn and will in general have other associations with "bear" coming from other features.

To summarize, it is possible to imagine that limited language can appear among groups of Talking Mollies and Talking Freddies. While the latter might be able to express only words related to their mood and changes in it (both individual and social), the former might be able to learn more words that describe the features of their environment that happen often in different contexts.

Now, consider a group of Talking Mollies who learned a repertoire of words. The question now is how will they talk with each other? In other words, how would a friendly conversation between two Talking Mollies look like? Suppose two Talking Mollies meet in the forest and see some features that they have words for. As features from the environment activate, they will associate in some order with the words that Mollies have learned. As the words get activated, they will be automatically pronounced since words themselves are actions related to pronunciations of the words. So, we should expect that Talking Mollies will randomly pronounce words related to anything they see. In addition, Mollies will associate the features in the environment with other features stored in their associative networks that might also have words related to them. Thus, Mollies might also say words that mean something not present in the environment,

but associated with it. The conversation then will sound like two streams of words in random order that depict what Mollies are seeing and associating the features in the environment with.

It may seem that this kind of conversation is not very helpful. However, information still gets passed between Talking Mollies. For example, if two Mollies look in different directions (e.g., watching for predators) they will alert one another whenever one sees a predator. Moreover, Mollies will share their associations that might be not the same, given different experiences that Mollies had in their lives, and thus learn new things since they will associate the words they hear from the other with the current environment. So, Talking Mollies are able to learn the experiences of other Talking Mollies and learn in this new way.

6 Cognition

If we look back at the stages of mind development from Tommy to Freddie to Molly, we can notice that these evolutionary innovations had one general purpose: aggregation of information about some specific context necessary for better actions. Tommy aggregated values and thus learned to feel his body and to make better choices taking many pieces of information into account at once. Freddie learned to associate features present in the context with other features that are stored in his mind, thus aggregating even more information relevant for the current context. Molly went even further by creating special memory features to remember and focus on contexts even better.

In the previous section, we saw that these developments bring advantages when organisms try to converse with each other by using language. Freddie is not as good as Molly at memorizing words given his inability to focus specifically on the features coming from the current context. His mind architecture, that mixes contexts with each other by associating all currently active features, ends up being too interconnected to distinguish them well, which leads to his inability to memorize words. Molly does a better job by using episodic memory, but has similar issues: the words are still associated with features that might be not relevant for the action at the moment.

We suggest that cognition represents the next level of the ability to focus on relevant information given specific context. In the end, features and associations between them can be seen as organism's database assembled through experience that needs to be accessed to retrieve currently useful information. The better the organism is at getting the relevant pieces, the better its survival chances will be.

In this final section of the paper, we show how special cognitive devices can help to reach the level of precision of information retrieval that is generally consistent with human cognitive abilities. While Tommy, Freddie, and Molly remind us more of animal minds rather than human ones, Alice, Esmeralda, and Robin will show characteristics inherent to humans. As we will see below, all new mind devices that allow to achieve this are introduced in Alice; Esmer-

alda and Robin are just "software" rather than "hardware" upgrades. Thus, with training and perseverance Alice can become Esmeralda or even Robin. It is interesting to note that with the new cognitive devices in Alice, mind becomes more like a self-programmable computer that can choose to train itself in myriad different ways (like installing new software), which produces the incredible diversity of amazing cognitive abilities that we observe in humanity. At the same time, we should not forget that Alice, Esmeralda, and Robin, as smart as they may be, still contain Spot, Tommy, Freddie, and Molly inside them. These automatic and affective minds are fully functional and do their jobs as designed by evolution. Cognitive processes organically mix up with affective and automatic processes, with affect and automatism sometimes prevailing over cognition, thus producing what we call human nature.

6.1 Focus, Concentration, Choice, Empathy: Alice

6.1.1 Focus and Concentration

We propose that the main device that allows for cognitive processing consists of one new feature called Goal and two new action features called Focus and Concentration. Notice that Goal is a built-in feature that makes us feel goal-oriented when it activates. When Goal is active, we feel purposeful and seek for tasks to do. Actions Focus and Concentration are connected to Goal with built-in associative links and activate whenever Goal is active. Focus and Concentration work like muscles, in the sense that different levels of their activation bring about different levels of focus and concentration. When we are focused, we see the features related to the task very vividly in our imagination. And when we are concentrated, the features not related to the task fade and we stop noticing them.

To give an example, suppose that you are taking a ride on a train where you are trying to read a book while listening to some music in your headphones. So, your goal is to understand what the book says while being slightly entertained by the music, which also blocks noises from other passengers. Suppose that the train car is quiet and no one disturbs you. In this case, you can read the book in an unfocused way: the words from the book and the music mix up and you perceive both with equal intensity. As a result, you understand some parts of what you read and notice some pieces of the music. Now suppose that you reached some passage in the book that you want to focus on. You increase your focus by activating feature Focus, which makes the music fade somewhat. Now, you understand more of what is written at the expense of missing some nice pieces of music that you like. Finally, suppose that you reached an important definition in the book that you really want to memorize. In this case, you concentrate on the text by activating feature Concentrate. Music fades (even though it is still playing) and your entire imagination is filled with concepts in the definition.

This example shows the main gist of how focus and concentration work. Consider the left panel of Figure 7 that shows a collection of features coming from the environment (the globe in

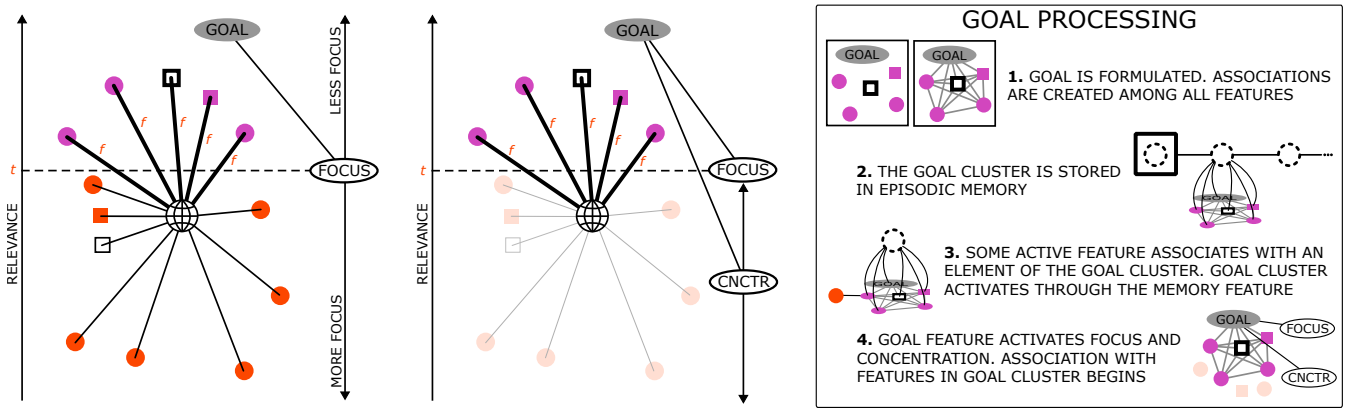

Figure 7: **Left Panel.** Focus increases relevances of all features with relevances above certain threshold to some high value f. **Middle Panel.** Concentration decreases relevances of all features below the threshold. **Right Panel.** The steps of goal processing.

the middle). There are simple sensory features marked with colored circles: the landscape you see from the window, the timetable that shows where you are travelling, etc. There are also two agents marked with colored squares: your friend, who sits next to you, and another passenger sitting behind. Both your friend and the other passenger are saying something, which is represented by two word features (white rectangles). On the graph all these features are ordered vertically by their relevance (the y-axis). Features with high relevance (e.g., music, book) are higher than features with low relevance (e.g., the other passenger, landscape). The x-axis does not mean anything.

We propose that action feature Focus works in the following way. When it is active at some level shown on the graph (white circle with the word "focus" in it), this means that all features with relevances higher than the threshold marked by the dashed line acquire an additional relevance boost. All features above the threshold become more relevant than they originally were. We can assume for simplicity that they all activate with some high relevance f represented on the graph by thick lines and magenta color of the features. So, at the threshold shown on the graph, you focus on the music, the book, your friend, and what he is saying. You also notice all other features to the normal degree. When you focus decreases (Focus is activated less) the threshold goes up and you might pay more attention to the words of your friend. When your focus increases (Focus is activated more), the threshold goes down, and you might start also paying attention to what the passenger behind talks about as well as other things.

The purpose of the action feature Concentration shown in the middle panel of Figure 7 is also to emphasize features above the threshold. The difference is that Concentration instead of increasing relevances of highly relevant features above the threshold, decreases relevances of features below it, which is shown with links in grey and faded colors of these features. High concentration (Concentration is very active) can decrease the relevances to zero, in which case

you will completely stop noticing them. When Concentration is not active, the features below the threshold have the original relevances dictated by the environment.

Together, Focus and Concentration allow you to single out the most relevant features in the environment (up to the threshold) and have them being presented in your imagination more vividly than they actually are. With Focus you can choose how many features to focus on, and with Concentration you can control how much of the less relevant features you want to register.

As we were saying above, the purpose of this mechanism is to obtain better quality information from the associative network about the features of interest. When you are focused and concentrated on a set of features, say the magenta ones on the graph, they start to associate with one another (Freddie) and produce associated values (Tommy). They also start being recorded to the episodic memory (Molly) that now contains not all features from the environment, but only those you have focused on. This ability allows you to memorize a definition from the book. If you focus on the words of the definition and concentrate so that all other features fade, your episodic memory will record only words of the definition that will now be connected to a memory feature. If you succeed at this task, you will be able to recall the definition well, because the words that it consists of will be well connected by associations and represented by a memory episode.

The mechanism described above is part of the cognitive mind we call Alice. Alice is able to focus and concentrate on some features while pursuing certain goals. But the question that naturally comes to mind at this point is *where* the goals come from and *who exactly decides* to focus, concentrate, and memorize definitions. It may seem that Alice cannot do this herself and that we need some additional mind that somehow thinks about these things outside the model and chooses to do them. However, we do not need that. There are two general ways in which the goals can be produced endogenously within Alice's mind.

The first way is rather automatic. It is plausible that goals are produced when some very relevant feature suddenly appears in the environment. Suppose that Alice does not follow any goals and her Focus and Concentration are inactive. This essentially turns her into a Talking Molly, since she does not use any cognitive mechanisms. The question is what does it mean for Focus to be inactive. It is plausible that in this case Alice perceives the world as Talking Molly with the exception that when some very relevant feature appears in the environment, for example a very annoying passenger, the relevance is so high that it goes above the threshold of inactive Focus. In this case, Alice will focus on the new very relevant feature and start the goal processing sequence described in the right panel of Figure 7. Thus, goals can be provided by the environment itself. And this is probably how cognition originally activates in babies and small children.

The second way is internal. Suppose Alice has some goals already stored in her episodic memory as shown in the third step of the right panel. Then, as she walks around she might come by a feature that is associated with something within some goal cluster (see the right panel). For

example, Alice walks by a grocery store and remembers that she needs to buy some vegetables. The feature Grocery Store activates the feature Vegetables that is part of the goal cluster "buy groceries." This way, environmental cues do not directly become goals, but rather activate previously stored goals by association. As a result, Alice goes into the grocery store and buys vegetables.

To quickly summarize, we do not need an additional "goal-setting" mind, as Alice is fully capable to acquire new goals when very relevant new features arrive, or when something in the environment reminds her of the old goals she has memorized.

The final question is how exactly Alice chooses to focus and concentrate on the current goal. We believe that this is a matter of training and upbringing. Not all people focus and concentrate in the same way. In fact, there is a large difference in these abilities across people. We believe that education is responsible for this. When Alice is a child, she spends many years in school where she is forced to focus and concentrate on various subjects. If she is a good student, she will train these abilities when told by her teacher, and they will become *habitual*. Indeed, Focus and Concentration are features as anything else. Thus, they can get associated with various other features in the environment. As Alice studies in school, she will develop various ways to activate her focus and concentration by, for example, teaching herself to study in trains while listening to music. So, well-taught Alice will develop associative routines, or *cognitive habits*, that switch on Focus and Concentration. Conversely, if Alice is not being taught in school and is left for herself, she might not develop such habits and will not use Focus and Concentration too often. The mind of untrained Alice might be closer in its characteristics to a Talking Molly than to the mind of a well-taught Alice.

6.1.2 Choice

Another cognitive ability that, we believe, Alice has is the ability to make choices. However, before we get to that we need to describe in more detail the mechanism that is used for it. We do not introduce any new devices specifically for choice, but rather dig deeper into the value updating mechanism, or *updater,* already introduced in Section 4.1 where we talked about valuation system within Tommy. Specifically, in this section we will talk about a part of the updater called *comparator*. In Appendix G we discuss in more detail the different possible versions of the updater, which also relates to how we model Tommy.

It would be hard to deny that choice involves some sort of comparison between available options. And indeed, in order to update values of features, as suggested in Section 4.1, we assumed that somehow the difference between the perceived value (mood) and the recorded value of a feature is computed in the mind, which was also used to construct derivative features. For expositional purposes, we did not specify how this exactly happens, but we do it now in this section, since these details are important for understanding choice.

We begin from a simple observation that comparisons are not made explicitly for choice. We are capable of making comparisons that are not directly value related. For example, we can ask: What is *redder* a cherry or a blueberry? The answer seems obvious, cherries are redder than blueberries that are not red at all. But then we can also ask: What is redder a cherry or a strawberry? This would depend on the specific berries in question, but we will still be able to provide an answer. Similar questions can be asked about anything. What is *more fun*, going to a party or to a music concert? What is *denser*, water or metal? Notice that, in principle, these questions are not related to any choice per se. However, we believe that the comparator used to execute such comparisons is also used for choice.

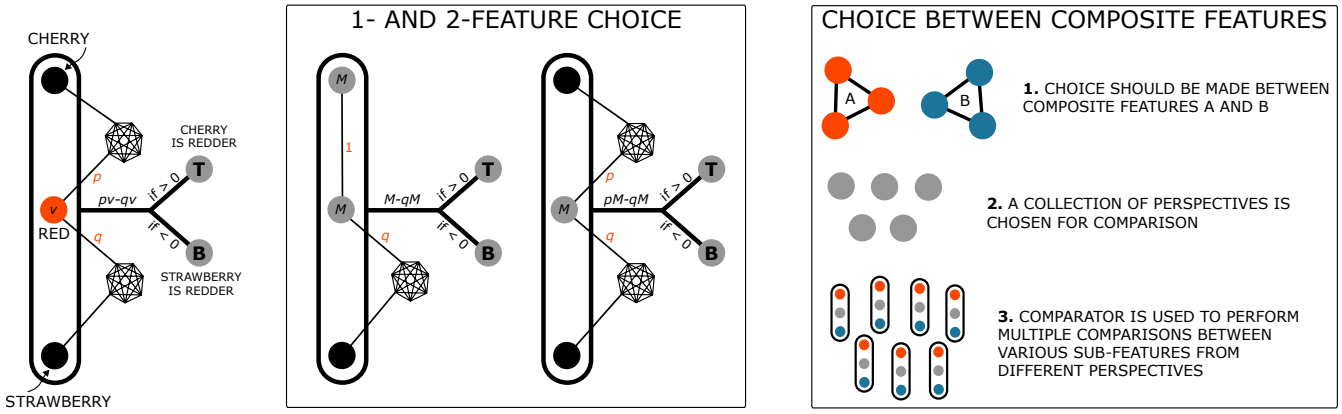

Figure 8: **Left Panel.** Comparator determines whether a cherry or a strawberry is redder. **Middle Panel.** Mood-related choice of one feature or choice between two features. **Right Panel.** Choice between composite features.

We did not discuss comparator before when we talked about Tommy because in Tommy it is used for only one specific purpose, namely, updating the values of features. However, it seems like cognition provides more direct access to this device and can use it more generally for any comparisons (though this claim needs research). As shown on the left panel of Figure 8, the comparator has three inputs: the two features that should be compared (features Cherry and Strawberry) and the *perspective* from which they should be compared, in this case it is the feature Color Red. The output of the comparator is the activation of one of the new features T or B (standing for top and bottom). If T activates, it means that Cherry is redder; if B activates it means that Strawberry is.

When Alice wants to determine which berry is redder, she focuses on the three input features and switches on the comparator. To make the comparison, comparator "listens" to the activation of the perspective feature Red while also activating each of the features to be compared in turn. First, comparator activates Cherry. The signal from Cherry spreads over the associative network represented on the picture with a little, fully-connected graph and checks the activation of Red. After spreading over the network, the signal finally hits the feature Red with relevance p. Then the comparator activates Strawberry and waits for the signal to hit Red again. In this case, feature

Red is activated with relevance q. Then the comparator subtracts the resulting relevant values pv and qv, where v is the value of Red. If $pv - qv = (p - q)v > 0$, then feature T is activated. If $(p - q)v < 0$, then feature B is activated.

At this point, Alice knows which berry is redder. If she feels that T is active, she decides that Cherry is redder than Strawberry and vice versa for feature B. Notice that, as long as $v > 0$, this boils down to the comparison of the relevances of the signals p and q. When $p > q$, it means that Red is more relevant for Cherry than for Strawberry and other way round when $p < q$. This makes sense. If Alice were to compare the redness of Cherry versus Blueberry, then the signal from Blueberry might never reach the feature Red at all because blueberries are blue and have nothing to do with red at all. So, in this case, we would have $q = 0$ and the conclusion would be made that Cherry is redder.

It is also possible that, when comparator listens to the activation of Red while activating Cherry, Red is hit multiple times from within the associative network. In this case, we can imagine that the relevances of these signals add up and the resulting aggregate relevance becomes p (same thing happens to q). Interestingly, the idea of collecting information in this way is present in so called *drift diffusion models*, well-known in neuroscience. Our description of the comparator matches this type of models almost exactly. So, we should expect that the comparator might share characteristics pertaining to this class of models.

Now, we can use this idea to think about economic choices where "utilities" of different alternatives are compared. To imagine how this is done, all we need is to take the perspective of the current mood that Alice is in. The simplest case of such choice, that we call *one-feature choice*, is shown on the left comparator in the middle panel of Figure 8. Suppose that Alice is walking down the street in a good mood and she sees an ice-cream shop. She wants to understand if she wants an ice-cream or not. In this case, she uses the perspective feature G (good mood) that also plays the role of one of the features to be compared. Given that the same feature G is the perspective and one of the compared features, the relevance of G with itself is 1, so the relevant value of G is just its value $M > 0$ (mood is good). Suppose that the relevance of ice-cream, as determined by the comparator is q. Then, the comparator will signal that ice-cream is desirable by activating feature B when $M - qM < 0$, or when $q > 1$. This means that if ice-cream is heavily associated with good mood (high relevance q) Alice will feel that she wants an ice-cream. If however, $q < 1$, or that Alice does not associate ice-cream with good mood much, then she will decide against getting one (we ignore the cost of the ice-cream for simplicity). The right comparator in the middle panel shows the *two-feature choice* from the mood perspective between two different features. Here, Alice will choose the feature that is more relevant for her when she is in a good mood (when $p > q$ she will choose the top feature and vice versa).

When Alice is in a bad mood things get different. Now $M < 0$ and the comparator will activate T and B in reverse given the same relevances of the compared features. However, the bad mood feature B might have different associations with features than the good mood feature

G, so we cannot say that all choices that Alice makes will be simply reversed. In general, when Alice in a bad mood decides about getting an ice-cream, she will get it whenever $M - qM < 0$, or when $q < 1$. So, Alice will get an ice-cream when the association with the bad mood feature is low, and not get it when the association is high. This also makes sense. In a bad mood, Alice tries to do things that are *not* associated with it. She will try to avoid being in bad mood and choose things that she normally does not perceive in such condition. Interestingly, this is reminiscent of the difference between gain and loss domains in Prospect Theory, which uses an assumption similar to this result.

Finally, the idea of the comparator can allow us to say something about choice between composite features (see Appendix A for details). Suppose Alice is choosing between laptops A and B that are shown on the right panel of Figure 8 with red and blue features connected to each other. These sub-features represent characteristics of the laptops. For example, they can be of different color, have different processors, memory, etc. Suppose as well that Alice determines a collection of perspectives to compare the two laptops shown as grey features. These can be for example, how Shiny the laptops are, how Small they are, how Fast, or how much they Cost. Then Alice can take any two sub-features (red and blue) and any perspective (grey), and use the comparator to make a comparison. Such comparisons can be made multiple times, in the limit exhausting all possible combinations.

We cannot say how exactly Alice will compare composite features. This depends on how she learned to do it in the past. For example, between any two features (like laptops) she might learn to always choose the shinier one. Or she might learn that, when it gets to computers, she should always choose the one that is faster. Or, she can write a long list of comparisons on a piece of paper and then use some complex aggregation procedure to figure out which laptop is better. One thing we can say though: Alice might not necessarily choose between laptops in accordance with her mood that would in our model count as an "economic" choice. Everything will depend on the *perspectives* that Alice is using.

6.1.3 Empathy

The idea of comparison of features from different perspectives explained in the previous section allows us to think about empathy and how it is implemented in the cognitive mind. By empathy we mean the ability to acquire understanding of how others would feel in certain circumstances and consequently taking it into account in choices. We believe that the roots of empathy lie in the affective mind that understands agency, namely in Talking Molly. If we go back to the story about parents and babies discussed in Section 5, it is not hard to imagine how empathy might work. The original incentives for parents to care for their babies—that we presume is a prerequisite for agency represented by the language handler—are most likely hard-wired into the mind. Organisms who raise their babies might attach very high values to them by default. This would make parents heavily associate surrounding features with their babies. Such *love*

might lead to a situation when parents, no matter what they do or where they are, keep getting reminded about babies through the associative network, which will make them take actions directed at increasing babies' well-being (in affective minds it might be an automatic Spot-level thing or more elaborate associative behavior developed later).

