Received: 20 October 2020 | Revised: 8 February 2021 | Accepted: 17 June 2021

DOI: 10.1111/jtsb.12311

ORIGINAL ARTICLE

WILEY

The psychology of ultimate values: A computational perspective

Francesco Rigoli

City, University of London, London, UK

Correspondence
Francesco Rigoli, Department of Psychology, City, University of London, Northampton Square, London EC1V 0HB, UK.
Email: francesco.rigoli@city.ac.uk

Abstract

Ultimate values can be defined as abstract rules or goals transcending specific contexts and defining the utmost purposes of existence. Although the literature about human values is vast, several fundamental questions about ultimate values remain open. What are the processes responsible for the formation of ultimate values? What is the impact of inbuilt affective processes and of learning, respectively? Regarding learning, what is the role of society? Empirical evidence shows dramatic variability in ultimate values pursued by different people. Why? These open questions suggest that a precise picture of ultimate values is lacking. This paper offers a computational theory of ultimate values. The key idea is that our brain represents values along a hierarchy where ultimate values are built upon experiences with inherent affective nature (basic values). Based on these representations, the proposal is that the brain infers the rules that foster basic values in a variety of contexts. These would become ultimate values and drive human behaviour independent of the ongoing context. We discuss how the theory can contribute to understanding a variety of aspects of human values, including morality, to what degree values are innate or culturally determined, and how values shape and are shaped by society.

KEYWORDS
Bayesian, computational modelling, ethics, goal, ultimate value

1 | INTRODUCTION

When comparing different times and places, it is striking how behaviour of different people can be driven by such diverse motives. Alexander the Great put all his energies in extending his rule to new territories. For most of his life, Anthony of Egypt voluntarily self-isolated in the wilderness of the Egyptian desert to be closer to God. The Greek thinker Socrates was deeply dedicated to philosophical discussions as a way to unveil the truth. Martial ethos, religious zeal, and pursuit of knowledge are three examples of ultimate values that guide human behaviour. These can be referred to as ultimate because they embody general and abstract rules or goals, which transcend specific contexts and define the utmost purpose of existence. As an example of how ultimate values work, consider Alexander the Great. Putatively, his day-to-day behaviour was often guided by goals such as eating, drinking, or sacrificing to gods. However, often (though not always) we can imagine that, in the king's mind, these were not ends in themselves, but they were instrumental to the fulfilment of the grand purpose of conquering new territory: food and drink were consumed to be fit for battle, and gods were prayed to win their favour for territory expansion. In this example, conquering land represents the ultimate value for Alexander the Great, because it subordinates all other values. It prescribes which subordinate values should be pursued in the different contexts (e.g., in the temple, pray the gods to win their favour). In other words, a hierarchy is implicated here (Carver & Scheier, 1998; Pezzulo et al., 2018; Powers, 1973; Talevich et al., 2017; Wicker et al., 1984), where ultimate values occupy higher levels and subordinate values occupy lower levels, and where subordinate values are constrained by ultimate values and by the current context. Although Alexander the Great, Anthony of Egypt, and Socrates are extreme examples of dedication to ultimate values, and most people are possibly less committed to them, yet ultimate values are arguably fundamental also for many individuals in their personal, social, and political life (Weber, 1904/2002).

What does research tell us about the nature of ultimate values? The literature about human values, motives, and goals is vast (e.g., Austin & Vancouver, 1996; Carver & Scheier, 1998; Fiske, 2004; Haidt, 2001; Inglehart, 1997; Laidlaw, 2002; Maslow, 1943; Nichols, 2004; Robbins, 2012; Rokeach, 1973; Schwartz, 2006, 2012); within this literature, some have highlighted the difference between abstract and concrete values (Carver & Scheier, 1998; Powers, 1973; Talevich et al., 2017; Wicker et al., 1984), which is critical to understand ultimate values. However, several fundamental questions about the nature of ultimate values remain open. What are the processes responsible for the formation of ultimate values? What is the impact of inbuilt affective processes and of learning, respectively? And regarding learning, what is the role of society? As the examples of Alexander the Great, Anthony of Egypt, and Socrates as well empirical evidence show (Rokeach, 1973; Schwartz, 2006, 2012), there is a dramatic variability in ultimate values pursued by different people. Why? These open questions suggest that a precise picture of the nature of ultimate values is lacking. The aim of this paper is to offer a theoretical framework to explore ultimate values, their development, the impact on motivation, and their relationship with subordinate values. The paper is structured as follows. The next section introduces the main tenets of the theory. To offer a precise account, this is followed by a section where the theory is presented in the form of a computational model. Next, fundamental aspects of the theory are examined in detail. The paper concludes with a section where more general issues are discussed.

2 | THEORY

According to our theory, the human brain represents different types of values along a hierarchy (Carver & Scheier, 1998; Pezzulo et al., 2018; Powers, 1973; Talevich et al., 2017; Wicker et al., 1984). The bottom level of the hierarchy reflects what we refer to as *basic values*, corresponding to stimuli or experiences with an in-built affective quality. These comprise basic physiological needs such as drinking, eating, having sex, as well as more sophisticated (but yet inbuilt) social experiences such as sense of protection, social esteem, or social bonding. In addition to these examples reflecting rewarding outcomes, other cases are characterised by punishment, such as when experiencing pain, threat, social despise, or abandonment. Moreover, basic values can encompass experiences relevant for self-interest (e.g., perception of pain) as well as outcomes relevant for the interest of others (e.g., perception of another individual experiencing pain), consistent with empirical evidence highlighting the importance of altruistic considerations in shaping human motivation and behaviour (Stich et al., 2010). Altogether, our notion of basic value is similar to the concept of unconditioned stimulus in associative learning literature (Pavlov, 1927). Both describe outcomes able to elicit an in-built motivational tendency. This ability is viewed as