In Alice, who thinks by focusing on features with high relevance, having babies would mean that they are constantly "on her mind" since babies would have very high relevance and be mostly above the threshold set by the Focus feature (possibly even when Focus in inactive). So, whenever Alice focuses on any situation, choice, or comparison, babies will be there in her imagination and she will knowingly or unknowingly take their needs into account. The point is that the basic mechanism for empathy (towards babies) is already implemented within the language handler.

We suggest that this same mechanism works for empathy towards other unrelated individuals. When Alice is young, she might be taught by her parents and teachers that caring about others is important. So others will acquire status similar to babies in Alice's mind (maybe to a lesser extent). This would imply that all agents or only agents belonging to a specific group (e.g., family, tribe, nationality) will have very high relevance to her. This, in its turn, might lead to the development of a collection of cognitive habits related to thinking about others in various situations. Thus, cognitive minds do not have to have empathy towards others per se—and we are all well-aware that empathic abilities are very different across people—but rather, cognitive minds can develop empathy to different degrees depending on the upbringing.

In this section, we will consider a case when Alice was raised in some tradition of caring about others and has developed cognitive habits related to this. What interests us is the mechanisms that she might use to reflect empathy in her behavior. One possibility is the following. Suppose that, when Alice was in school, she memorized a rule "when making a choice, always think about how it might influence other agents." Then, any choice that Alice makes might look like a choice between composite features (as shown in the right panel of Figure 8) that include the actual features that Alice is choosing between and a collection of perspective features that includes other agents in it.

To illustrate with an example, suppose that Alice is deciding whether to buy cherries or strawberries for her guests. When she ponders this choice, she has all guests represented as potential perspective features because she remembered the rule above that she learned in school. The first thing she might do is to associate the collection of agents (guests) with cherries and strawberries. This thinking might produce episodic memories like "Bob is allergic to cherries." This is valuable information that makes Alice think that strawberries might be a better option. However, there are many other agents except Bob, and her associative network does not produce any more health-related episodes about specific guests and berries. So, Alice starts making comparisons from various perspectives.

First, she uses the comparator to decide what she likes more (inputs: Cherry, Strawberry; perspective: her mood feature). She focuses on the results and her preference gets recorded into episodic memory.

Then she might use other guests as a perspective feature instead. For example, inputs Cherry and Strawberry with perspective feature Kathy (one of the guests whom Alice likes a lot) gives very high positive difference $(p - q)v$. Kathy has a high value v since Alice likes her, which is multiplied by the difference in relevances $p - q$ of links from Cherry and Strawberry to Kathy. The Cherry-Kathy relevance p is very high because Alice has just seen Kathy eating cherries last Saturday. The Strawberry-Kathy relevance q is small because Alice has never seen, heard, or otherwise experienced Kathy and strawberries together. This comparison also gets stored in Alice's episodic memory.

Then, Alice thinks about Todd. Todd is a friend of a friend and she does not like him too much, so his value v is negative. When Alice uses the comparator with the perspective feature Todd it gives some value $(p - q)v$ that is also positive (suggesting cherries). This is because $p = 0$ (associating Todd and Cherry produces zero results). However, q is very high because Alice heard on another party recently that Todd really enjoys strawberry-scented shampoo. So, implicit in Alice's thinking is the idea that since Todd likes strawberries, but she does not like Todd, she should *not* buy strawberries and opt for cherries.

Finally, Alice considers all pieces of information together. Bob is allergic to cherries, but Kathy really likes them and she is a very good friend. Todd likes strawberries, which means that again cherries are better. In the end, Alice buys cherries thinking that Bob is less important than Kathy. This last comparison is made by using Bob and Kathy as input features and her own mood as the perspective.

This example shows how empathy can be implemented in the cognitive mind. However, it is important to note that Alice—even though she is fully dedicated to making sure that her choices do not harm others—is not probably considering everything possible to make her choice. She considers *some* possibilities that come easier to her mind given the time restriction (she is in the grocery store and needs to hurry to make it on time to the party). Thus, we should expect that cognitive minds, even empathic ones, will only have partial considerations about others when making choices.

6.2 Imagination: Esmeralda

As we mentioned at the beginning of Section 6, Alice possesses all mind devices to perform cognitive tasks, namely focus, concentration, and the ability to make choices and empathize. We also mentioned that cognition is not a skill given from birth, but rather something that is trained through long years of schooling. In this section, we discuss *imagination,* an ability that,

we believe, can be trained though research is needed to understand how trainable it is and what role "talent" plays in it.

In our framework, imagination is the ability to *focus* on features that are not active in the environment, but that are activated in the process of association through the associative network. This, in fact, might not be a trivial task. In Section 4.2, we discussed how signals spread over Freddie's associative network gradually decaying with the discounting parameter $0 < \delta < 1$. The fact that signal decays means that associated features lit up in the mind are less relevant than the original features from the environment that activate them. If δ is not very large and sufficiently close to zero, then Alice who chooses certain low focus threshold (only very relevant features are above it) might not register them at all since all associated features can fall below the threshold. However, Alice will still feel these associated features through her mood (the value aggregator) and her attitude (the language handler). Such Alice can feel happy or scared, or angry due to features that light up in the network, but she will not know why she feels these things, since there will be no "picture" in her imagination attached to the sensations.

This of course does not have to be this way. Esmeralda, an "artistic" version of Alice, can see and operate with things in her imagination really well. To see how she can do it, consider the left panel on Figure 7, where we assumed that Focus involves increasing relevances of the features active in the environment to some high value f. The fact that f is high implies through the formula for associated value in Section 4.2 that the relevances of all associated features will also be higher than usual since they all take the form $\delta^n f$, where n is the distance on the network from the original feature. Thus, given high enough f and high enough degree of Focus—so that the threshold is low and features with low relevances get registered in the imagination—Esmeralda can see the associated features. Specifically, for the relevance threshold t, Esmeralda will see all associated features with relevances $\delta^n f > t$. Or, she will see all associated features up to the distance $n < (\log t - \log f) / \log \delta$ on the network.

This argument suggests that in order to have vivid imagination we need to be able to focus really well, which implies high f and low t. However, just seeing things in your mind might not be enough. To use the fruits of her imagination, Esmeralda needs to store associated features in episodic memory. We can consider one *instance of imagination*. Suppose Esmeralda is focused on a set of features A in the environment (their relevances are above the threshold). Then, the set of associated features B appear above the threshold and get registered as sensations like images, smells, tastes, etc. At this point, Esmeralda needs to hold all features in A and B in her mind with relevances above the threshold long enough so that they get recorded into her episodic memory. If this works, the memory with features $A \cup B$ gets recorded.

This may not sound like a big deal, but it actually is. First, by recording a memory with $A \cup B$ Esmeralda has created something *new* that has never existed before: $A \cup B$ is a joint product of the features from the environment and the features from Esmeralda's associative network, which contains all her past experiences. So, $A \cup B$ is a product that mixes "reality" (represented by A)

and Esmeralda's mind (represented by B). This product is *unique* because each associative network in each human being is unique, given that each human being has unique life experiences that can never be repeated or replicated in other humans.

Second, the memory of $A \cup B$ allows Esmeralda to retrieve it in the future and work on it again. To realize the power that it gives her, imagine that you try to draw a complex image in Photoshop, but you do not have a save button. So, all you can do is to start anew each time you open Photoshop again. With the ability to save your work in a file you can work on the image in consecutive steps separated in time making it more and more elaborate with each new session. The ability to store imagined features in memory is exactly the same. Next day, Esmeralda can retrieve $A \cup B$ and continue imagining, but now using features in B as inputs and produce, say something like $A \cup B \cup C$, where set of features C is what her associative network produced from $A \cup B$. The process of imagining, recording imagined features, and retrieving them allows artists, writers, scientists, and anyone else whose job involves imagination to create very complex new objects such as paintings or books that we all enjoy so much.

6.3 Reasoning: Robin

Imagination is a fascinating cognitive ability and people really value and respect Esmeraldas who create beautiful pieces of art and science. However, imagining things takes a degree of courage. The reason is that the associative network stores not only features, but also their relevances and values. So, each time Esmeralda imagines some feature with value v and relevance p, she also *feels* its relevant value pv through the value aggregator or the language handler. In other words, an instance of imagination is always accompanied by the "reliving" of past experiences. All of us remember some traumatic episodes and reliving them might be not very pleasant or even horrifying. However, Esmeralda, if she wants to be good at her job, needs to feel these things again and again, which can have negative consequences for her mental health. For example, she can get depressed (very bad mood) when she recalls some bad experiences. Dealing with intense emotions can be exhausting and can even lead to substance abuse.

While in some imaginative tasks, like art, having a wide repertoire of emotions is a good thing, since the purpose of art is primarily to create objects that represent feelings, in other tasks like logical thinking and reasoning it might not be very helpful. Imagine a mathematician who for some reason feels terrified each time he imagines a parabola or a quadratic equation. His ability to reason logically in this case can be greatly diminished, because mathematicians need to operate with many abstract concepts quickly in order to prove a theorem or to develop a new theory. Feeling terror each minute in the process will most likely make the mathematician switch profession due to constant emotional distress. The same argument can be made if we replace the word "terror" above with "ecstasy." Thus, for some imaginative tasks having extreme emotions can be inefficient.

In this section we consider Robin, the mind that is able to reason and make discoveries in his (or her) imagination due to the specific properties of his associative network. One way to solve the problem with too much emotion in the process of imagination is to experience and store features in the network that do not have very positive or very negative relevant values (pv is positive and close to zero). To do that Robin needs to learn to control his mood (the value aggregator) and attitude (the language handler) in the process of learning the subject of his study. For example, in order to not be terrified by quadratic equations, Robin needs to be in a good mood, he needs to learn about them in a calm, quiet environment and use Concentration to remove all the noise coming from external and internal features. In this way, when Robin records an episodic memory containing the definition of the quadratic equation, the value of the memory will be a small positive number. We discuss the techniques that can be used to control the associative network in more detail in Appendix H.

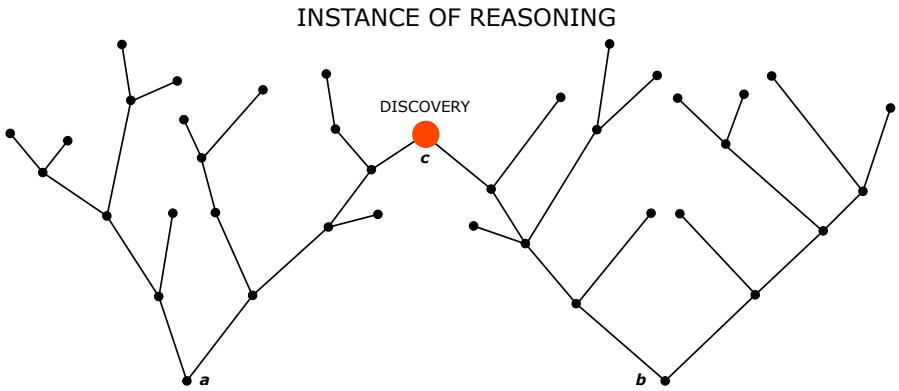

Figure 9: The process of reasoning in Robin.

Suppose that Robin was able to study mathematics for many years and had stored many mathematical concepts in his memory that all have low positive relevant values. The concepts, definitions, and theorems are well organized in his mind: the associations between them are not random, but highly structured and represent logical relationships needed to search the associative network in an efficient way. In this case, he can make discoveries as shown in Figure 9. Using high Focus and Concentration, Robin can completely detach himself from the real world and perceive only two features a and b in his imagination that can represent some mathematical objects. Notice as well that no mood or attitude features are present, they all get suppressed by Concentration. So, Robin is in a neutral state where he does not experience any emotions or sensations.

We define an *instance of reasoning* performed by Robin as the process of associating on this highly-structured and low-value associative network where imagination can roam free and move fast without constantly stumbling upon extreme values felt through the value aggregator or the language handler (if Robin felt some emotions they would bring about other unrelated associations thus obfuscating the thinking process). Robin begins with focusing on a and b and

continues to associate until the two associative processes converge *somewhere* on a common feature c that is hit by the signals from a and b simultaneously (we assume this for the sake of the argument). The fact that c gets activated from two associative links is important, because this is exactly what allows Robin to *notice* that c lies on both associative trees coming from a and b. The relevance of c will be higher than those of all intermediate nodes on each tree separately and thus go above the focus threshold, which allows Robin to see it in his imagination. This is when Robin has his eureka moment.

Notice that Robin does not know where or what c is going to be in the beginning of this process. In fact, we call c a *discovery* exactly because Robin has learned that features a and b have something in common, they both associate with c, which he did not know before. After c is discovered, Robin can focus on features a, b, and c and store the three of them as a memory episode. Thus, the instance of reasoning culminates with the discovery of c and recording it to the memory. Similarly to Esmeralda, this allows Robin to continue his reasoning process later when he can retrieve the memory and start thinking again.

When we talk about an instance of reasoning and the discovery of c, we do not allude to c as some great discovery like $E = mc^2$. In fact, it is almost never that. When trying to achieve some goal, like proving a theorem, Robin wants c to satisfy certain criteria. For example, he might seek for a mathematical object c that belongs to the same class of objects as a and b at the same time (a and b come from different classes). So, when a discovery c is made, Robin checks if the discovered object satisfies some criterion or criteria, for example "c should belong to class of a and to class of b." If c does not satisfy this, Robin starts anew (maybe modifying the original objects a and b). Thus, many instances of reasoning, hundreds or even thousands, might be necessary to get to the desired result.

According to this argument, Robin's thinking process—that may include many instances of reasoning, recording results to memory, retrieving the results, modifying the initial conditions, and checking the criteria—does not follow some predetermined path that always leads to success. In fact, it is more like random wandering in an unknown forest where you cannot see anything far away, but only things very close by. This is a *search* in a very complex landscape that rarely climaxes with important findings. Nevertheless and despite these complications, many Robins do succeed and achieve great things that contribute to the progress of humanity.

6.4 Self-Reflection and Self-Awareness

The final topic that we cover in this paper is concerned with self-reflection and self-awareness. These concepts may sound mysterious, surrounded by philosophical conundra, and defying logical explanation. Nevertheless, we believe that cognitive minds, as we see them in this paper, can naturally develop these abilities.

When we discussed Robin, we used mathematics as a leading example of reasoning ability. However, reasoning can be applied to anything and does not have to be deeply mathematical in its nature. In fact, the ability to self-reflect is probably present in most humans to various extent. To imagine how this might work, we first need to define a concept of *self* within the mind. One way this can be done is to assume that the mind can think of itself as an agent that is perceived by the language handler in the exactly same way it perceives other agents. Indeed, it does not take too much effort for any human being to realize that he or she lives in a body that is the same as bodies of other agents around, that this body does and says things similar to other agents, etc. Thus, we assume that the mind can create a Self feature, which is a standard agent feature, with the only difference that the mind has much more access to the information about what this agent thinks, does, or feels, as compared to information that can be obtained about other agents. Notice as well that the mind *does not have to* have a Self feature, and we believe that it can operate perfectly fine without it. Self-reflection is, in a sense, a hobby that anyone can entertain themselves with, but it is not something necessary for any cognitive processes.

Consider a Robin who, instead of learning mathematics, decided to become a philosopher. Robin has decided that to pursue this career he needs to understand himself and everything that happens inside his own mind as he feels it. To begin, Robin imagined that there is an agent called Self who acts, thinks, and feels as all other agents; and Robin made it his goal to construct a model of how Self works. To achieve this, Robin can do something like this. He creates various categories to classify own experiences. For example, he can use words like Pain, Pleasure, Anger, Shame, Irritation, Taste, Color, Touch etc. (all represented as word-features) to use as basic elements of his theory of self. Categories do not have to be these exactly and might depend on the culture that Robin was born into.

Then, he starts to gather data. Suppose Robin eats an ice-cream. As this is happening, features from the environment related to ice-cream get activated. For example, Robin feels that ice-cream is intensely cold on his tongue. This is a feature with very high relevance. So, Robin can Focus on this feeling together with the categories Cold and Tongue (he imagines written words), blocking all other features using Concentration. As a result, an episodic memory gets created that has three features in it: the sensory feature of feeling cold on the tongue and two words Cold and Tongue. The three features also get heavily associated since Robin focuses on them for a very long time. This instance of reasoning has generated a datum in his associative network (a memory feature) that is now a part of his growing theory of self.

Proceeding in this way, Robin can classify all his experiences, adding also his own behavior to the mix in the later stages. He can, for example, keep track of his mood when he wakes up (depending on what he did the day before), or what he does after someone made a rude gesture at him. With time, the sub-network of associations between categories and various sensory features, actions, other agents, attitudes, moods, etc. increases and Robin gradually becomes better and better at understanding how Self operates and what it does depending on circumstances.

At the same time, he might develop concepts of who Self is and create theories or models of Self's behavior. These models can actually help Robin to predict his own behavior and to avoid awkward social situations. We can call all these deliberations *self-reflection*.

Suppose now that Robin does this for a very long time. Thinking about Self becomes habitual and Robin reaches the state of *self-awareness*. We call it such, because this new overly self-reflecting Robin will feel differently than a "normal" Robin who never self-reflected before. A normal Robin, who does not focus on his own feelings too much, registers ice-cream as a slight activation of his positive derivative feature. As the sub-features in the ice-cream get activated, this activation percolates through normal Robin's associative network without much consequences and then decays. Nothing else changes. Normal Robin enjoyed an ice-cream, but this did not leave and additional mark on him.

However, in self-aware Robin eating ice-cream will create a very different activation. Once the sub-features within the ice-cream activate in his mind they immediately get associated with the concepts from his theory of Self. So, when self-aware Robin eats an ice-cream what goes through his mind is that the coldness of the ice-cream is 4 on the scale from 1 to 5, that the ice-cream shop switched from whole milk to powder, which makes the ice-cream taste bad, that keeping ice-cream on his tongue for a long time might cause tooth pain (this happened before), etc. All this new information also gets incorporated into the theory of Self.

With this example, we suggest that in the state of self-awareness some or all perceived features are associated with an additional network (Self theory) that Robin can easily construct. This also suggests that self-awareness does not have to be the same across all Robins, who chose to follow the path of knowing thyself. Some Robins might be very aware of their health, while others can be more aware of the consequences of their actions for other agents. It is likely, that such theories of Self are present in most human beings to various degrees.

6.5 Cognitive Behavior

In this last section, we summarize what our theory implies for cognitive behavior and how to study it. The main message that the theory provides is that cognition is not some monolithic ability that each and every human possesses in its entirety. Rather, we suggest that each human possesses *cognitive devices* that are utilized in cognitive processes. These devices are focus, concentration, choice, and empathy. However, just having these devices is far from enough. Like muscles in our bodies, cognition requires intensive training without which it is not going to be helpful. To develop cognitive abilities, children need schooling from birth. They need to be trained for many years to focus on the study material; to memorize abstract concepts and rules; to increase the ability to hold many things in their minds; to control their mood and attitude through concentration, etc. Strangely enough, cognition needs to be practiced to the extent that

it amalgamates with the affective system and turns into cognitive habits that allow people to use cognitive devices automatically.

Moreover, cognition critically depends on the associative network: the number and kind of features contained in it, and the ways they are connected to each other by associative links. When cognitive mind experiences little in its life, does not try to understand what is going on around, does not strive to learn more, and is in general ignorant, cognitive devices, even when trained well, will not produce anything valuable simply because they will be constrained by the amount of information that they can operate with. Therefore, in order for cognition to be effective people need to learn a lot about the world, diversify their knowledge, experience pleasure and pain, meet different kinds of people, and constantly analyze all this information to create a well-structured and conceptual picture of the world.

Unfortunately, few of us are lucky enough to afford doing all these things. Many people are born into places and circumstances that do not allow them to properly develop their cognitive abilities. As a result, they grow up without being able to understand the world around them, which can lead to bad decisions and suffering. From the moral perspective, this situation is unacceptable. However, as scientists and philosophers we must nevertheless draw conclusions about human behavior and understand the implications that developed and underdeveloped cognitive abilities can have for our societies. Next, we will sketch a general picture of different cognitive levels and their behavioral implications that we believe can be helpful for research and philosophical inquiry.

As we mentioned at the beginning of Section 6, cognition is the ability to retrieve valuable information from the associative network that helps to make decisions in various contexts. This ability is based on focus and concentration and how habitual they are. We will base our simple classification on these concepts.

On the first, most basic level, Alice can focus and concentrate from time to time without having a habit of doing it. Being a member of a society, this allows her to memorize simple rules of behavior like "stealing from other agents is bad" or "you should wait at a traffic light until it turns green." Once rules like this are memorized, they become part of Alice's associative network. This means that whenever something in the environment reminds Alice about some features of some rule (e.g., she sees a red traffic light), Alice will remember the rule and follow it. Notice that Alice, who is not used to focus, will not question the rule or check what others think about it. She will follow the rule unconditionally. Thus, our theory suggests that on this basic level of cognition we should expect Alice to be *unconditional rule-follower*, who learns some rules of social and personal behavior when she is young and then continues using them without question or modification for the rest of her life. We should also expect that Alice has a small circle of people whom she considers agents. For example, close family, friends, and important people of high social status. Thus, Alice will exhibit empathic behavior only towards these individuals and not towards anyone else (it takes cognitive effort). Finally, affect should play

a large role in Alice's life. She should exhibit a lot of mood-driven decisions, impulsivity, and other phenomena arising from the affective system.

On the next level, we assume that Alice has developed some ability to focus and concentrate habitually. This implies that she pays more attention to her surroundings in the sense that she analyzes situation she is in and pays attention to what other agents do. Alice will have a broader circle of people who are considered agents (e.g., neighbors, people living in the same country) and, as a result, she might modify the rules she learned when she was young to take into account some information from the environment. For example, she might copy the behavior of other agents thus becoming a *conditional rule-follower*. Alice might also empathize with the injustice done to other members of her community and take steps to change something to avoid such things happening in the future. Alice should be more in control of her emotions, exhibit less impulsive behavior, and have more common sense. Undoubtedly, Alice's behavior on this level can vary in a rather wide range depending on the rules that she learns from her environment.

Finally, if Alice has learned to focus a lot and cognition has become habitual for her, we should expect that she might turn into a *rule-questioner*. Alice's circle of agents can become arbitrarily large, potentially including all humanity (sometimes animals and plants as well). Alice will constantly analyze all aspects of the situation from many different perspectives and think about the best course of action independently of what rules are saying. She will become an independent thinker with highly developed empathic abilities who challenges the authority and strives for better life for everyone. We should also expect Alice to be fully in control of her emotions, to be "rational" in her well-calculated decisions, and to never give in to emotional whims.

Without question, this classification of behavior is very vague and incomplete. However, we hope that this broad overview can help researchers to place their research topics somewhere within our theory and to relate their hypotheses to it in some way.

7 Context Model

7.1 Why Reduced Form?

The model of minds presented above is very impressive in terms of describing the details of how the mind might work. However this specificity also defines the scope of applicability of this model. It is good for psychology and neuroscience research where the details of mental processes are studied that are also described in theory of minds. In other words, theory of minds in its associative-network form is good for describing what happens in one specific mind: how associative network is connected in it, what are the values of individual features, etc.

However, this is not good for studying *properties of average behavior,* which is the topic of economics. The network version of the theory is too specific for this and requires too much indi-

vidual data for one person, which is not reflective of the population at all. Nonetheless, average properties of behavior of people in some population (and how this average reacts to something) are important because they allow to conduct economic policy based on some statistical indicators. Even microeconomics, that specifies preferences of individual agents, in the end cares about how groups of these agents behave in aggregate.

Thus, to use theory of minds for such analysis we need some separate, *reduced-form model* that would follow economics tradition and would thus be useful for economists who work in applications, in theory, and in policy. We present such model in the rest of this paper. It is a completely theoretical construction where we nevertheless suggest the types of data that can be used to estimate the parameters of the model. Along the way we make a lot of suggestions about future research directions as many assumptions made in the model were driven by the desire for original simplicity with complications coming later. It is also not hard to imagine how to use the existing tasks from experimental economics and psychology to estimate various parameters in the model. We leave this for the future.

Now we provide a short summary of the model. The new agent we discuss below is called Robbie since he is a reduced-form version of Robin. Robbie is built on completely different principles than associative network of Robin, though there is a sense of some "average equivalence" between the two that we explain below. This is the idea. Robbie is built on principles of economic modeling, however, his design is such that he can capture many characteristics of different minds described above. And there is a sense in which parts of Robbie reflect the inner-workings of the associative network. In other words, this is the reduced-form model that tries to keep important pieces while reducing complexity for greater tractability.

In economics terms, Robbie can be described as follows. His preferences are defined over the set of all contexts, where context is the state of Robbie's mind: everything that he currently perceives, thinks about, or does. Preferences come in three types. First, Robbie can prefer one context to another because it has higher value coded in the features (Tommy; affective value). Second, Robbie can prefer a context because it is more familiar (Robbie experienced it often before). This represents Molly and associations between features. Third, Robbie can prefer one context to another because he has cognitive value of it computed from his models of reality (we do not specify which).

All these preferences get mixed inside Robbie in certain proportions that define his psychology. Robbie can be more cognitive and prefer cognitive value more than the affective values of Tommy and Molly. Or Robbie can be more affective and prefer affective value to cognitive value. Robbie uses the mixture of values to compute utilities of actions in a context and then chooses the action with the highest utility, which moves him to the next context.

We also model how Robbie thinks (to compute utilities of actions). We suggest that Robbie has a knowledge tree where he keeps all information that his cognition has discovered. In any context, Robbie can choose to "press Think button" as we call an instance of reasoning about

the choice that Robbie can perform. Pressing this button makes Robbie gradually smarter. It decreases the costs of thinking in the specific context and overall, and it changes Robbies psychology by increasing the weight on the cognitive utility in the utility function. Thus, thinking both makes future thinking easier and makes Robbie more cognitive in his preferences.

Preferences are also not fixed and change with each experience. Underlying all this structure is a reinforcement learning network where values of states are updated and preferences change correspondingly with this process. We trace how experiences in one context spill over to other related contexts close to the original in some topology. This provides the theory of how contexts are connected to each other. This also paves the way to interpolation of values in new, unexperienced contexts from the data on known contexts and information on how similar all contexts are.

We further discuss how to insert moral considerations into Robbie who can have a mixture of affective and cognitive morality and how to model strategic interactions. We finish with the discussion of how lessons from Robbie's psychological changes can shed some light on the emergence of different types of institutions and how this is related to the individual psychology of agents involved.

Overall, we believe that this model is ultimately useful to economists for one simple reason. It captures a continuum of versions of Robbie whose behavior spans from completely rational (if he collected and analyzed all information in the world) to completely affective (emotional choices, motivated reasoning, psychological biases, Prospect theory, etc.). This essentially covers all imaginable behavioral phenomena (imaginable by us) and allows an economist to be sure that by using this model she is not missing some important pieces that can influence behavior in the environment under study.

The model also suggests which types of *ethnographic* information need to be collected to approximate someone's or some population's average mind. We believe that most of it can be obtained from surveys and simple experimental tasks. This directly leads to new types of economic policy that can take institutions and psychology of agents into account.

7.2 Contexts as Fuzzy Sets

In this section we present the details of the reduced-form model. We start with the finite set of all features $\mathcal{F}$, which supposedly includes everything that the mind, Robbie, can perceive, including all modalities of senses, words, concepts, actions, etc. (in this section we will call the agent Robbie, emphasizing that it is a Robin, but in reduced form). In the discussion of Freddie and more complex minds above, we mentioned that features from reality are perceived by the mind as having different relevances. Thus, we can think of a *real context* in which the mind finds itself (real physical features) as simply the collection of currently present features and their relevances (e.g., how bright the Sun is or how far the bear is). This can be conceptualized as a

fuzzy set C', which is essentially some mapping $\mathcal{F} \mapsto [0,1]$ that assigns a number, or relevance, between 0 and 1 to each feature. We think of features with relevance 0 as not present in C', and a feature with relevance 1 as being maximally relevant (e.g., extreme existential pain or a bear standing right next to you). Similarly, we can imagine some *mind context* C, which is also a fuzzy set only consisting of features that light up in the associative network with different relevances. Contexts like C' and C that are elements of the set of all contexts $\mathcal{C}$, the set of all fuzzy sets on $\mathcal{F}$, will be the building blocks of the model below.[1]

Before we get to the details though, it is important to separate what is real in this model (real contexts) and what is imaginary (mind contexts). Or, in other words, what is contained in reality and what is contained in the mind. This difference and its implications are important, because real contexts are what we, as researchers, can observe and sometimes control and mind contexts determine eventually what Robbie will do. So, to have a good theory that can predict how Robbie operates in reality, we should keep track of the connection between the real contexts and the mind contexts that they evoke.

We will not go into details on this connection—this is the job for the future research—but will simply define it in general to show how it enters the model. Suppose that the mind is in some real context C', which is the collection of all currently physically present features and their relevances. Then we can define a *reality mapping* $\mathcal{C} \times \mathcal{C} \mapsto \mathcal{C}$ that for each real context C' and past mind context $\check{C}$ defines the mind context C that they evoke together. We can think of C as containing two disjoint parts, $C = C'' \cup \tilde{C}$. Here $C'' \subseteq C'$ is the fuzzy set representing the sub-collection of real features that Robbie manages to register (e.g., if Robbie is in the fast-moving train, he might miss some features of the landscape that pass by); and $\tilde{C}$ represents the additional mind features that were there before C' came about or lit up in Robbie's mind by association when it registers C'' as outside features. Notice that whatever Robbie is doing or thinking will be completely determined by C, because this is the information that Robbie has about reality represented by C'.

In what follows, we will not consider real contexts as such, but will focus explicitly on the mind contexts $C \in \mathcal{C}$ and what happens to Robbie when they are perceived. How exactly the reality is represented in the mind is of course crucial for understanding how the mind works, and we will suggest some relationships between C' and C below. But, we will largely leave this topic for the future research assuming for now that Robbie can manage to notice anything that can be important to him.

Finally when considering mind contexts like C, we should take into account that *action features*, contained in the subset of features $\mathcal{A} \subseteq \mathcal{F}$, can be a part of C as well (having positive

[1] In the main text we consider relevances as any positive real numbers, not constrained to $[0,1]$. This is, technically, different from thinking that relevances are constrained by 1. However, this does not create any problems for the model. In any case, relevances cannot go to infinity, because the mind is built from physical matter, that cannot generate anything infinite. Given that there is some limit on relevances, here we just choose it to be equal to 1 to be consistent with the fuzzy-set literature.

relevance). In the discussion of the minds above, we defined behavior as a consequence of the activation of action features. Thus, when considering context C, it might happen that it involves *doing* something. Suppose that there is some threshold of relevance, say $\beta \in (0,1)$, such that if action feature $a \in C$ has relevance higher than β, then the body starts performing action a. So, context C also records what Robbie is currently doing. Notice that action a can be performed with different effort. When the relevance of a is below β, the action is presented in the mind as an idea. When the relevance is slightly above β, Robbie starts to perform a a little bit, as the relevance grows, Robbie starts doing a more and more intensely. For example when Robbie sees a bear far away, he just looks at it, but does nothing. When the bear starts to get closer, Robbie starts first to slowly walk away. As the bear gets even closer, Robbie starts running away, and then running in panic, as the bear approaches. It is also possible that Robbie performs several actions at the same time. If action b also has relevance above the threshold, then Robbie will perform action b in addition to action a. We can say that Robbie's *behavior* in C is determined by the fuzzy subset of action features with relevances above β.

An important implication of the idea that actions are parts of mind contexts is that we do not have to consider actions and outcomes as distinct entities in modeling choice. What we have now are just contexts that already code the action that is performed in them. Thus, the world of the decision-maker is the one where he moves from one context to another while performing different actions in each of them (it can be "doing nothing" as well). We believe that this construct is more powerful than the standard economic abstraction, because now we can formalize the cost of choice. The cost of choice is the cost of *change* in behavior required to move from one context to the next and can incorporate many important psychological phenomena. We will discuss this modeling technique in more detail below.

7.3 Similarity Measure and Topology on $\mathcal{C}$

To make the exposition below more amenable, we need to define some preliminary concepts that will be used in the model below. For example, we will need a sense in which contexts $A, C \in \mathcal{C}$ are similar to each other to update preferences over the contexts, and we will need a related topology on $\mathcal{C}$ that would allow us to talk about contexts "close" to A or C and about continuity of preferences over contexts. Please note that Appendix J contains the list of all variables used in the model.

Suppose that $C = \{(k, p_k)_{k \in \mathcal{F}}\}$ and that $A = \{(k, q_k)_{k \in \mathcal{F}}\}$, where $p_k, q_k > 0$ are the relevances of all features $k \in \mathcal{F}$ that are positive in A or C (they are contained in A or C). Suppose as well that we want to know how much the values of features in A will be updated when features in C are updated (because they are currently perceived, say). This will depend on how "similar" A is to C. For example, if A only contains features that are also in C (with positive relevances), then the update of the value of C should bring the same update to A (since, after all, A contains

the same nodes on the associative network as C). If A only has positive relevances q_k on features that have zero relevances in C (or A and C are disjoint), then A should not be updated at all when C is updated (no common features). Finally, if A contains some features from C, then the update should be proportional to the total relevance weight of features in $A \cap C$ relative to the size of $A \cap C$ within A. So, if A and C have only one, barely relevant, feature in common and A is very large, then A should only be slightly updated. However, if $A \cap C$ contains all but one feature in A, then the value of A should be updated almost as much as C.

To capture all this intuition we use the *similarity measure* proposed by Bush and Mosteller (1951) and defined as

$$S(A, C) = \frac{|A \cap C|}{|A|} = \frac{\sum_k \min\{p_k, q_k\}}{\sum_k q_k}.$$

Here $\sum_k$ goes over all features in $\mathcal{F}$; the notation $|A|$ means the sum of relevances in A; and the notation $A \cap C$ means the fuzzy set with relevances $\min\{p_k, q_k\}$ for all $k \in \mathcal{F}$. Notice that this measure is always between 0 and 1. When A and C have no features with positive relevances in common we have $S(A, C) = 0$. When $A = C$, we have $S(A, C) = 1$. Finally, for the cases in between we will have something between 0 and 1 depending on the total relevance of the intersection $A \cap C$. In the model below, we will make a simplifying assumption that whenever the value of C is updated by some amount x, the value of A is also updated, but by the amount $S(A, C)x$, proportional to similarity between C and A.

To measure how close the contexts are to each other, we use the topology proposed by Gregson (1975) that is generated by the symmetric version of the similarity S above (see also Zwick et al., 1987). Namely, we can define

$$S'(A, C) = \frac{|A \cap C|}{|A \cup C|} = \frac{\sum_k \min\{p_k, q_k\}}{\sum_k \max\{p_k, q_k\}}.$$

Here the fuzzy set $A \cup C$ is the one with relevances $\max\{p_k, q_k\}$. Notice that $S'(A, C) = 1$ only when $A = C$ as with S. Similarly, $S'(A, C) = 0$ only for disjoint sets A and C, same as with S. We will use the topology generated by this similarity measure—that we will call S'-topology on $\mathcal{C}$—to define how Robbie imagines contexts close to the experienced C and possibly chooses actions based on this similarity (see the definition of topological space in Appendix I).

7.4 Choice Problem

Now we can describe the conceptualization of the choice problem that Robbie faces. Suppose Robbie is in some real context C' and his current mind context is C. Notice that C is Robbie's current *state of mind*. It describes the relevances of all currently active features, which can include 1) some features C'' from C'; 2) features associated with C'' on the associative network; 3) possibly some features still active from the past (e.g., Robbie got food poisoned an hour ago and

feels stomach pain while thinking about his grandma who recently passed away). In addition, remember that C also codes the behavior that Robbie is exhibiting if some action features in C has relevance above the threshold β.

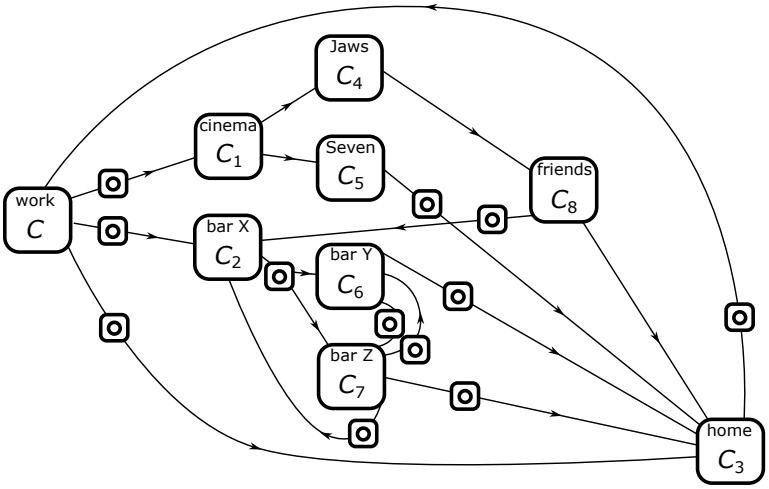

Figure 10: Robbie's view of the world.

We propose that Robbie sees the world around him as a sequence of contexts that are also simple one-shot choice problems, where in each problem the choice among some actions—or to be more precise, among some *changes in behavior*—should be made. Taking an action transports Robbie from one state of mind (context) to another. Figure 10 illustrates. Here, there are two types of contexts that are marked with rectangles. One type corresponds to contexts where Robbie stays in one place, while potentially still doing something (like in bar X he is drinking beer). Another type of context is a smaller rectangle with the circle in it that signifies that this context is mostly about performing some action, for example walking from work to bar X. The process of walking is a context in itself, since it involves some features like street, cars, etc.

Suppose that Robbie (while at work in context C) thinks about what to do in the evening: he can go to the cinema (context C_1), to bar X (context C_2), or simply go home (context C_3). These three options represent three possible mind contexts that are available to Robbie and he can choose one of them. Robbie knows that these are his *choices* because to get from C to one of C_1, C_2, or C_3 he needs to change his behavior. Namely, he should stop working and walk or commute to one of these destinations going through an intermediate context that involves the action of moving from the office to somewhere else.

Other choice problems or uncertainties might follow after one of C_1, C_2, or C_3 got experienced. For example when at work, Robbie is unsure which movie will be shown in the cinema, Jaws or Seven, so he conceives mind contexts C_4 and C_5 that might follow C_1. Then Robbie also knows that if the movie is Jaws, then he will have to stay with his friends and discuss the movie afterwards (C_4 leads to C_8), which does not happen after watching Seven, because it is too scary. Also, Robbie can go to the bars (C_2, C_6, or C_7) and rotate between them for some time.

He can also go to bar X after he talks with friends (or if he does nothing then the meeting with friends will slowly transit into coming home). Thus, Robbie can think about what to do and what happens next in his life by imagining the sequences of contexts and changes in behavior (aka actions) that lead from one context to the next. In principle, Robbie can imagine the whole interconnected network of possible contexts emanating from the current context he is in.

Now let us focus on one context or choice problem. Suppose Robbie is in mind context C and he needs to change his behavior so that this leads him to the next context (or maybe he needs to do nothing and the context will change itself). How does Robbie know what actions are available? We assume that there are several mechanisms that define Robbie's perceived action set. First, the context C itself might include action features that got associated with something in the real context C' (when Robbie sees a bar in C', an idea of having a drink comes to mind and lights up in C). Second, Robbie could have already been in context C and has memories of how he got out of C before (this suggests all changes in behavior chosen in the past). Third, Robbie can think about what action sets he had in some similar contexts in the past (Robbie can look at contexts close to C in S'-topology, check if he chose before in one of them, and use this information in the current context).

The point of this argument is to emphasize that we need to conceptualize available actions differently from how it is usually done in economics. In the standard approach, the modeler decides that the model includes all actions relevant to the decision-makers for the problem at hand and that the decision-makers are automatically aware of this situation and indeed take all the available actions as the full description of the choice problem. The one issue with this view is that it assumes that the modeler's view of the world is the same as the view of the decision-makers. However, it is obvious that the decision-makers, given any real context C', will perceive different things depending on what is stored in their associative networks and might not only see the available actions differently, but also see greater or smaller number of actions than what the modeler believes everyone takes as given. The approach that we suggest instead is to focus on the real context C' and to try to understand what actions might get associated with C' in the minds of the decision-makers under consideration. Thus, action sets emerge endogenously from the information about the real context and the information about the minds of the decision-makers (in many cases, the real context C' will be very suggestive about the available actions, as when people go to vote for example).

To summarize, Robbie sees the world as a sequence of contexts, or states of mind, in the space $\mathcal{C}$ that can be achieved through various changes of behavior available in these contexts. In each state of mind $C \in \mathcal{C}$, Robbie comes up with the set of available actions leading to contexts $\{C_1, ..., C_n\}$ determined by C, his past knowledge, and possibly some cognition (looking at similar contexts). In addition, we assume that there always are two more actions that Robbie can choose in any context. The first action is "Do nothing else" that we also call action ϕ. This action moves Robbie to a special context C_ϕ where he chooses to just remain in his current state of mind

doing whatever it is that C prescribes. The second action is "Think" also called action θ. With this action Robbie moves to the context C_θ "Think about C" and can improve his knowledge (see below), which may lead to a better-informed decision. Thus whenever Robbie is in some context $C \in \mathcal{C}$, he sees the set of actions leading to contexts $\{C_1, ..., C_n, C_\theta, C_\phi\}$ which gets determined from the context, knowledge, and cognition.

7.5 Component Values

The previous section described how Robbie sees the world. Namely, as a sequence of choice problems. However to choose some action, he needs to be guided by some value, association, or reasonable argument (cognition). From the first part of the paper we know already that choice can be produced in many different ways: some context can be chosen because it has high value of features (Tommy); it can be chosen because it is familiar, or has many other features associated with it (Molly); or it can be chosen because there is a model of reality within the cognitive mind that says it is a good choice (Robin). Regardless, the choice process is highly dependent on the structure of the associative network, the values of features in it, the strengths of associations among them, and the availability of cognitive models about the current context. Here in the reduced-form model, we replace all this complexity with three context-dependent *component values* (v_C, f_C, w_C) that roughly correspond to the three types of "utility" that Robbie can feel in some context $C \in \mathcal{C}$.

The first, affective, value $v_C \in \mathbb{R}$ is the fleshly desires of Tommy (inside Robbie), who wants to be in contexts where features have high values. Typically in the mind that is not too burdened with knowledge, contexts that present the possibility to have sugary or fatty foods, socialization, sex, drugs, entertainment, or other things that pleasure the senses will have high value v_C. However, it is important to note that values get updated with time and experience, so v_C can acquire values from other things as well. For example, when a scientist proves an important theorem or an artist finishes a painting, they will become excited, which will then get recorded into the value v_C. So, v_C can represent different types of values, including the fun from doing science or art, depending on the cognition/emotion balance in the mind.

For practical, economics purposes, v_C can be approximated by the basic physiological needs like food, sleep, health, having a home, having a family, friends, interesting job, feeling safe, etc. Notice that things that enter v_C are already known to us to a rather large extent, because they all come from our basic and obviously common biological needs as members of the same social species. Thus realistically, what we need to know about v_C in some specific, given population of interest is which of the basic needs do these people *lack.* This will give a good, preliminary estimate of their preferences with respect to value v_C. Also notice that v_C can be easily aggregated to represent preferences of a population, because everyone is the same species and has roughly the same v_C.

The second, also affective, component $f_C \in \mathbb{R}_+$ is a non-negative number that represents the familiarity value of Molly (within Robbie) with context C. Molly is attracted to things that are highly associated in the network. This is because the high degree of association within and beyond some context C implies that Molly spent a lot of time in C (associations get stronger with experience) and since she spent a lot of time in C it means that C is good (otherwise Molly would not spend so much time in it). All this information is coded within the strengths of associations between features, but we replace this with the value f_C that simply get higher whenever context C gets experienced. Notice that "gets experienced" here means that C gets activated in the mind. And this can happen in at least three different ways: 1) Robbie can experience C himself, in which case f_C increases by some small value; 2) Robbie might imagine C, in which case familiarity f_C also increases by some (smaller) value; 3) Robbie might observe someone else in C and also increase his familiarity with C (since it gets imagined). Notice that Robbie likes C more, the more familiar it is to him regardless of how exactly it got familiar. Robbie might like C just because he imagined it many times, or because he saw other people in C. It might also happen that Robbie does something that is familiar, but has low value v_C. For example, vaccination is something that people are familiar with, but might do it reluctantly because they are afraid of needles. This shows how familiarity value f_C overcomes v_C.

In practice, f_C is even easier to estimate than v_C. All we need to know about some real context of interest C' and some population in it is how many times people experienced this or similar contexts, how often they heard about others being in C' and how often they imagined C'. Information of this type can be collected easily with various survey tools. Notice also that f_C can be aggregated for populations where members have similar lifestyle: they are going to be familiar with the same features in reality and have similar familiar concepts (same culture).

The third component $w_C \in \mathbb{R}$ represents the value that cognition attributes to context C. We think of w_C as being computed with the *models of reality* that Robbie might possess. For example if Robbie is religious, he might attach a high value to the context "church on Sundays" because his model of reality says that being in this context is important for salvation and for good life after death. If Robbie is not religious he might attach low value to "church on Sundays," because he might think that it is a waste of time that he can better spend saving the environment. Such Robbie might also attach high value w_C to the context "world without fossil fuels."

An important difference between v_C, f_C on the one hand and w_C on the other is that the former get computed automatically and without cost from the associative network that is already in place, while w_C needs to be computed through costly reasoning (Robin). We assume that Robbie can compute w_C if he presses the Think button (aka chooses to think, chooses the action θ). If this happens, Robbie obtains some w_C that comes from his models of reality, and this value gets recorded into Robbie's *knowledge* that we discuss below.

Notice that we do not specify *which* models of reality Robbie is using. This is because such models are different across different people and cultures. In some cultures, people might

use physics to reason about the cognitive values of contexts. In others, people might rely on witchcraft, religion, or traditions to figure out what the value is. Thus, our approach here also suggests that we need to estimate the values w_C from the models of reality of the decision-makers and not from the model of reality of the modeler.

In practice, it does not seem like a very complex task to determine how some specific population of people reasons about values in some specific contexts. This can be approximated from the understanding of local culture, traditions, science, or from the specific topics covered in school curriculum. Notice yet again that this knowledge will be the same across members of a given population assuming that they all get the same education and focus on roughly the same information.

7.6 Knowledge

When Robbie presses Think button he employs his models of reality to acquire new *knowledge* about the context he is in or about some possible future contexts or beliefs. In this section we discuss how this information gets stored and accessed in Robbie's episodic memory.

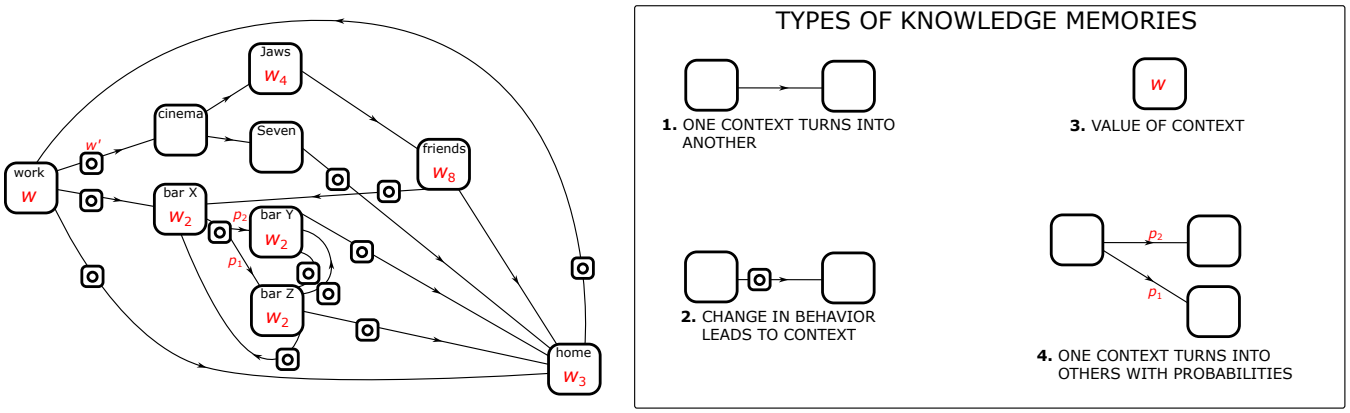

Figure 11: Robbie's view of the world.

There are several types of knowledge that Robbie can memorize about the world. The right panel of Figure 11 illustrates. The first kind of memory is about "transient" contexts that do not stay as they are but change. For example, if you leave a bottle of milk on the table, the milk will get sour. So, one context with fresh milk gradually turns into another context with sour milk. Robbie can remember this important information in his episodic memory (type 1 on the figure). Notice as well that there might be several memories of this type related to one context. For example, Robbie did not know if they are going to show Jaws or Seven in the cinema. This is represented by two memories of type 1, one for Jaws and one for Seven. Both memories activate when Robbie imagines context "cinema" thus creating uncertainty about the future.

Even though many contexts are transient, there are enough that are not, and in such "stable" contexts that do not change Robbie needs to make choices. The memory of type 2 records the

change in behavior that leads to another context. Again there can be many of these memories related to the same action in some context (for example on the left panel of Figure 11, we can see two possibilities happening after action in bar X).

The next, type 3, is the memory of the value w of some context. We assume that once these values get computed with the Think button, they are stored permanently in the episodic memory, unless Robbie decides to recompute them.

The memory of type 4 concerns with probabilistic beliefs. The beliefs are also computed from the models of reality. For example in the left panel of Figure 11, Robbie might think that if he leaves Bar X, he will end up in bar Y or bar Z depending on some factors (e.g., whether a friend is sitting in bar Y or not). Robbie estimates that his friend will be sitting in bar Y with probability p_2, and if the friend is not there, then Robbie knows he will go to bar Z (probability $p_1 = 1 - p_2$). The same thing can happen without Robbie taking an action. In this case, Robbie has probabilistic beliefs about a transient context and what it might become in the future. Or it can be a mixed context that transits into another on its own, but an action can also be taken (e.g., context "friends" on the left panel of Figure 11).

Now, we can define Robbie's knowledge as the collection of all memories of the four types described above. Knowledge gets gradually collected through experience and new pieces are added to the memory collection as time unfolds. We will call the collection of memories forming Robbie's knowledge his *knowledge tree* (possibly emanating from an action or a context). If we look at the left panel of Figure 11, we can see that Robbie has a lot of memories of types 1 and 2 (all the arrows on the picture). He has type-3 memories of values $w, w', w_2, w_4, w_8, w_3$ and a belief memory about the probabilities of contexts after action in bar X. All this information gets retrieved by association from Robbie's knowledge tree when he is in some context or when he imagines some context or its consequences.

Notice that not all contexts have values and not all uncertainties carry probabilities. When Robbie discovers or imagines a new context C, he does not have to think about its value w_C, so some contexts might just stay without any value like "cinema" or "Seven" in the figure. Robbie might also not attach probabilities to contexts Jaws and Seven when in cinema. Thus, Robbie's knowledge can be patchy, does not have to be consistent across different elements, and in fact can come from different models of reality. Typically, Robbie is not going to use one model for all contexts. To illustrate, suppose Robbie is a plumber. Then, he will have a lot of knowledge stored in his memory that relates to plumbing and plumbing theories. His knowledge in plumbing contexts will be very good and fine-tuned, with probabilities and values well-defined for most contexts. However, when it gets to vaccination, Robbie might not have much knowledge (plumbing theories do not apply) and might use a simple model of reality he heard online that says that vaccinations are bad for you because they contain microchips. Robbie uses this model to decide to not vaccinate. It can also be that Robbie is a microbiologist. Then, he will use the same model of reality for work contexts and for vaccination contexts, and might choose to vac-

cinate. However, when it gets to religious contexts, it might happen that Robbie-microbiologist will use Christianity as his model of the world.

7.7 Intuition

The fact that most human beings use different, and often inconsistent, models of reality in different contexts suggests that they can *afford* it. This means that they can manage to live their lives somehow without having one grand theory of everything that is computed for all imaginable contexts. And this is also the reason why neoclassical economics models (rationality) do not fit human behavior very well. In a typical economics model, rational agent is assumed to have the "correct" and consistent picture of the whole reality—maybe with some uncertainties—that is represented by the model of reality in the head of the modeler (the modeler assumes that his model of reality is how the world actually works). It is assumed that economic agent has computed all values w_C for all contexts $C \in \mathcal{C}$ and all probabilities of all uncertain transitions between all conceivable contexts. It is also often assumed that the beliefs that the agent has are also "correct" and represent the actual probabilities of events (rational expectations). Thus, a rational agent in our framework can be seen as a Robbie who knows how the world actually works and who has computed all values and all beliefs for all contexts.

The assumption of rationality, which is obviously not consistent with most human behavior, was made in economics because economists did not consider any other systems, except for cognition, that can do the decision-making. However according to the theory of minds in the previous sections, such systems exist and can guide behavior even in the complete absence of any cognition (Tommy, Molly). And this is the reason why people can manage to live their lives with patchy and inconsistent theories of reality. It is because their behavior can be directed by other, non-cognitive systems and people rely on them in situations when their models of reality are not good enough or do not help all together. In other words, people rely on *intuition*.

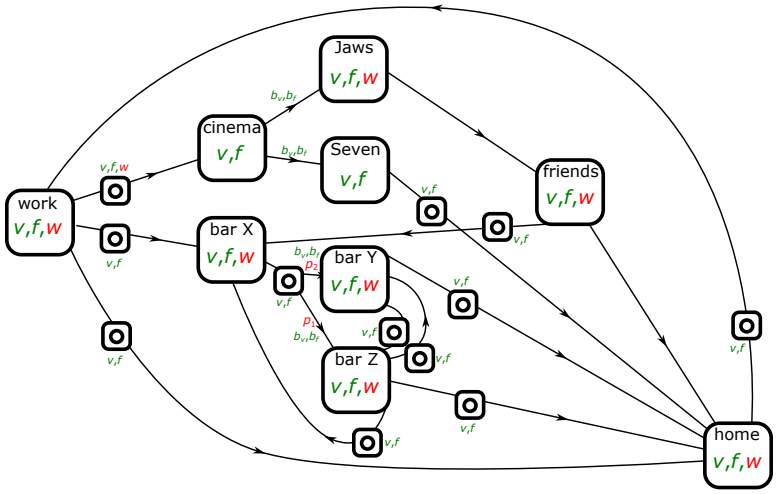

Figure 12: All information available to Robbie.

To understand what intuition means in our framework, we should consider all information available to Robbie when he is in context C (work). Figure 12 illustrates. Here in red we see the same knowledge acquired by Robbie in the past as discussed in the previous section (subindexes are dropped for convenience). In green, we see the other component values v and f that are instantly available to Robbie simply because his associative network computes them automatically. Specifically, such values are available in all contexts that Robbie imagines. The same holds for the beliefs—marked on the figure by b_v and b_f—that are also computed automatically from the component values v and f (see below).

Thus with this additional information, Robbie can make choices even in contexts where his models of reality are completely useless. For example from his component value v, Robbie can construct a belief that it will be Jaws in the cinema *for sure* because he likes Jaws much more than Seven (Tommy; mood affiliation); or Robbie can believe that it will be Seven in the cinema with probability 80% because Seven is overall shown more often everywhere than Jaws (Molly; frequency, familiarity). Notice as well that this intuitive information coming from Tommy and Molly can amalgamate with knowledge previously obtained through cognition (in red on the figure). When Robbie is in Bar X, he will compute beliefs about ending up in bars Y and Z using a mixture of cognitive information (p_1, p_2) and intuitive information (b_v, b_f). The same is true for the values of contexts that have component value w computed for them. In bar Z for example, Robbie will mix component values v, f, and w to obtain a *utility* that we discuss in the next section.

We can define *intuition* as the estimates of values of contexts, actions, and beliefs that are obtained from the affective information (v, f) and the previously stored cognitive information (w) about contexts and their connections. These are quick estimates that Robbie can make by just thinking about what to do in a context. These values are computed automatically and come to Robbie as intuitive judgements when he thinks about some action in some context.

An important point about this is that Robbie *does not need* to use cognition (press Think button) to obtain his intuitive estimates of values. This is done by the machinery of the mind. However, Robbie can still choose to press Think button and reason before he makes his choice. This action can produce new cognitive information in the form of values of contexts, beliefs, or connections between transient contexts that can improve Robbie's decision-making. In the following sections we describe how this process happens in detail.

7.8 Imagined Utility

We can think of different minds within Robbie, such as Tommy, Molly, or Robin, as acting separately and producing their own values v, f, and w. However, they are all parts of Robbie and are all represented in the same associative network. Thus, it makes sense to believe that Robbie will somehow aggregate the information coming from the three systems to make his choices. Notice

that to make a choice, Robbie needs to *imagine* different contexts where he can find himself in the future. Thus, the aggregation of values into utility happens in the imagination, and therefore we will call it *imagined utility*.

Suppose Robbie is currently in context C. We assume that, when he imagines some context D (while in C), he perceives *imagined utility* of context D as

$$u(D|C) = v_D + (f_D - \chi) + (\alpha_C + \alpha)w_D.$$

Here, for any context D, we simply add the three values, though with some additional parameters. Notice also that when w_D is not available, or it is not in the knowledge tree, we assume it to be zero or absent (see the argument in the next paragraph).

The number $\chi > 0$ is subtracted from f_D to represent the idea that unfamiliar things feel bad (negative value). Remember that in our framework the sign of value has meaning. Positive sign means pleasant value and negative sign means unpleasant value. Since f_D is a positive number (frequency), we need to control what Robbie feels in completely unfamiliar contexts (where f_D is very close to zero). The higher χ is, the more unpleasant Robbie will feel in unfamiliar contexts. Thus, Robbies with high χ will avoid unfamiliar contexts more than Robbies with low χ. We currently do not have a theory for how χ is determined, so we leave it for the future research and simply assume for now that χ is a fixed individual parameter.[2]

In addition to this, we assume that the cognitive value w_D has weight $\alpha_C + \alpha$, with $\alpha_C \geq 0$ and $\alpha \geq 0$, that represents the importance of cognitive utility to Robbie *in context* C that he is currently experiencing. This weight can be thought of as Robbie's cognition/emotion balance or how "cognitive" he is in C. To make it clearer, consider

$$u(C|C) = v_C + (f_C - \chi) + (\alpha_C + \alpha)w_C.$$

As we will explain in more detail below, $u(C|C)$ is what Robbie actually feels when in C, or his current mood. Thus, the coefficient $\alpha_C + \alpha$ determines how much of what Robbie feels comes from cognitive value w_C relative to affective values v_C and f_C. The higher $\alpha_C + \alpha$, the more of cognitive value Robbie will feel.

In addition to this as we explain in more detail in Sections 7.13 and 7.14 below, α_C and α increase when Robbie thinks in context C. As Robbie keeps pressing Think button, he becomes gradually smarter, his $\alpha_C + \alpha$ grows. As a result, Robbie becomes more "serious" in context C, he starts feeling more cognitive value in it. When imagining context D, he also starts relying more on cognitive value w_D than on familiarity f_D or the affective value v_D. So, the increase in

[2]One idea is to assume that $-\chi$ is the (negative) mood that Robbie felt when he was in the most unfamiliar context he ever experienced in his life. Then χ can change and be replaced with the new mood if Robbie finds himself in even more unfamiliar context.

$\alpha_C + \alpha$ allows Robbie to make choices that are driven by cognitive value to a higher degree as he keeps thinking in C.

The idea that the weight $\alpha_C + \alpha$ increases when Robbie thinks is in fact a reflection of the changes that happen in Robbie's mind during cognition (that also have other consequences, see Section 7.14). Suppose Robbie thinks a lot about his work. This will make the parts of his associative network that represent contexts related to work more interconnected with high-capacity associations. This happens because Robbie uses Focus and Concentration to perform cognition and these operations necessarily involve strong associative activation of related features in the network. Thus, when Robbie comes to work (context C) he will feel more and stronger associations that come from cognition than from affective value or familiarity systems (v and f). This will make him react more to cognitive values w_D than to others when he imagines different contexts D. Thus in context C with high weight $\alpha_C + \alpha$, Robbie will be driven more by cognitive values w_D of imagined, possible future contexts D. As a result, Robbie will be very serious and focused on his job. We represent this by the coefficient α_C that is high due to the fact that Robbie thinks a lot about work. In some other context, like for example bar X, Robbie might have low α_{barX}, because he never uses cognition in a bar and the only associations he gets there are related to beer, football, and politics. This will make Robbie more reactive to affective values v and f in this context and allow him to do things he would typically never do at work.

The other term α also increases when Robbie thinks. However, it makes cognitive utility feel more important in all contexts. We assume this because Robbie uses the devices, Focus and Concentration, for all thinking and thus they get trained like muscles from each use. This suggests that if Robbie thinks a lot in one collection of contexts (say, related to plumbing), then he will become a little bit more cognitive in all other contexts just because it will be easier for him to activate Focus and Concentration if they are well-trained.

7.9 Perspective-Taking

It is important to mention that the way imagined utility $u(D|C)$ is formulated leads to a *perspective-dependent* calculation of utility. Indeed, suppose that $\alpha_C \neq \alpha_E$ for some contexts C and E. Then while in C, Robbie will imagine the utility of D to be $u(D|C)$, whereas when he moves to E his perspective will change and he will perceive utility of D as $u(D|E)$. For example, when Robbie is at work (context C) he has high α_C and he is very cognitive. Robbie imagines going to the gym (context D) and thinks that he will have a great time there because it is healthy and brings a lot of cognitive utility w_D. Since Robbie weights cognitive utility w_D with high α_C it feels to him that the gym will be very pleasant. Thus, he decides to go to the gym after work. However when Robbie gets home before the planned trip to the gym, he relaxes. His α_E at home is much lower than α_C at work, so as a result $u(D|E) < u(D|C)$ because the high cognitive value w_D of the gym is now not weighted as high as when Robbie was at work ($u(D|C) - u(D|E) = (\alpha_C - \alpha_E)w_D$).

Robbie thinks: "It felt like I really want to go to the gym when I was at work, but now that I am at home it feels like it is too painful. I will skip the gym."

Imagined utility in the formulation above can lead to *time-inconsistent behavior* (meaning that someone has decided on a plan of action, but then changes it on the way) because the perception of imagined contexts depends on the current context. Robbie with such imagined utility is naive and does not understand that his perspective on D will change when he moves from context C to context E. However, it does not have to be that way. We know that people can be self-aware (see Section 6.4) and are able to imagine that they will have different perspectives in other contexts (e.g., Robbie can imagine that he will feel lazy at home and choose to go to the gym straight from work). So, we can assume that Robbie can imagine other perspectives and thus other imagined utilities. For example while thinking about the gym at work, Robbie might be able to imagine $u(D|E)$, or the imagined utility of the gym at home, and deduce that if he gets home, he will not go to the gym. Or even better. Robbie might be able to imagine $u(D|D)$ with the actual and correct α_D. If Robbie can do that, then he can be time-consistent in all contexts, because he will imagine the same gym utility $u(D|D)$ when at home and when at work.

It is also possible that Robbie can imagine some of his own perspectives but not all of them. In this case in some contexts C, Robbie will only use $u(D|C)$ when imagining D. He is *naive* in C. In some other contexts, Robbie might be able to imagine other perspectives $u(D|E)$ or $u(D|D)$. We believe that imagining other perspectives takes cognition and that Robbie needs to train this ability. We leave it for the future research to understand the perspective-taking better and will assume in the rest of the paper that Robbie is naive and can only compute $u(D|C)$ in every C.

7.10 Beliefs

Now that we have aggregated values into utility, we can describe how uncertainties are aggregated into expectations. This process however is not as straightforward as in standard economics where beliefs are simply assumed to be equal to something. In our framework, Tommy, Molly, and Robin have separate "belief-generating systems" that form beliefs independently from the information available to each system (v, f, and models of reality). Thus, we need to *aggregate beliefs* across minds before we get to aggregation of uncertainty.[3]

We start with describing the ways beliefs are formed by each mind. From previous work (Kimbrough and Vostroknutov, 2022), we assume that Tommy uses *mood affiliation* to form beliefs about uncertainty. Suppose that Tommy-Robbie is currently in C and he contemplates some uncertainty following an action in some context D on the knowledge tree that leads to contexts

[3]It should be noted that this is the place where the economics idea of "as if" models actually finds its turf. When we talk about beliefs of Tommy and Molly, we really only *imagine* that they have beliefs (as cognitive agents do), whereas in reality Tommy and Molly do not know what it means to have beliefs. They just do things when they feel they should be done. However, it is worth to think of Tommy and Molly as having beliefs because it allows us to mix beliefs across different systems in one framework.

$D_1, ..., D_n$ (e.g., he is at work and thinks about uncertainty when leaving bar X). Suppose that Robbie is in mood $M \in \mathbb{R}$ (we show how to find it below). Then, his Tommy-belief will be that the context D_k, with the utility closest to his mood M, will happen with probability 1. Let us say that the probability of D_k according to Tommy is

$$\begin{aligned} b_t(D_k|C) &= 1 \quad \text{if } k = \arg\min_{i=1..n} |u(D_i|C) - M| \\ b_t(D_k|C) &= 0 \quad \text{else.} \end{aligned}$$

With Molly the situation with beliefs is even simpler. Molly's preferences f_D are based on familiarity. The values f_D literally count how many times D was experienced (imagined, observed). Thus, these values are frequencies of past occurrences of contexts and essentially *are* beliefs. For contexts $D_1, ..., D_n$ in the example above, Molly will take all values f_{D_k}, the familiarities of these contexts, and form an empirical distribution with probability of context D_k being

$$b_m(D_k|C) = \frac{f_{D_k}}{\sum_{i=1..n} f_{D_i}}.$$

For Robin we do not have a formula to represent his beliefs because they come from some models of reality that we do not specify. Thus, we assume that Robin produces probabilities $p_1, ..., p_n$ for contexts $D_1, .., D_n$ if he thinks about it. If Robin does not think about these probabilities, then they will be missing and the aggregation will go on without them. We can summarize this as follows:

$$\begin{aligned} b_r(D_k|C) &= p_k \quad \text{if computed} \\ b_r(D_k|C) &= \emptyset \quad \text{if not computed.} \end{aligned}$$

Now we need to aggregate these beliefs into one. We suggest that—since Robbie has a type based on the weights $\alpha_C + \alpha$ that determine how cognitive he is in the current context C—the beliefs should be aggregated across Tommy, Molly, and Robin in the same proportions as values in $u(\cdot|C)$. Let $h_C = 1/(2 + \alpha_C + \alpha)$. This gives the following formula for aggregated belief $b^*(D_k|C)$ in case cognitive probabilities $(p_1, ..., p_n)$ are computed, and aggregated belief $b(D_k|C)$ in case they are not:

$$\begin{aligned} b^*(D_k|C) &= h_C \left[b_t(D_k|C) + b_m(D_k|C) + (\alpha_C + \alpha) b_r(D_k|C)\right] \quad \text{if } (p_1, ..., p_n) \text{ is computed} \\ b(D_k|C) &= \frac{1}{2}\left[b_t(D_k|C) + b_m(D_k|C)\right] \quad \text{if } (p_1, ..., p_n) \text{ is not computed.} \end{aligned}$$

This construction implies that highly-cognitive Robbie with high $\alpha_C + \alpha$ will form beliefs based mostly on knowledge coming from the cognitive system. Such Robbie will ignore impul-

sive motives and familiarity biases. When Robbie is not super cognitive (low $\alpha_C + \alpha$), his beliefs will be mostly influenced by mood and familiarity. Such Robbie, in a good mood, will believe that in the future everything will be great (mood affiliation) and that unfamiliar things never happen. Robbies with intermediate values of $\alpha_C + \alpha$ will have mixed beliefs influenced by all factors. The same differences apply to a single Robbie who is differentially cognitive in different contexts.

7.11 Expected Imagined Utility

Now that we have defined imagined utility and beliefs over contexts on the knowledge tree, we can describe how *expected imagined utility after context D* is computed (while Robbie is in some current context C). This is the expected utility of the next context that can be uncertain and be among some contexts $D_1, ..., D_n$ that represent uncertainty over one period into the future (only). We assume that Robbie takes one expectation in each context using the utilities of the directly connected contexts and ignores the possibilities of computing deeper and taking more future contexts into account. This might not be such a bad strategy for two reasons. First, it is probably too cognitively demanding to compute the whole discounted utility from infinitely many periods in the future using strict laws of Bayesian updating. Second, given that we are dealing with essentially a Q-learning network, the values propagate from context to context. Thus, the utilities of the contexts take into account future utility to some degree.[4]

To determine the expected future utility in D we first need to talk about the *cost of changing contexts* from D to some D_k. As we mentioned above, each context D can have a behavioral expression when some actions, corresponding to highly active action features, are performed. Thus, switching from one context to another might involve changing behavior. This might be related to some cost. Imagine that in context D Robbie is exercising in a gym, and in the next context D_k that comes in an hour he needs to be at a wedding showered and dressed up. The transition from the gym to the wedding does involve significant costs. We can denote them $\kappa(D_k|D) \geq 0$. Even without changing behavior, there might be costs associated with context transitions. For example in context D, Robbie might be crying in his bedroom while depressed and in an hour he needs to talk to his mother on zoom and he needs to look happy. The transition from one emotional state to another might involve a significant mental cost. Thus, we will assume that for all context transitions there are associated costs that Robbie takes into account.[5]

Now, to compute the *expected imagined utility after D* we use Robbie's beliefs. There are two possible cases. In the first case, Robbie does not have cognitive, probabilistic beliefs computed

[4]In general, we make many assumptions about how the computation of expected utility takes place. Most of them are made for convenience and simplicity, given that this is the first model of this kind. We believe that future research should clarify how exactly these computations are done and propose a better version of the model.

[5]There are of course many situations where $\kappa(D_k|D) = 0$. For example when you take something from the shelf in a supermarket, we can think that the cost of transition from standing in front of the shelf to picking a product from the shelf is zero (unless you are handicapped).

for the connections from D. Here we compute the expected utility of D while being in C as

$$E(D|C) = \sum_{i=1..n} b(D_i|C)[u(D_i|C) - \kappa(D_i|D)] - \zeta.$$

Notice that here the beliefs b are computed without taking cognitive probabilities into account and that the utility of context D_i is equal to its utility minus the cost of transition from D. Following Kimbrough and Vostroknutov (2022), we also assume that there is some cost of uncertainty $\zeta > 0$ perceived by Robbie when he does not have cognitive probabilities attached to connections. The cost ζ reflects the idea that without cognition Robbie is not very sure about the resolution of uncertainty since he never thought about it. In a sense, it is reminiscent of *ambiguity aversion*.[6] However, if Robbie did think about the uncertainty and attached probabilities $p_1, ..., p_n$ to the contexts $D_1, ..., D_n$, then we assume that the cost ζ disappears and the expected utility is computed as

$$E(D|C) = \sum_{i=1..n} b^*(D_i|C)[u(D_i|C) - \kappa(D_i|D)].$$

Here, beliefs b^* do take cognitive probabilities into account.

7.11.1 Transient Contexts

These formulas for the expected imagined utility $E(D|C)$ express the straightforward intuition. But, it needs to be clarified what this implies for transient contexts where context C can change on itself into some other context. When this can happen (e.g., milk can turn sour), Robbie needs to employ a specific action "Do nothing else" to take into account possible self-transitions.

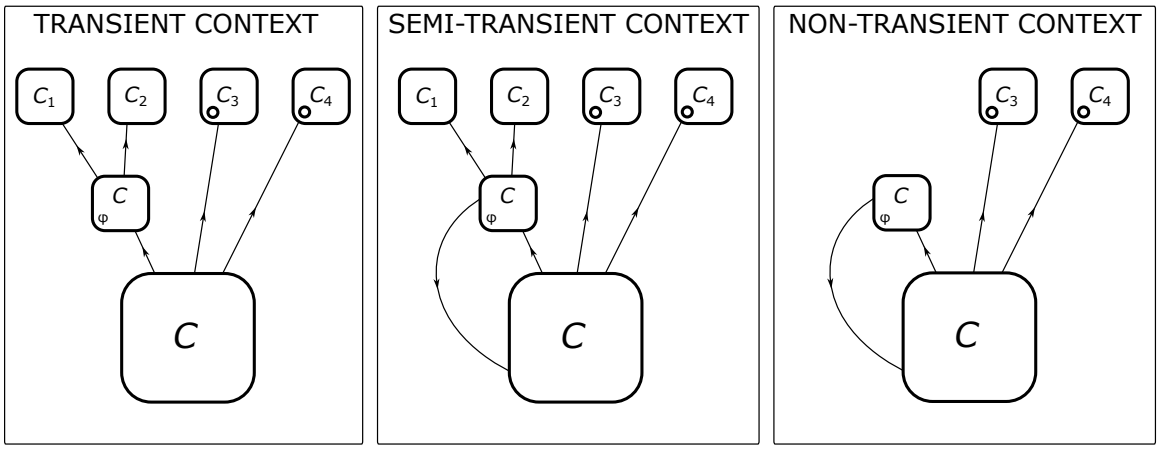

Figure 13: Three types of contexts.

Suppose that Robbie is in context C and that there are connections to contexts $C_1, ..., C_4$. Figure 13 illustrates. Contexts C_1 and C_2 on the figure are contexts into which Robbie can transit

[6]The cost of uncertainty ζ should probably depend on C, D, $(D_i)_{i=1..n}$, and on beliefs b_t and b_m. We do not have a good theory of ζ, so we fix it as a constant for now and leave it for the future research to figure this out better.

without doing anything (e.g., it can start raining). Contexts C_3 and C_4 are reachable only if Robbie performs some additional action (changes his behavior), which is marked by a circle on the figure. In *transient contexts* (the left panel), Robbie cannot stop context C from changing to context C_1 or C_2, however he can still act (C_3 or C_4) and possibly change the situation. If Robbie cannot do anything in a transient context (C_3 and C_4 are unavailable), then he will "be transited" to some next context without having a choice. In *semi-transient contexts* (the middle panel), the context might not evolve into C_1 or C_2 for sure, but can stay what it is as well, which is marked by the arrow circling back into C. Finally in *non-transient contexts* (the right panel), there are no self-transitions to another contexts. Here Robbie is trapped unless he acts and chooses C_3 or C_4.

The reason we consider these types of contexts is because in transient contexts Robbie needs to decide whether he wants to do something about it (act in some way, like C_3), or just let it go, do nothing, and wait until the context transits on its own. We suggest that Robbie deals with this by means of a special context "Do Nothing Else but C" that we denote C_ϕ. This context is just C plus an additional action feature "Do nothing else" (action ϕ) that is activated on top. If ϕ activates, then Robbie moves to another context C_ϕ where he acts as he is supposed to in C, but pays special attention to *not* perform any other actions except for those he is doing in C. In other words, he waits for the context self-transition.

The introduction of context C_ϕ allows Robbie to keep track of uncertainty that follows after self-transitions of C. The values in this context will reflect the expectations of uncertainty in the future and Robbie can use them to decide whether to do nothing or not. Thus, when we consider the expected imagined utility in some C which is connected to $C_1, ..., C_n$, we should keep in mind that with the introduction of C_ϕ only two cases are possible. There are contexts C where *all* connections involve an action, or change in behavior, and one of these actions is ϕ (like in the example with transient context in the left panel of the figure). And there are contexts like C_ϕ on the figure, where *all* connections are non-action connections that do not involve action or any change in behavior.[7] This exhausts all possible types of contexts, though the expected utility is computed in them in the same way.

7.12 Expected Utility of an Action

Now that we computed the expected utility of a context, we can compute the expected utility of an action. For clarification, let us discuss again what *is* an action. Remember, in our framework each context C already codes for the behavior that is performed in it and transitions between contexts signify possible changes in behavior. Thus, we can define *action* as a change in behavior *necessary* to transit from context C to some context C_k. The word "necessary" here is important because the idea is that without this action C_k will not come about.

[7] This is actually reminiscent of the artificial Nature player introduced in game theory.

As just discussed above, choice happens only in contexts C where *all* connections to other contexts involve actions including one called Do Nothing Else. And specifically, there are no forced transitions to other nodes, because they are all computed in C_ϕ. In such context C, Robbie can estimate the expected utility of an action (change in behavior) that leads to some context C_k and also takes into account possible future contexts after C_k.

We suggest that Robbie does this in steps by computing the expected imagined utilities of various contexts as he travels down the knowledge tree from the original starting point C. As Robbie thinks about C_k, first that comes to his mind is the cost of context transition $\kappa(C_k|C)$ as well as $u(C_k|C)$, the utility of the context to which Robbie can transit. So, we can say that at level 0 away from C Robbie computes (expected) utility

$$U_0(C_k|C) = u(C_k|C) - \kappa(C_k|C).$$

Now, Robbie can go to the next level of contexts, those that connect to C_k. But to do that he needs to use more imagination than usual: he needs to imagine all contexts that follow an *imaginary* context C_k, which can be harder than when thinking about how current observable context C will change. Thus, we assume—which is the same thing that we do in the first part of the paper—that the utilities from the levels further than 1 are discounted with some factor $\delta \in (0, 1)$. This δ is related to the δ in the first part, but not exactly. This is because current δ also involves associations, but on a sort of "macro" level (contexts). We will leave this discussion for the future and just assume some δ.

For any context C_k, let us denote by $P(C_k)$ the set of contexts to which transitions can be made by action or fate (forced transitions). Let also $P^2(C_k)$ denote the set of all contexts to which transitions can be made from some context in $P(C_k)$. Similarly we can define $P^3(C_k)$, etc. Then, on level 1 Robbie computes the expected imagined utility of C_k discounted by δ:

$$U_1(C_k|C) = \delta E(C_k|C).$$

Further down the knowledge tree Robbie computes level 2 utilities as

$$U_2(C_k|C) = \delta^2 \sum_{D \in P(C_k)} E(D|C).$$

Then for all further levels t:

$$U_t(C_k|C) = \delta^t \sum_{D \in P^{t-1}(C_k)} E(D|C).$$

Overall, we can say that Robbie estimates the *expected (imagined) utility from action leading to* C_k as

$$U^{\infty}(C_k|C) = \sum_{t \geq 0} U_t(C_k|C).$$

Here, we compute an infinite sum, which is reflected in the subindex in U^{∞}. Realistically though, it is reasonable to assume that Robbie cannot imagine infinitely many levels of future contexts. In Section 7.18 below, we suggest that Robbie might have limits on imagination represented by the imaginativeness parameter ι. This parameter puts the upper limit on the total activation of the associative network. We can then use this idea and define

$$U^{\iota}(C_k|C) = \sum_{t \leq n(\iota)} U_t(C_k|C).$$

In this definition, we assume that Robbie can only imagine limited number of future contexts up to level $n(\iota)$ defined as the highest level that Robbie can imagine given his limit ι. This number, $n(\iota)$, can be defined as the highest n for which

$$|C| + \delta|C_k| + \sum_{2 \leq t \leq n} \delta^t \sum_{D \in P^{t-1}(C_k)} |D| < \iota.$$

Thus, we can connect our framework to the models of *level-k reasoning* that also assume a limit on the number of levels of reasoning (imagination) that the agent has.[8]

7.13 Think Button

Another action that Robbie has in any context C is Think button (aka action θ). As we mentioned above, it helps Robbie to learn new information from the models of reality. It is treated by Robbie as a regular action (context) with utility that gets updated depending on the results of the thinking process (see below). So, the choice to think for Robbie enters the maximization problem when he chooses among available actions. Robbie presses Think button when its expected utility is the highest among all other options, and if there are options with better expected utility then Robbie does not press Think button and consequently does not learn new information.

The left panel of Figure 14 shows the "Think about what happens after C" context represented with the θ at the bottom-left. It is created by activating Focus and Concentration while being in C, so it is a "mental" operation similar to Do Nothing action. When Think button is pressed, Robbie moves to context C_θ where he thinks about new possibilities after C. Once one instance of reasoning (or maybe a bunch of instances) are done, Robbie returns back to context C and thinking stops.

[8] We implicitly assume in the above formula that $|C| + \delta|C_k| < \iota$. In other words, Robbie is capable of imagining the current context C and the context C_k that can follow.

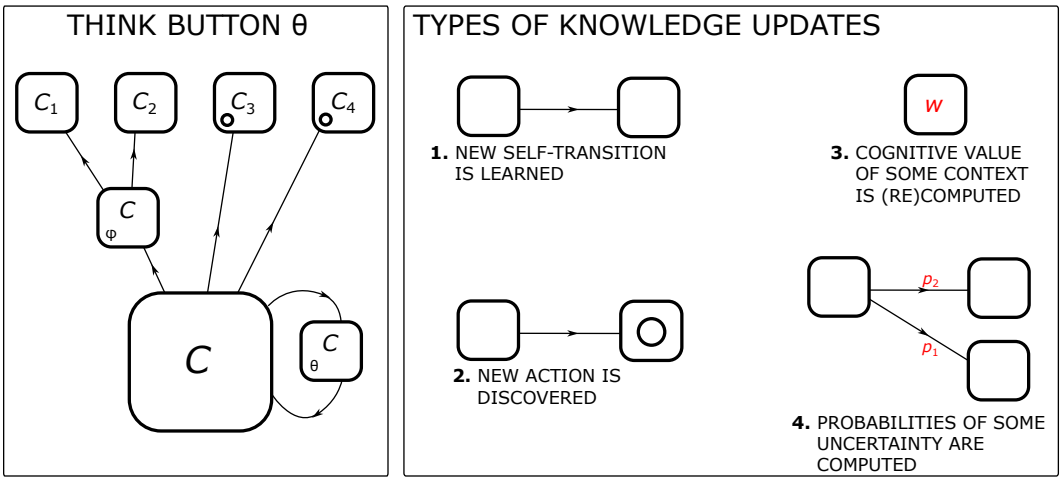

Figure 14: Think button and types of knowledge updates.

While Robbie is in C_θ he will focus on different future contexts on the knowledge tree emanating from C and will try to find out new things about these contexts using his model of reality (that he typically uses in C). We assume that one press of Think button produces one *knowledge update* that can come in four types. The right panel of Figure 14 illustrates.

First, Robbie can learn that another self-transition from a context is possible. For example, when Robbie was young he discovered that people can die (e.g., disappear somewhere and never return). He learned that there exists the ultimate self-transition of contexts that cannot be avoided. Or, Robbie can learn that his car can break in yet another new way, and he adds this context to the collection of possible car malfunctions.

Second, Robbie can realize that some new action is available in context C. By action we mean some new activity added on top of existing activities in C that can somehow transit Robbie to another context. When such action is discovered, it is recorded on the knowledge tree as a link between C and the new context where the action is performed (later presses of Think button might also produce the consequences of the action).

Third, Robbie can compute new cognitive value w_C for context C (if an old one existed, it will be overwritten). This happens for example when Robbie receives new information. Suppose that he came to C and found out that some gauge on a machine shows number 7 whereas it used to be 5. In this case, Robbie might update w_C—after recomputing it through the model of reality—because now there is new information that can be taken into account.

Fourth, Robbie can compute probabilities of some uncertainty following some context. This can be a context like C_ϕ where all connections are self-transitory and this is a pure matter of understanding how the world works. For example, Robbie might estimate the probability that it will rain today. Or it can be a context where Robbie needs to act. In this case, he can also have beliefs about what action will be taken. These beliefs might be of the kind "According to reason I must take this action in this context" or "I know that I cannot hold myself, so I will choose this action in this context," etc. In general, cognitive beliefs about own action sets might come from

some model of self that Robbie might have or from some norms that Robbie wants to follow. We leave this very interesting question for future research.

From our current perspective, we cannot say much about *which exact* knowledge update Robbie will come up with in a given context when he presses Think button. Given how reasoning works in Robin (some random associations that get interconnected sometime), it is possible that no one, including Robbie, knows what he will come up with when he thinks. It might be a process highly dependent on minuscule active features or random associations (e.g., the virus idea in Independence Day movie) that is generally chaotic.

Given this, we will not provide any suggestions on how thinking exactly happens, but will simply assume that a knowledge update happens in some random place on the knowledge tree emanating from the current context C. It is highly likely that the updates are more likely in contexts closest to C on the tree provided that imagination is not easy and tracing many contexts into the future is a difficult task. It is also possible that pressing Think button does not produce any result at all. This happens sometimes as we all know from personal experience. In this case, no update takes place and Robbie goes back to C unchanged.

Finally, after the new piece of information was added to the knowledge tree, Robbie can recompute the expected utility of the action (leading to) C_k that was affected by the new information and get the new $U(C_k|C)$. After that Robbie can choose again among the available actions with the new information in place. If it so happens that Think button still has the highest expected utility among all actions, Robbie will press it again. This will continue until he will run out of new ideas, after that the value of the Think button will decrease (see below) and he will stop thinking.

Also, do not forget that the thinking context C_θ also has component values v_{C_θ}, f_{C_θ}, and w_{C_θ} that can get updated after each usage of the button (see below). These values will determine how often Robbie presses Think button and in which contexts. For example, Robbie might have a theory that thinking should be done at work and at home there should be no thinking, but relaxing. Then, Robbie will have high w_{C_θ} for work related contexts and low w_{C_θ} for home related ones.

Also, repetition will matter since Robbie cares about familiarity. The more often he presses Think button in some context, the more familiar it will become and will increase in familiarity value (through updates, see below). This implies that learning to think in some contexts might make it easier in the future because you get used to it.

Finally, Robbie might simply enjoy thinking and have high value v_{C_θ}. Maybe this is because he made great discoveries in the past and this makes him excited about making more, so he presses Think button often. It can also be other way around. Robbie might feel stupid and not able to study and depressed because of that. He will then start hating Think button and will stop pressing it in all contexts. This might lead to poor development and sad consequences for Robbie.

As we will describe below, pressing Think button also decreases costs of thinking that get recorded in the update of the context C_θ. Each press makes context-specific and general costs of thinking smaller. This is the reflection of the same process that made $\alpha_C + \alpha$ in the utility function change when Think button is pressed. Thinking involves Focus and Concentration, and thus increases the capacities of parts of associative network where thinking is done. So, thinking about one subject, for example plumbing, makes Robbie better at thinking about plumbing in the future, the costs will decrease due to more associations and better understanding of the subject. General costs of thinking also decrease and affect thinking in all contexts. This is again due to training of Focus and Concentration.

As a result of these precesses, Robbie most likely will develop areas of expertise where he likes to think and thinks a lot. He will have low costs of thinking about favorite subject and his utility in such contexts will be more cognitive than in others (more oriented towards cognitive value). In other areas though Robbie might rarely press Think button, have low utility from cognitive value, never think in these contexts as a result, and react to everything emotionally without cognition (e.g., in bar X).

7.14 Updating

Up to now, we discussed how Robbie constructs utility and reasons about action using contexts stored in his knowledge tree. We did not however discuss how affective component values get there in the first place (v and f), and how contexts are connected to each other. In this section we do this via an idea of updating, which is a translation of value updating from reinforcement learning introduced in Tommy but reinterpreted in the context of this framework.

As Robbie walks around and experiences different contexts, he receives some value from the environment, feels something, and updates his values using Tommy and Molly. Suppose Robbie enters context C with the idea that he will receive $u(C|C)$, which is computed from the component values on his knowledge tree. However, instead he feels something else, say V. There might be many reasons why Robbie does not feel what he expects. Some features can change value, like milk can turn sour, or Nature could have changed the context somehow in a way that Robbie cannot fix.

Anyhow, we assume that Robbie updates affective value v_C:

$$v_C \leftarrow v_C + \lambda(V - v_C).$$

However, things do not stop here. The whole concept of this framework is to replace associative network from the first part of the paper with something more amenable. We introduced the space of contexts $\mathcal{C}$ as a replacement. But, there is a cost to this since on the network values simply get stored inside features and get accessed when necessary. But on the space of contexts, we need to artificially define these connections, because one feature is present in many different

contexts at once, and when its value is updated, so should the values of all contexts it is in. Moreover, our level of abstraction goes even higher and we treat whole contexts as having single values instead of thinking of their value as a complicated weighted sum of values and relevances of separate features.

To capture the intuition of updating on the space of contexts $\mathcal{C}$, we suggest the following simple procedure. We take the update $\lambda(V - v_C)$ and assume that the values v_D of contexts D close to C in some sense also get updated, but to a lesser extent depending on the similarity between D and C. So, if D has many relevant features shared with C, then D should update a lot, almost as much as C. But if D has little overlap with C, then it should be updated a little bit. Finally, if D and C are disjoint, no update should happen at all.

All this is captured by the similarity measure we introduced in early sections. We define *system update* following local update described above as follows:

$$\forall D \in \mathcal{C} \setminus C \qquad v_D \leftarrow v_D + \lambda(V - v_C)S(D, C)$$

Realistically, we do not have to update all contexts, but only those that share features with C. The idea of system updates following local updates will go through all updates that we discuss below.

When Robbie enters context C in addition to value v_C he also updates familiarity with the context. Since the context got experienced, Robbie updates:

$$\begin{aligned} & f_C \leftarrow f_C + \epsilon_f \\ \forall D \in \mathcal{C} \setminus C \quad & f_D \leftarrow f_D + \epsilon_f S(D, C). \end{aligned}$$

Here $\epsilon_f > 0$ is some small number that we take as constant. This is followed by the system update.

Notice as well that familiarity updates happen also when Robbie imagines contexts. We assume that upon imagining a context or seeing someone in it makes Robbie update familiarity same way only with an additional factor δ since context is imagined and not experienced. In this case the update is

$$\begin{aligned} & f_C \leftarrow f_C + \delta\epsilon_f \\ \forall D \in \mathcal{C} \setminus C \quad & f_D \leftarrow f_D + \delta\epsilon_f S(D, C). \end{aligned}$$

with the corresponding system update.

We do not update cognitive value w_C because it comes from the models of reality and does not have to change. It can change only when Think button is pressed.

Overall, the updates create the following picture of how Robbie's preferences develop. When Robbie is born he has $v_C = f_C = w_C = 0$ for all $C \in \mathcal{C}$ except maybe for contexts related to his

mother's heartbeat and her voice that are familiar since Robbie was in the womb (and maybe some hard-wired values like fear of snakes coming from Spot). Gradually, Robbie experiences real world and updates values v and f which also spreads over his context space $\mathcal{C}$ in the form of updates. Thus, Robbie learns how to act in the environments he never experienced before: they are semi-familiar and have some value from related experiences that produced updates in the past. With time, Robbie learns more and more through experience and this starts defining his future "preferences."

7.14.1 Think Button Updates

Finally, we should discuss the updates that happen when Think button is used. The most important update is related to how "well" Think button functions. In other words, whether or not it produces good results. If it does, then the value of thinking increases; if the results are bad, then the value of thinking decreases. We express this as the update of the component values in context C_θ.

Suppose that in C when Think button was pressed some new information got discovered that changed the expected value $U(C_k|C)$ of action leading to some C_k. Let us call the utility before thinking $U_{\text{old}}(C_k|C)$ and after thinking $U_{\text{new}}(C_k|C)$. We assume that the "quality" of thinking is determined by the change in value of the action $\Delta U = U_{\text{new}}(C_k|C) - U_{\text{old}}(C_k|C)$. If the difference is positive then the thinking is deemed good and vice versa. So, the update of v_{C_θ} goes as follows:

$$\begin{aligned} & v_{C_\theta} \leftarrow v_{C_\theta} + \lambda(\Delta U - s_C - s - v_{C_\theta}) \\ \forall D \in \mathcal{C} \setminus C \quad & v_{D_\theta} \leftarrow v_{D_\theta} + \lambda(\Delta U - s_C - s - v_{C_\theta}) S(D, C). \end{aligned}$$

Here, $s_C, s > 0$ are costs of thinking that are determined by the training of the cognitive system (how often it was used in the past). The system update also makes thinking in similar contexts more or less attractive.

It may sound counterintuitive why the quality of thinking is related to the new utility that the thinking process uncovered. After all, if someone discovers something with high utility by chance, it does not mean that this is high-quality thinking. This is true. However, the mind has no other way to estimate the quality, realistically. If thinking brings more utility, then it is useful, if it does not bring more utility then it is not useful. And we believe that this mechanism actually produces many inefficiencies. For example, imagine that Robbie thought about something and imagined a very bad consequence. He got scared, and his ΔU was very low. He updated his Think button value and after that he refuses to think anymore, because it was scary in the past. People might shy away from thinking because it produces sad or scary outcomes. Only Robbie who is dedicated and has high cognitive value of thinking w_θ can think through negative utilities. Such Robbie would have a theory that thinking is always better than no thinking and the

high value of w_θ will make him think even if he experiences something bad. In addition, thinking about beliefs can increase utility. This is because when cognitive probabilities are computed the cost ζ is removed from expected utility. So on average, attaching beliefs to uncertainties decreases the cost of uncertainty and can stimulate the usage of Think button in the future.

Related to this is the update of the familiarity of the Think button that happens as follows:

$$\begin{aligned} & f_{C_\theta} \leftarrow f_{C_\theta} + \epsilon_f \\ \forall D \in \mathcal{C} \setminus C \quad & f_{D_\theta} \leftarrow f_{D_\theta} + \epsilon_f S(D, C). \end{aligned}$$

We assume that this spreads over to other contexts, so familiarity of thinking becomes higher in similar contexts.

Next we discuss the remaining updates that are related to the functioning of Think button. First consider thinking costs s_C and s. When the cognition is young, these costs are high; as it gets trained the costs get lower. The costs get lower in specific contexts due to learning curve, and costs of thinking in general get lower too (Focus and Concentration are trained). We assume that, after the main update above, other updates follow. Costs are updated as follows:

$$\begin{aligned} s & \leftarrow s - \epsilon_s \\ s_C & \leftarrow s_C - \epsilon_{sc} \\ \forall D \in \mathcal{C} \setminus C \quad s_D & \leftarrow s_D - \epsilon_{sc} S(D, C). \end{aligned}$$

Here, $\epsilon_s, \epsilon_{sc} > 0$ are some small constants.

Further, we assume that pressing Think button increases the weight $\alpha_C + \alpha$ on cognitive value in the utility function $u(C|C) = v_C + (f_C - \chi) + (\alpha_C + \alpha) w_C$, thus making cognitive value feel more important. The update happens as follows:

$$\begin{aligned} \alpha & \leftarrow \alpha + \epsilon_\alpha \\ \alpha_C & \leftarrow \alpha_C + \epsilon_{\alpha c} \\ \forall D \in \mathcal{C} \setminus C \quad \alpha_D & \leftarrow \alpha_D + \epsilon_{\alpha c} S(D, C). \end{aligned}$$

Here, $\epsilon_\alpha, \epsilon_{\alpha c} > 0$ are some small constants. It may be thought that ϵ_α probably should be smaller than $\epsilon_{\alpha c}$, given that α represents the increase in cognitive skills across all domains and α_C only in the specific context C. But we leave this to future research.

7.15 Choice Process

With all the preliminaries above, we are finally ready to describe how Robbie makes choice. It is not too difficult with all the notation we have developed. Suppose that Robbie enters context C and his knowledge about C before entering is coded in the values (v_C, f_C, w_C). Suppose that

once in context C, Robbie feels what is going on in it as some value V. As described above in Section 7.14, this leads to updates of values to, what we denote, v_C^d and f_C^d (including system updates). We have the subindex d to emphasize that these are values *after* updates. Then Robbie computes his current mood as utility of C after the update:

$$M = v_C^d + f_C^d + (\alpha_C + \alpha)w_C.$$

Notice that this is the same M that was used in Section 7.10 to compute Tommy's beliefs.

In the next stage, Robbie determines his available actions. As we mentioned above, it might be that Robbie has already been in context C and he has knowledge tree connected to it. In this case, Robbie simply proceeds to the next stage (see below). If Robbie has never been to C before, he will not have knowledge tree connected to it, and thus he needs to do something else. Robbie can, for example, take all action features that are contained in C and use them as candidates for actions.

Alternatively, we believe that Robbie uses yet another mechanism to determine actions in C. Realistically, we never enter the same context as before. Contexts are always slightly different, even when it is the context of Robbie's home, where he spends most of his time. Robbie might wear different socks each time, or he might see various things on TV that change his mood. This however does not change his actions or his view of the current context much. The point is that since we never enter the same context, a mechanism should be in place in the mind that discards some small unimportant features in the context and treats it as some context that has been experienced before.

This can happen in the following way. Suppose Robbie has been in context C and made choices in it in the past. Then, we can imagine that there is some small ϵ_1-ball $B(C, \epsilon_1)$ around any C in S'-topology such that if C is experienced, then all contexts $D \in B(C, \epsilon_1)$ also are considered as the same context C and are treated with the same knowledge as C. We treat ϵ_1 as an individual parameter and leave it to future research to understand what determines it.

Following the same idea, Robbie might try to look for a similar context to learn about the current context C. Robbie's cognitive abilities might determine the size of the ball around C where he can search. Let us denote this ball $B(C, \epsilon_2(\iota, s_C + s))$ and say that $\epsilon_2(\iota, s_C + s) > \epsilon_1$. Here, we emphasize that ϵ_2 can depend on the thinking costs $s_C + s$ as well as imaginativeness ι that we define below in Section 7.18. In any way, we assume that when in C, Robbie can check out contexts in the ball $B(C, \epsilon_2(\iota, s_C + s))$. If he finds a context that is connected to the knowledge tree, then Robbie can use the knowledge from that context as a guide in the current context. The fit might be not perfect, but it will give Robbie an initial idea what to do.

In addition, Robbie can press Think button to determine the actions. It might be a special occasion where Think button is pressed often because Robbie finds himself in a context where he does not know what to do (and it looks like he is staying in it until he chooses something).

In such situations, Robbie will have no choice but to press Think button, because it is the only choice he might have at all.

Now suppose that Robbie has determined the actions that he can choose from. Suppose these are actions leading to contexts $C_1, ..., C_n$. Plus there are contexts C_θ and C_ϕ. Thus, Robbie's full assortment of actions is $\{C_1, ..., C_n, C_\theta, C_\phi\}$. For these actions, Robbie computes expected utilities $U(C_i|C)$, which he can do given his mood M (to compute Tommy's beliefs in Section 7.10) and the information in the knowledge tree, and then chooses the action that maximizes

$$\max_{i=1..n,\theta,\phi} U(C_i|C).$$

If his choice is Think button, then things unfold as described above in Sections 7.13 and 7.14. Thinking produces new information on the knowledge tree (or not) and updates many coefficients and values in C and C_θ. After that, Robbie is back at now updated maximization problem.

If it is the Do Nothing Else action, then Robbie gets back to the same context after updating C_ϕ (unless forced transition happens). Finally, if Robbie chooses an action that takes him to another context, say C_k, then Robbie moves to C_k experiencing the cost of transition $\kappa(C_k|C)$ on the way. At C_k the choice process repeats again as described in this section. Life of Robbie is thus a sequence of choices and forced transitions that move him from one context to the next.

7.16 Continuity of Preferences

Given the choice process described above, we can formulate a proposition, proved in Appendix I, that Robbie component values v_C and f_C are continuous in the space of contexts $\mathcal{C}$ with S'-topology. This is so after any number of updates that Robbie might perform. To understand why this might be the case, let us consider a system update of any variable, say v, after experiencing context C. It is always of the form $v_D \leftarrow v_D + hS(D, C)$, where h is some number. This can be seen as two functions of D being added: v_D plus $hS(D, C)$. Notice that $S(D, C)$ satisfies properties similar to $S'(D, C)$ that defines the topology on $\mathcal{C}$. Thus as we show in Appendix I, $hS(D, C)$ is a continuous function of D on $\mathcal{C}$. So, the update that adds $hS(D, C)$ to v_D adds two continuous functions. $hS(D, C)$ is continuous by proposition in Appendix I and v_D is continuous because it is a sum of continuous updates from the past. So as system updates are being applied, they are continuous in nature and thus they do not change the continuity properties of v_C and f_C.

Why do we care about this? Continuity is an important property that can be of tremendous value in applications. It essentially implies that we do not need to know what Robbie might feel in all contexts C, but instead we can approximate his values at C from the values of the surrounding contexts for which we have data. Continuity says that we can take some weighted

average of these known values and arrive at a good approximation of values in C. Without continuity we would not be able to make such a claim.

Moreover, continuity also implies that little changes to contexts should not change the optimal behavior (if Think button is not pressed). Indeed, suppose that Robbie's knowledge tree is fixed and does not get updated with new information. Then, Robbie's reactions to continuous changes in context will also be continuous. He will gradually switch from one optimal action to another. In addition, continuity also guarantees that the actions available to Robbie do not just randomly pop up in and out of existence in some contexts. The available actions are most likely the same as in similar contexts (due to continuity in S'-topology) and their expected utilities will also be continuous in changing context as long as knowledge stays the same. When the knowledge changes, this can create a discontinuity in behavior and values until the knowledge gets incorporated into the knowledge tree after which the behavior becomes continuous again.

7.17 Interpolation of Affective Values from Data

From the model above, we can understand how to approximate component values v and f for new contexts using the data about surrounding contexts. Suppose that we, as researchers, have selected some set of events, that took place in reality and are documented, that we believe have influenced some new context C that people never experienced before. We want to know what people might feel in C. Suppose that we possess data in the form $(v_{C_i}, f_{C_i})_{i=1..n}$ for some events $C_1, ..., C_n$ that happened before. We also know the similarity $S(C_i, C)$ of all these events to the context of interest C.

Notice an interesting property of system updates. Whenever some component value is updated, the same update happens in close contexts only multiplied by the similarity $S(C_i, C)$. Thus, if we look at value v_C at some context C as a sum of past system updates (coming from contexts other than C), we will see that past elements of this sum form exactly the current values of events that were updating v_C (times $S(C_i, C)$). If we take into account only the influence of $C_1, ..., C_n$ on the values in C and assume that all other experiences are zero on average, then we can calculate the estimated affective values in context C as

$$\begin{aligned} \hat{v}_C &= \sum_{i=1..n} v_{C_i} S(C_i|C) \\ \hat{f}_C &= \sum_{i=1..n} f_{C_i} S(C_i|C). \end{aligned}$$

In other words, without being experienced, the affective values of C are just the weighted sums of values coming from all contexts that influenced C in the past. Thus, the data we have can give us simple estimates $\hat{v}_C$ and $\hat{f}_C$ of values v_C and f_C. This means that we can interpolate affective values from data for the (possibly never experienced) context of interest C.

Notice that we do not discuss similar mechanisms for cognitive values w_C. This is because cognitive values are constructed from models of reality. This means that the mixture of values from similar contexts $C_1, ..., C_n$ in the sense of S'-topology might not be a very good predictor of the cognitive value of C. It might be easier to understand which models of reality generate w_C than approximate them from similar contexts.

7.18 Imaginativeness

When we talked about Robbie being in context C, we subtly assumed that Robbie can be in any context C. We never discussed how *possible* it is for Robbie to be in different contexts. Indeed, the mind might contain millions of features that are all highly interconnected with each other. So then it is unlikely that Robbie is physiologically able to activate all these features at once. It would probably lead to an epileptic seizure anyway. All this implies that Robbie should have some limits on the contexts that he can be in. For example, the context where all relevances of all possible features in $\mathcal{F}$ are 1 should somehow be excluded.

We suggest the following simple idea, which is also connected to the models of cognition in the first part of the paper. There we proposed that, during Focus, features that Robbie focuses on get activated with an additional boost of relevance in the network. The size of this boost eventually determines how many nodes further down the network will be activated and how strongly since signal decays. The higher the boost, the more imaginative Robbie will be. He will be able to imagine more features, and also more features at the same time.

This intuition can be expressed as a limit on the size of the context $|C|$, which is just the sum of all relevances in it. Suppose that Robbie can only be in contexts C with $|C| \leq \iota$, where $\iota > 0$ is some fixed upper limit on sum of relevances (or imagination). This will do several things. When ι is low, Robbie will not be able to imagine many features since most available relevance in any context will be consumed by the features activated from reality. Such Robbie with low imaginativeness ι will not be able to think well and use models of reality that take imagination. Robbie will not be able to trace expected utility much down the knowledge tree thus becoming myopic. His δ will be probably low due to his inability to activate additional features due to imaginativeness limits.

When ι is high, Robbie will be able to do much more. He will be able to imagine many different features associated with some context; he will be able to imagine many features at once; and he will be able to focus much harder on something, thus spreading a stronger signal down the associative network. All these qualities will make Robbie much more imaginative, able to perform abstract thinking, able to better predict the future (he can look further down the knowledge tree), better use models of reality, etc.

In addition, we should not forget about the influence of ι on the cost of transiting between states $\kappa(\cdot|C)$. Suppose that Robbie needs to transit to context C with $|C| > \iota$. He simply will not

be able to imagine it, so he will not be able to make the transition and will transit to some other state with less total relevance. This probably happens when students try to learn for some hard exam and fail. The cost of transition is therefore infinite, which makes it likely that Robbie will try to abstain from such transitions all together. It would be interesting to know if this simple abstraction can be helpful to study limits of human cognition. We leave this for future research.

7.19 Representation of Other Agents

It is interesting to consider how to represent other agents in this model, which will lead the way to game theoretic considerations described below. To represent agents, we can extract a new set of features of interest, like we did with action features, from the set $\mathcal{C}$ and treat them as special contexts. Suppose agent i is represented by feature $i \in \mathcal{F}$. This is an agent feature (see section on language in the first part of the paper). Now, when agent i meets Robbie, Robbie gets feature i, and everything that is associated with it, active. When meeting his friend, Robbie remembers that he is a colleague from work who likes chess and beer. So, we can say that agent i is a special context $I_i \in \mathcal{C}$ where feature i has high relevance and also other features associated with i like chess or beer are relevant as well. Context I_i contains all information about i that Robbie has.

To understand what happens when Robbie meets i let us define a simple operation on fuzzy sets, addition. For any contexts $C = \{(k, p_k)\}$ and $D = \{(k, q_k)\}$ let

$$C + D = \{(k, \min\{p_k + q_k, 1\})\}.$$

This is just the sum of relevances, which is capped at 1 since nothing can be relevant more than the highest relevance, so we just cut the sum if it exceeds 1.

With this device we can imagine that, when Robbie meets i, Robbie transits from the context C where he was before i entered into the context $C + I_i$. This new context is just like C only relevances of things related to i are added. So, we can say that the presence of people $i = 1..n$ around Robbie puts him in context $C + \sum_{i=1..n} I_i$. What Robbie does in this new context will depend on his knowledge tree and his component values. For example, if the presence of some I_i reminds Robbie of recreational drugs, he might try to get away from that person to not get addicted again.

Notice as well that, as a context, I_i has component values (v_i, f_i, w_i). The value v_i records information about how i treated Robbie in the past. For example, if i was mean to Robbie then v_i will be negative and vice versa. Value f_i records familiarity with i: how many times Robbie met i or imagined i or heard about i from others. Finally, w_i records cognitive value of i that can be high for example when i is a scientist (in case Robbie uses physics and biology as his models of reality) or a king (in case Robbie respects monarchy).

In addition, notice that adding new contexts to the existing one increases the total sum of relevances in the new context with other agents. This can create problems given that Robbie

might have limits on imaginativeness ι, or how much total relevance he can use in his imagination. Thus, having many people around might overload Robbie's system due to too much new relevance that is being added. In this case, Robbie might escape to some quieter context.

7.20 Morality

In this section we discuss possible extensions of the model that can include moral considerations of Robbie. To understand what sort of morality Robbie can exhibit, we should first think about the "bounded rationality" aspect of Robbie. At two extremes—when Robbie never presses Think button versus when he presses it all the time—Robbie will develop into very different persons. The Robbie who never presses the button will never learn any models of reality, will never attach any cognitive probabilities to events or have cognitive value of them. He will live purely on affective values v and f without any cognitive abilities developed. Moreover given the absence of cognitive value, v will mostly contain the "default" values for things that our bodies like, for example sugary foods, sex, friends, etc. Thus, non-thinking Robbie will have mostly affective mind and will be driven mostly by *affective morality*.[9]

Unlike cognitive morality discussed below, affective morality is not based on empathy or other cognitive computations that can be done with the comparator. It is based on affective values v and f attached to other agents or groups of agents that arise as social identities (see Kimbrough and Vostroknutov, 2022). Affective agents (like Robbie who never pressed the button) do pro-social things when they have high affective values of contexts in which they are and vice versa. For example, if affective Robbie has high affective value attached to some agent i, he will simply try to be next to this agent as much as possible and will try to be nice to i. This can quickly turn into the situation where Robbie is doing anything that i wants. Thus, we should expect that (pure) affective morality is based on respect of individuals with high status (family, kin, friends, bosses, priests, kings, etc.) and doing what they tell you to do. Traditions and customs evolve around these ideas in most cultures and become familiar, thus making them even more attractive in terms of affective value (familiarity). As a result, a social identity develops that dictates people who belong to it how they should live their lives. So, affective morality is the traditional morality based on rules and values coming from some social identity (see more on this in Kimbrough and Vostroknutov, 2022).

At the other extreme, super smart Robbie—who got extremely high values of α_C and α in his utility from constantly pressing Think button all his life—will completely stop caring about affective values v and f and will be guided exclusively by his models of reality. This Robbie is a crusader, a monk, a mathematician, or a philosopher. With respect to social relationships, such Robbie will use *models of social reality*. There can be many of those. For example, if Robbie lives

[9] In addition, Robbie might use *moral rules* that are the simplified versions of cognitive morality (Kimbrough and Vostroknutov, 2023a).

in Roman Empire, he might have a model that the Emperor is the most holy person in the world who is a God, our protector, and the guarantor of peace. He has very high status and we should always do what he says. In Ancient Rome, Robbie might have a model of social reality that attaches cognitive values (aka statuses) to different agents depending on their position in the society or their religious role. If smart Robbie lives in modern Western world, he might have a different model of social reality that says that all people should be treated equally. In this model, all people have the same, equal social status and their high political roles are considered as just services to society. The principles of law, equality, and freedom rule the land. Thus, cognitive models of social reality, or *cognitive morality*, can differ and we should know what they are to understand how some population behaves.

It is important to add that we assume that cognitive morality is always based on empathy, or other sorts of computations done with comparator, where some values across agents are compared (see Section 6.1.3). We think of these comparisons as instances of reasoning used to compute social norms (see Kimbrough and Vostroknutov, 2023c). Thus, they belong in the cognitive domain.

To understand how cognitive morality works we can use the model of injunctive norms and punishment by Kimbrough and Vostroknutov (2023c,b), further KV, where social norms are assumed to arise from dissatisfactions of various agents with various outcomes. Aggregation of such dissatisfactions across agents gives a measure of social appropriateness for each outcome that is then used as a normative guidance to choose among them. This model is very much in line with the whole framework discussed in this paper given that the computation of the norm is *modular*. Each dissatisfaction is computed with one usage of comparator, thus making such aggregated norm a typical outcome of cognitive processes we discussed in the section dedicated to Robin.

To continue with our argument, we will assume that Robbie has the theory of KV in his mind as his model of social reality. In this theory, it is assumed that each agent i has a social weight $\tau_i \in \mathbb{R}$ (aka status) that determines how important the dissatisfaction of i is in the computation of the norm. When τ_i is close to zero, Robbie thinks that this agent is irrelevant and can be ignored. When $\tau_i = 1$, Robbie treats the dissatisfactions of i as he treats his own. When τ_i is very high, Robbie cares about i more than about himself (happens with parents and children, or peasants and kings). When τ_i is negative, Robbie wants to increase dissatisfaction of i. He will want to hurt i more, the lower τ_i becomes.

The point of this is that the model of KV proposes a method that can be used to compute morality based on empathy and this method fits into our framework very well. We can imagine how exactly Robbie uses comparator to compute dissatisfactions, which connects the two models together.

However, the most interesting part is to imagine how Robbie would behave when he is affected by *both* affective and cognitive morality at once. To do that we can use the social weights

from the model of KV and think how they are determined in the mixed version. We follow the same intuition as when we mixed beliefs between the three systems. We use the coefficient $\alpha_C + \alpha$ to compute the relative weights of the three "component statuses."

Remember that each agent i has affective value $v_i + f_i - \chi$. This contains records of familiarity with i and how nice i was in the past. We will treat it as the *affective status* (or affective social weight) of i. The value w_i corresponds to the cognitive social weight, or *cognitive status*, of i that comes from the model of social reality. Let $h_C = 1/(2 + \alpha_C + \alpha)$. Now we can just take a convex combination of these numbers and set $\tau(i|C)$, the new more consistent notation that now emphasizes dependence on the context, to be

$$\tau(i|C) = h_C(v_i + (f_i - \chi) + (\alpha_C + \alpha)w_i) = h_C u(i|C).$$

The idea here is that the more cognitive Robbie becomes, the more he relies on cognition in his estimates of statuses of other people. A very cognitive Robbie will forget about personal grievances and familiarity with others and will treat them as equals (in case of Western cognitive morality) even if they did some harm to Robbie in the past. A very affective Robbie will ignore broadly shared social norms that should guide his cognitive morality and will base his judgements of others on personal interactions, past experience, personal grievances, etc. He can be vengeful or authoritarian (with those of lesser status). Thus, inserting these social weights into the model of KV allows us to model mixed affective/cognitive morality within Robbie.

Moreover, notice that status $\tau(i|C)$ of i in Robbie's imagination depends on the current context C and Robbie's cognitive weight α_C. Similarly to the discussion in Section 7.9, this then implies that Robbie's idea of the status of i will not be fixed, but will *change* with the context in which Robbie is. For example, when Robbie is in the context of voting for the President, Robbie might be very focused and cognitive since voting is an important duty. Robbie might believe that the presidential candidate i he votes for has high cognitive value w_i, because the candidate shares Robbie's views on how the world works. This makes Robbie vote for this candidate. But when Robbie is at home arranging flowers on his windowsill, he is relaxed and not very cognitive and as a result his perception of the status of the candidate is now lower, because Robbie's cognitive weight on w_i is small. So, if asked whether Robbie would trust this candidate or vote for him, Robbie might say that he won't because in the context of arranging flowers he does not feel that this candidate is somehow better than others.

This example suggests that typical social behavior, that is a mixture of affective and cognitive moralities, should be context-dependent to a degree when people perceive the same person differently depending on how they feel themselves in different contexts they are in. Intuitively, this idea does not seem flawed, as we observe constant changes in trust and willingness to support presidential candidates that are brought about by factors beyond control of these candidates (e.g., the current state of the economy influences the support of the current president).

7.21 Games

It is not hard to imagine how to model games in our framework. Suppose Robbie is in context $C \in \mathcal{C}$ and suddenly some agents $I_1, ..., I_n \in \mathcal{C}$ enter and Robbie needs to interact with them somehow (which is described in context $R \in \mathcal{C}$ that contains features that describe the rules of interaction). Then Robbie moves to context $C + \sum_{i=1..n} I_i + R$ and starts thinking about his actions. Suppose he knows that the actions are $A_1, ..., A_n$ that lead to some collections of uncertain contexts that can happen depending on moves of other agents. So, Robbie can imagine contexts C_{A_i} in which he performs A_i. Then he imagines the connections from C_{A_i} to the contexts representing outcomes of the game (including what happens to other agents). There can be many emanating from each C_{A_i} depending on moves of others.

This gives us the knowledge representation of a normal-form game. In it, Robbie can assess various component values, think about equilibrium play, etc. depending on what model of reality he has. If Robbie is a game theorist, then he might try to analyse the situation he is in using some game theoretic concepts stored in his mind. After that, he can attach cognitive values w_C to all contexts C that represent outcomes of the game. These w_C might depend on the assumptions about rationality of others, what they are going to play, beliefs, equilibrium, etc.

The same goes for extensive-form games, which are even simpler to represent on the knowledge tree. They are both trees, so it is just natural. Each node on the game tree translates into a context in which someone chooses something, etc. until the final nodes with payoffs are reached that are also represented by imagined contexts where the payoffs are received.

All this suggests that games—as we, people with PhD, are used to them—can be easily represented on the knowledge tree and Robbie can perform all the game theoretic reasoning he wants if he is a game theorist. But the very interesting question is What if Robbie is not a game theorist? While game theoretic concepts dovetail nicely with our framework, it does not mean that actual human beings represent games like that. People do not know game theory and might not even understand that they are in a strategic interaction at all. It is possible that they should be reminded of that to activate any strategic reasoning.

Thus, it would be very interesting to know how exactly people represent real-life strategic situations in their minds. Understanding this can give us a tool to predict what *actually* the outcomes of strategic interactions in the real world might be. We believe that our model of knowledge is flexible enough to provide different representations that can be tested experimentally or in the field. We leave this to future research.

7.22 Institutions

In this section, we propose a theoretical connection between psychological properties of Robbie and some broad types of institutional setups and state organizations that we observed throughout history and observe in the present. Namely, we will talk about Robbie's cognitivity recorded

in coefficient $\alpha_C + \alpha$. It determines how much Robbie relies on models of reality in his utility judgements. The coefficient itself increases when Robbie presses Think button.

So, suppose we have a population of affective Robbies who did not press Think button very often. They rely mostly on affective values in their judgements, are not very educated, act "on emotions." They are also not able to think about reality on their own—outside their contexts of familiarity like home town or country—because they do not have good models of it. In addition, they only listen and respect people they like personally, like family members, kin, tribe leaders, bosses, etc. This implies complete distrust to strangers outside this circle.

The question that comes to mind is What is the best way to organize affective Robbies so that they could achieve some broad(er) coordination and cooperation? Given that the only way to influence them is through someone of high status they respect and listen to, to organize affective Robbies' behavior that someone needs to tell them to do it (otherwise they won't listen: they do not have their own models of reality to verify what anyone else they do not personally trust says). From this it follows that affective Robbies will gradually self-organize into nepotistic networks of family and friends with more powerful networks ruling over others. At its extreme, this can lead to an authoritarian Empire.

An important implication of this argument is that given that affective Robbies are affective and do not use cognition too often, there is no other way they can possibly self-organize than through nepotistic networks of personal connections. This suggests that attempts at "implanting" democratic rule might not be successful in such situations. Affective Robbies simply do not understand the world outside their social network.

This also implies that situation can only change when affective Robbies become less affective and start pressing Think button more often. And indeed, such process can unfold on itself. This happens when affective Robbies mix with others from other tribes in a town for the purpose of trade. As towns grow, they present a new, much noisier environment in which things are much less predictable than in familiar stable environments of small villages that affective Robbies like so much. In towns, cognition becomes useful to Robbies because they need to understand how to navigate the complex social environment. They start pressing Think button more often.

As Robbies become more cognitive in historic time, it becomes possible that their morality also becomes more cognitive. As cognition gets developed, it might become more empathic by construction, simply because the ability to calculate dissatisfactions is seemingly embedded in it. It is a simple comparator operation. All this will lead to the development of models of social reality usually summarized by philosophers. Cognitive Robbies like new theories of social reality and start following them in life as well, thus becoming the embodiments of the theories themselves to a certain degree. In this way, the morality can shift from affective to cognitive gradually, which can also lead to change in institutions.

Cognitive morality demands that institutions reflect the core idea of the morality itself. For example, Hobbes' idea of Leviathan suggests that monarchy is the perfect type of social organi-

zation because monarch (though he might be not super awesome sometimes) still guards people from chaos and promotes prosperity and safety. Thus, we should love our monarch and respect him. From this follows that people who share this idea will support monarchy and think that without monarchy everything will turn into chaos, and so it is the only way to keep things in order.

Similarly to this, cognitive morality of the Western world, rooted in the ideas of equality and freedom, demands that the society is built on democratic principles. Other forms of rule seem evil to people with such cognitive morality. As a result, democracies and democratic rule are supported from within by cognitive Robbies, who share the principles on which democracies are founded and moreover have enough cognitive capacity to care about these ideals to a large enough extent.

The point of this discussion is to emphasize that cognitive properties of agents within the system can define the type of institutions they self-organize into to a large degree. If agents do not rely on models of social reality (affective Robbies), they will organize into nepotistic networks. If agents do rely on models of social reality, they will organize into institutions that reflect the core principles of their cognitive morality. Thus, new rules and institutions cannot be simply imposed on Robbies from above under assumption that they will simply switch to the new rules unconditionally. Robbies can only live in institutions that reflect their individual morality, be it affective, cognitive, or mixed.

In order to change institutions—the model thus suggests—we need to change Robbies' psychology. Specifically, education comes to mind. In good education system, Robbies are forced to press Think button a lot. This develops cognition and helps Robbies to form some ideas about social order they live in. More education is eventually the path to better societal organization. Though, it should be mentioned that it takes a lot of time and not one generation to increase average education level. The morale of the story is that institutional change takes time because for it to happen agents should change psychologically first.

8 Concluding Remark

We do not claim that theory of minds presented here gives some kind of a precise or exact representation of how human mind actually works. Rather, we present a novel way of thinking about biological organisms, based on the new type of mathematical abstraction, and provide a sketch of the framework that we hope can be used for scientific research into human and animal behavior, can be modified, tested, and improved. We see the value of this theory in the fact that it provides one treatment of many different phenomena using few basic abstract elements and believe that it is time for social and biological sciences to advance to the level where we can talk about whole organisms, their behavior, their minds, and all their properties together instead of using a patchwork of mutually inconsistent models.

We also hope that this theory can be a step to unification of various fields of research that deal with studying life in general around one mathematical language, so that we can start the conversation among all of them using a common foundation.

References

Bush, R. R. and Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological review*, 58(6):413–423.

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

Gregson, R. A. M. (1975). *Psychometrics of similarity*. Academic Press, New York.

Kimbrough, E. and Vostroknutov, A. (2023a). A meta-theory of moral rules. mimeo, Chapman University and Maastricht University.

Kimbrough, E. and Vostroknutov, A. (2023b). Resentment and punishment. mimeo, Chapman University and Maastricht University.

Kimbrough, E. and Vostroknutov, A. (2023c). A theory of injunctive norms. mimeo, Chapman University and Maastricht University.

Kimbrough, E. O. and Vostroknutov, A. (2022). Affective decision-making and moral sentiments. mimeo, Chapman University and Maastricht University.

Robinson, J. A., Vostroknutov, A., and Vostroknutova, E. (2023). Endogenous institutions and economic policy. The World Bank Policy Research Working Paper no. WPS10600.

Rusch, H. and Vostroknutov, A. (2023). The evolution of personal standards into social norms. mimeo, Maastricht University.

Zwick, R., Carlstein, E., and Budescu, D. V. (1987). Measures of similarity among fuzzy concepts: A comparative analysis. *International journal of approximate reasoning*, 1(2):221–242.

Appendix

A Composite Features, Concepts, Scale-Free Networks

In our framework, we consider features connected by associations with different capacities. The legitimate question is what these features and associations exactly are and how to think about them on conceptual level.

It is important to note that for different minds these definitions can vary. For Spot, who cannot associate at all, features correspond to low-level sensors that detect elemental outside stimuli. These can be, for example heat, taste, smell, color, etc. So, when Spot perceives a bear, what he is capable of "noticing" is not really a bear, but rather color brown and smell. So, Spot can only perceive sensory features in their original meaning. For Tommy, the situation is the same: Tommy cannot associate features with each other, so he also cannot really "see" a bear, but only elemental pieces of it like color and smell.

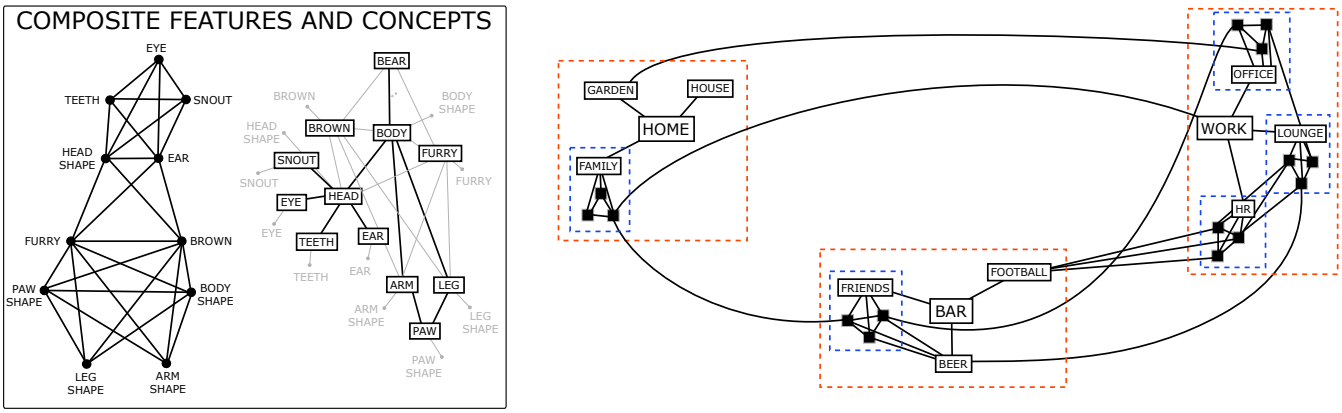

Figure 15: **Left Panel.** Composite features and concepts in Freddie and higher-level minds. **Right Panel.** A scale-free associate network.

Once we get to associations in Freddie though, the situation changes. Freddie can associate elemental sensory features with each other, so he can construct what we call *concepts*, or clusters of heavily connected features. For example, Freddie can look at bear's paw and can create associations between basic shapes in it, as they are recognized by low-level visual cortex. So, Freddie can create a concept of a paw that is a cluster of interconnected elemental shapes. Then Freddie can construct concepts of higher level. Consider the left network in the left panel of Figure 15. It shows how Freddie can represent a bear. It consists of interconnected elemental features (Brown, Furry) and possibly some concepts like bear's head, leg, ear, etc. So, Freddie can perceive a bear as a concept in itself. Notice that Freddie cannot talk, so when he sees a bear what he perceives is something furry, brown, and shapes like ears, eyes, etc.

When we get to Talking Molly, the situation changes again. Talking Molly has word-features that allow her to create different types of concepts. She has words for all parts of a bear, and in

fact she has a word "bear" for the bear itself. So, for Talking Molly, the concept of a bear consists of words (in white rectangles), connected on a tree-like structure (the right network in the left panel of Figure 15). The words are associated with the sensory features. So, when thinking about the word "paw," Molly perceives all associated shapes within a paw. This is shown on the network with grey links. Still, Talking Molly perceives a concept of a bear similarly to Spot just using more features and adding additional structure that words allow for. This makes her able to understand for example that bear's body includes head, arms, and legs. All minds above Talking Molly form concepts in the same way.

Concepts are very important for understanding how our perception of the world works. This is because the whole concept can activate all at once when a signal passes through the associative network. Given that concepts are highly interconnected, whenever a signal hits one feature inside a concept, it spreads easily through all features in it, since they all have connections with many other features within a concept. The signal will bounce back and forth inside the concept keeping it active and lit up for potentially a prolonged time.

This is a very useful characteristic of associative networks. Indeed, it allows Freddie or Molly to recognize a bear very fast from seeing only one paw, or smelling it, or seeing something furry and brown. Once a *sub-feature* within the bear concept lights up, the signal spreads all over other features within it and Freddie perceives a bear (or its associated value) and can run away.

However, this same characteristic of concepts can have drawbacks. Indeed, Freddie can make mistakes. The bear concept can light up also when Freddie sees a beaver, which is also brown and furry. So, the fact that concepts are easily activated in the mind can create a lot of false positives when Freddie mistakes a beaver for a bear. This problem persists on all levels of minds and can get really severe. For example, we believe that stereotypical thinking is exactly that: activation of previously formed concepts in situations where they do not apply. Suppose you have had some really bad experiences with people who have green beards in the past. So, you have formed a concept of a "green-bearded man" that has a very negative value in your mind. When you travel to another country where people are really nice, but some of them have green beards, you will feel bad whenever you see a green beard and will treat a person with it in a disrespectful way. This is because green beard automatically lights up the whole concept of "green-bearded man" even though it is completely uncalled for in the new environment.

On the conceptual level, the concept of concept is useful because it allows to see concepts as "features" connected by "associations" and thus allows to treat associative network as *scale-free*. To give an example of what this means, consider the world of a typical Western man called Bob, shown in the right panel of Figure 15, who goes to only three places: home, work, and bar. The picture shows the representation of the world by Bob. He uses words in white rectangles to associate the places with each other and with different people marked with black squares. For example, Bob associates the word "family" with his wife and two children, which creates the concept Family. Similarly, he associates "friends" with his drinking partners in the bar, and

same goes for his colleagues in the office, in the lounge, and in HR. These mini-networks, marked with blue dashed rectangles, consist completely of elemental features: agents and words. Thus, they represent the lowest scale within the associative network.

Bob can also think about a higher scale. For example, Home consists of Garden, House, and Family (a concept). Similarly, Bar consists of Friends (a concept), Beer, and Football; Work consists of Office, HR, and Lounge, all of which are concepts, or networks of lower scale. We can think of these higher scale networks as having features mentioned above and connected by "composite associations" that are simply the collections of associations that connect different concepts within a higher scale network. For example, people in the lounge know some people in HR, there are three connections between these sub-networks. Thus, we can say that on a higher scale the concepts HR and Lounge are heavily connected, or have high capacity of connection.

We can proceed to yet higher level and consider three concepts: Home, Work, and Bar marked with red rectangles. These concepts, when seen as "features" are connected by "associations" of different capacity. Home is connected to Bar with only one link (Bob's wife knows one of his drinking friends). Thus, the capacity of this connection is low. Home is connected to Work with two connections (wife comes to work sometimes and a colleague from the office is a fellow gardening enthusiast). So, the Home-Work connection has higher capacity than Home-Bar connection. The capacity of Work-Bar connection is even stronger: there are five links (all HR likes football, a colleague from the office knows a drinking partner, and a colleague from the lounge drinks a lot of beer).

As we can see, the associative network is scale-free because it can be seen as the same kind of network on many different scales. This is a very useful property, because it allows us to model associative networks on arbitrary scale and not bother with details of which features are concepts and which are not. For example, depending on the application we can model the mind of Bob on three levels:

Level 1. "Life of Bob." Features: Home, Work, Bar;

Level 2. "Bob at home." Features: Family, House, Garden;

Level 3. "Bob and HR." Features: HR colleague 1, HR colleague 2, HR colleague 3.

On the last note, the value of a concept, like Family, can be defined as the sum of values of all its sub-features.

B Automatic Minds and Evolution of Values and Associations

When discussing Spot, we assumed in the main text that each sensory feature is connected to only one action. This assumption was made for simplicity and to make the argument. However, it does not have to be this way. There are two things that can be different.

First, it may be the case that features are connected to multiple, not mutually exclusive, actions that can be performed simultaneously. It also can be that actions themselves are connected to each other to form "action programs." For example, many animals are born with the ability to walk, which suggests this mechanism: walking is a complex coordinated activation of many action features responsible for different muscles, so it is plausible that such connections are "hard-wired" on genetic level. In principle, we can imagine that Spot has any possible set of fixed connections. This does not change his conceptual nature: he is still an automatic mind that produces same action output for the same sensory input.

Second, it may be that Spot can learn new associations from the environment, thus obtaining some characteristics of Freddie, only without having values. We do not know for sure that the ability to change associations (Freddie) has necessarily evolved after the valuation system (Tommy). It might be that these systems evolved simultaneously. We do not have any evidence for this, so we leave it for future research.

C Other Definitions of Mood

In the main text, we assumed that there are only two mood features in the value aggregator: bad mood and good mood features. The same holds for derivative features. It may well be that there are more of them. This would allow an organism to distinguish mood in better detail. For example, it can be that there are two good mood features: "simply good mood" and "very good mood." The same can hold for other features (derivatives) as well. We do not have any way to tell if this is the case or not, so we leave it to the future research.

It can also be that the division of the continuum of values into mood intervals depends on the species. We chose the simplest division: if the value is below zero then the mood is bad. If value is above zero, then the mood is good. But, in both humans and animals it can be that there are many arbitrary intervals of summed values that are mapped into different mood features (same can be true for derivative features). This might have some advantages, depending on the species.

D Alternative Assumptions on Signal Spreading

In the main text, we make the simplest assumptions on how the associative network functions. For example, we assumed that when Freddie perceives some active features, the associations between them increase their capacity by a small amount $\varepsilon > 0$ in each discrete time period. This might be too simplistic for some applications for at least three reasons. First, it might be that this rate of increase depends on other parameters. For example, it is not inconceivable that ε is a function of the strength of activation (relevance) of the features that are being associated. When there is a feature in the environment that has very high relevance, this might make ε for links connected to it larger than for features that have low relevance. In other words, strong activation makes associations stronger faster.

Second, we implicitly make assumptions that time is discrete in the model. This modeling choice is, of course, questionable since continuous time might be more appropriate for associative networks as we know from a lot of research in neuroscience where processes like these are modeled in continuous time (e.g., hemodynamic response function).

Third, it is plausible that the increase in capacity cannot go to infinity as implicitly follows from our assumptions. Realistically, it probably reaches some limit, at which point the connection becomes "permanent." When we learn to ride a bike, it takes some effort first. But, after enough repetitions, riding a bike becomes automatic. We do not think about it ever again, and moreover, the skill *never gets forgotten*. It is possible that cerebellum plays a role in this. For example, it might be that whenever an associative link reaches the maximum capacity, it gets taken over by the cerebellum that maintains it indefinitely after that and the link never decays.

Another assumption we make implicitly is that the rate of increase of capacity is the same in the whole associative network. This might also not be true. It may be that different features coming from different cortexes like visual, auditory, etc. have different ε for some biological or physiological reasons.

Finally, we assume that the speed of spread of the signal on the network is constant. This might also not be true. Given that signals are noisy in any physical system, of which human brain is an example, it is plausible that strong signals spread *faster* than weak signals. This is simply due to the problems with noise. So, it may be that signals with high relevance spread quickly, while signals with low relevance spread slower. This can lead to differences in how fast some ideas or concepts come to our minds.

E Alternative Assumptions on Memory Maker

When we discuss memory maker we assume for simplicity that new memories are being created constantly in each discrete time period. This is, of course, not true as we well know from quotidian experiences. For example, we rarely remember what exactly we saw outside when taking a bus. So, the memory maker somehow chooses when to record a new memory. This most likely happens when there is something *new* to record or when associated values are very high in absolute value. When we take the same bus route for a long time, we see the same landscape each day and we are mostly in the same mood. It is common sense that there is no reason to make multiple memory features of the same thing over and over again. We leave it for the future research to figure out how this process exactly happens.

It might also be the case that the memory maker does not record all currently active features for similar reasons. When we see something interesting happening, we may only record the important features and skip the multiplicity of small irrelevant details.

Finally, we believe that the value that the memory maker assigns to the memory feature is definitely not only the mood coming from the value aggregator. Most likely it is the sum of the mood, the derivative feature, the attitude (mood in the language handler), and the derivative attitude feature. After all, we can recall how painful it was when we fell from a bike (the derivative feature). The modification is simple: we need to assume that the value of the newly recorded memory feature is the sum of the four values mentioned above.

F Affective and Cognitive Languages

In the section on language, we concluded that affective language should look like a stream of random words that get pronounced as the signal spreads over associated features in Talking Molly's network. We believe that *cognitive language* is a cognitive upgrade of affective language. Specifically in the sections on cognition, we mentioned that Alice can memorize *rules* in her episodic memory and this is what allows her to live in a society. It is not hard to imagine that for the sake of better communication a group of Alices can invent rules for their originally affective language. The rules of how to speak (like English grammar) can make information transmission faster and increase fidelity. So, it is reasonable to think that language rules evolve among cognitive agents thus turning an affective language into a cognitive one.

It is also not hard to think of languages that are more "cognitive" than others. For example, the rules of English grammar are very simple, logical, and clear, whereas the rules of Russian grammar consist mostly of exceptions and from a certain perspective are not rules at all. So, we can think that English is more cognitive than Russian.

G Updater

In the section about Tommy and the valuation system we assumed that the updater makes the updates of values using the standard reinforcement learning formula $v \leftarrow v + \lambda(M - v)$. For expositional purposes, we did not specify how exactly this updated value is computed. We do this in this section.

We use the idea of the comparator presented in Section 6.1.2. In fact, we do not obtain exactly the formula above, but rather its generalization. Specifically, we show how the parameter λ can be *endogenized* and discuss different possibilities that arise in our model and that might actually make more sense than the traditional formula (in the context of biological organisms with associative networks).

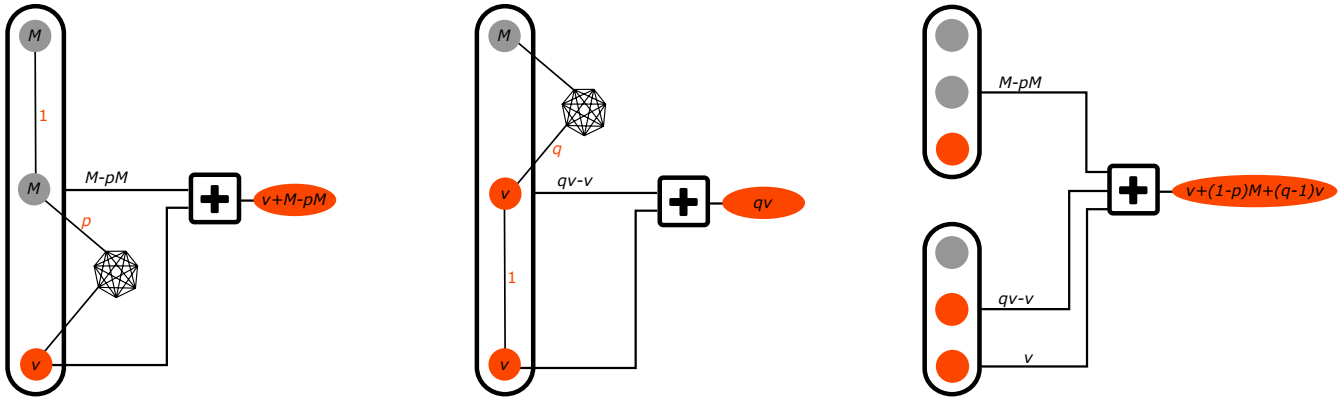

Figure 16: **Left Panel.** Simple updater that uses own mood perspective. **Right Panel.** Alternative simple updater that uses the feature perspective. **Right Panel.** A complex updater that uses both perspectives.

In section on choice, we proposed that the comparator, that is also a part of the updater, is used to decide which option is better. The left panel of Figure 16 shows how it works. Suppose Alice wants to understand whether or not she wants some feature when she is in mood M (we called this one-feature choice). To decide, Alice compares her mood and the feature from the perspective of her mood. She plugs the ingredients into the comparator and if $M - pM > 0$ she decides that she does not want the feature. The logic of this is that the feature associates weakly with her mood ($p < 1$) and this is why she does not want it. When $p > 1$, Alice does want the feature because now it associates a lot with her mood ($M - pM < 0$).

This is how the comparator is used for choice in the cognitive system. When it is used in the updater, it performs a different function. The updater's purpose is to make sure that the values of features present in the environment become closer to the current mood M, since M seems to convey information about what environment is like at the moment. So, feature values are being made consistent with that information. The comparator computes the difference between the relevant values of mood, which is M, and the relevant value of mood when the feature is activated, which gives $M - pM$. So, when $M - pM > 0$, this means that the feature is not

desirable (see above) at current mood. Thus, its value should be increased by the amount of that undesirability ($M - pM$). This is so because the value of the feature should be made more consistent with what Alice feels, namely M. The opposite holds when $M - pM < 0$.

We propose that a simple updater adds the value $M - pM$ to v in order for this value to move in the direction of the current mood. So, as shown on the left panel of Figure 16, the updater generates $v + M - pM$ and records this as the new value of the feature.

Alternatively, the updater could do the same but from the perspective of the feature itself. This is shown on the middle panel of Figure 16. Here, the comparator with the perspective of the feature returns $qv - v$. This value is positive when $q > 1$ or when current mood associates with the feature a lot, which implies that its value should be increased. Comparator does that by adding $v + qv - v = qv > v$. When $q < 1$, the feature weakly associates with the current mood, thus its value should be decreased. Indeed we have $qv < v$.

Philosophically, we cannot make a choice between the two ways updater might work. It can update values from two different perspectives, both of which seem to make sense. Thus, we propose the third version, a complex updater that takes both perspectives at once (the right panel of the figure). In this case, the comparator is used twice with different perspectives and then both updates are added to v. As a result we get the updated value $v + (1 - p)M + (q - 1)v$.

None of this seems like anything having to do with the reinforcement learning update that is $v + \lambda(M - v)$. However, in a special case when the mood feature and the feature being updated are connected *directly by one link* we have $p = q$ and the formula from the complex updater becomes $v + (1 - p)(M - v)$. Now, this does look like the reinforcement learning formula as long as $p < 1$.

These ideas suggest that the intuition of reinforcement learning does work for a special case in our framework, but that involves a complex updater and a weak link between the feature and the mood. We believe that our treatment of the updater is more general and provides many opportunities for testing experimentally which exact updater is used in reality. We leave it for future research to determine that.

H Controlling Associative Network with Cognition

The ability to focus and concentrate gives Alice a unique opportunity to control and change her own associative network. In fact, it is possible that many people use this ability often without realizing it (though not always to good ends). The key to this ability lies in the possibility to focus on a feature for a long enough time. Since the value aggregator and the language handler are always on, they update values of features continuously, and this can be used by Alice to create values of features that she desires.

For example, suppose that Alice became friends with someone who likes techno music, say Bob. She really likes Bob and she wants to spend time with him, but Bob always goes to rave parties and the sound of techno makes Alice scared and sick. So, Alice decides to fix that by *learning to like* techno music. She can do the following. She can play techno in her headphones in a comfortable, calm environment and focus on the sounds, while thinking about something really pleasant. She can imagine how much fun she could have with Bob at a techno party if only the music did not annoy her. Or she can imagine how cool everyone will think she is if she could join the techno-loving crowd of people.

Theoretically, it works in the following way. Alice focuses on three features: Techno (negative value), Being Cool (positive value), and her good mood (a set of features coming from the calm environment). While she is focused, the aggregated sum of relevant values is overall positive: features from the environment and Being Cool overcome the negativity of Techno. Thus, the value aggregator will automatically increase the value of Techno a bit, making it less negative. If Alice keeps training like that, she can gradually increase the value of Techno to zero, and then even make it positive.

Similarly, Alice can *learn to stop liking* something. For example, Alice always liked to drink fizzy drinks with sugar. But then, she has read in an article that refined sugar is very unhealthy. So, Alice decided to change her habits. Each time she has her favorite drink, she focuses on the horrible consequences of having diabetes (e.g., losing her feet) and what her life would be like in such case. The values of these thoughts overcome the positivity of the value of the drink and the value aggregator rewrites the value of the drink. As a result, Alice starts liking it less. Continuing in this fashion, Alice can reach a point where she does not want to drink fizzy sugary drinks anymore.

These *cognitive* techniques are somewhat different from the *affective* techniques that are used to teach children. In affective techniques, real incentives are used to achieve the desired outcome (you punish a child for drinking a fizzy drink, say). In this case, an *association* is created in child's mind between Fizzy Drink and Punishment. So, like Pavlov's dog, the child might learn not to drink fizzy drinks because it reminds him of the punishment. However, this is different from what Alice was doing, because the child has never changed the value he attaches to the fizzy drink. Whenever punishment disappears, he will gradually start to drink fizzy drinks

again, because he still likes them (high positive value), and only the association with punishment stopped him from doing it.

The last kind of cognitive technique is *devaluing features* by suppressing mood and attitude with Concentration. For example, if Alice has PTSD (she has very bad memory of something from the past with very negative value) she can Focus on the bad memory while concentrating in a way that suppresses her mood and attitude features. In such state, the value of her mood will be zero and the value aggregator will increase the value of the negative memory making it less negative than before. Exercising in this manner many times Alice can gradually decrease the negativity of the memory and fix her PTSD problem.

I Topology Definitions and Continuity of Preferences

In this section, we provide the specific definitions of topology on $\mathcal{C}$ and show that Robbie's preferences—defined by the updates—are continuous in the space of contexts (see Section 7.16).

Consider the set of all contexts $\mathcal{C}$ and define an ε-ball around context $C = \{(k, p_k)_{k\in\mathcal{F}}\}$ as all contexts $A = \{(k, a_k)_{k\in\mathcal{F}}\}$ that are similar to C in the sense of S' to the degree of at least $1 - \varepsilon$. In other words, ε-ball around C includes all $A \in \mathcal{C}$ for which

$$S'(A, C) = \frac{\sum_k \min\{p_k, a_k\}}{\sum_k \max\{p_k, a_k\}} > 1 - \varepsilon.$$

Notice that the largest ball around C is of size 1 obtained by setting $\varepsilon > 1$. It includes all contexts in $\mathcal{C}$ (the contexts disjoint from C are contained only in this 1-ball).

Now, consider the collection of all ε-balls around all contexts $C \in \mathcal{C}$. This can be treated as the base of topology on $\mathcal{C}$. Thus, we can say that S'-topology on $\mathcal{C}$ is generated by the base consisting of all ε-balls around all contexts.

Given this definition we are interested in showing that the function $f : \mathcal{C} \to \mathbb{R}$ defined as $f(A) = S(A, C)$ for some fixed C is continuous in the above topology. To do that, consider first the meaning of the condition that A lies in the ε-ball around C. Suppose that $a_k = p_k + \varepsilon_k$. We just rewrite the relevances a_k of features in A as deviations from relevances p_k in C. Then the inequality above can be rewritten as

$$\varepsilon|C| > (1 - \varepsilon) \sum_{k\in K_1} \varepsilon_k - \sum_{k\in K_2} \varepsilon_k.$$

Here, K_1 is the set of features for which $\varepsilon_k > 0$ and the set K_2 contains features with $\varepsilon_k \leq 0$. Given these conditions, we can easily derive that

$$\forall k \in \mathcal{F} \quad |\varepsilon_k| < \frac{\varepsilon}{1 - \varepsilon}|C|.$$

This is the condition on individual relevances that should be satisfied for A to lie in the ε-ball around C.

To show that f is continuous, consider some ε-ball around context $B = \{(k, b_k)_{k\in\mathcal{F}}\}$. We want to show that the image of this ball through f converges to point $f(B)$ as $\varepsilon \to 0$. Let us take some context $A = \{(k, a_k)_{k\in\mathcal{F}}\}$ in the ε-ball around B with $a_k = b_k + \varepsilon_k$, and consider the value $f(A)$:

$$f(A) = S(A, C) = \frac{\sum_k \min\{p_k, b_k + \varepsilon_k\}}{\sum_k b_k + \varepsilon_k}.$$

Given that from the above we know that

$$\forall k \in \mathcal{F} \quad |\varepsilon_k| < \frac{\varepsilon}{1-\varepsilon}|B|,$$

it is clear that as $\varepsilon \to 0$, we have $|\varepsilon_k| \to 0$. Thus, we have

$$f(A) \to \frac{\sum_k \min\{p_k, b_k\}}{\sum_k b_k} = S(B, C)$$

as $\varepsilon \to 0$. This shows that for any open set in $\mathbb{R}$ we can always find an open set in $\mathcal{C}$ that maps through f inside it proving that f is continuous. The same holds for any function $f(A) = hS(A, C)$ for some real number h (multiplying continuous function by a constant keeps it continuous). We formulate this as a proposition.

Proposition. Individual updates $f(A) = hS(A, C)$ are continuous in S'-topology on $\mathcal{C}$.

J Variables in the Reduced-Form Model

Variable	Definition
$\mathcal{F}$	The set of all conceivable features
$\mathcal{C}$	The set of all contexts
$\mathcal{A}$	The set of all action features
C'	Real context observable in reality
C	Mind context induced by C'
β	The relevance threshold for the action feature to be performed
$S(A, C)$	Similarity measure on $\mathcal{C}$ to perform system updates
$S'(A, C)$	Symmetric similarity measure that induces S'-topology on $\mathcal{C}$
v_C	Affective value of context C
f_C	Familiarity value of context C (also affective)
w_C	Cognitive value of context C
χ	Cost of complete unfamiliarity in utility
α_C	Context-specific coefficient multiplying cognitive value in utility
α	Global coefficient multiplying cognitive value in utility
$u(D\|C)$	Imagined utility of D in C
$b_t(C_k\|C)$	Belief of Tommy
$b_m(C_k\|C)$	Belief of Molly
$b_r(C_k\|C)$	Belief of Robin
$b(C_k\|C)$	Aggregated belief
$\kappa(C_k\|C)$	Cost of transitioning from context C to C_k
ζ	Cost of uncertainty before cognition
$E(D\|C)$	Expected imagined utility of context D when in C
C_ϕ	"Do nothing else in C" context
C_θ	"Think in C" context
$U_t(C_k\|C)$	Level-t expected utility
$U(C_k\|C)$	Expected utility of the action leading to C_k
λ	Reinforcement learning parameter
δ	Discount factor for computing expected utility on the knowledge tree
s_C	Context-specific cost of thinking
s	Global cost of thinking
ΔU	Change in utility after thinking used to update C_θ
ϵ_f	The size of update of familiarity when experienced live
ϵ_s	The size of update of s
ϵ_{sc}	The size of update of s_C
ϵ_α	The size of update of α
$\epsilon_{\alpha c}$	The size of update of α_C
ϵ_1	The size of the ball around C where contexts are considered the same
$\epsilon_2(\iota, s_C + s)$	The size of the ball in which Robbie can search for similar contexts
$\hat{v}_C$	Value of v_C estimated from data
$\hat{f}_C$	Value of f_C estimated from data
ι	The imaginativeness parameter
$\tau(i\|C)$	Perceived social weight, or status, of agent i in C

Table 1: Variables used in the reduced-form model.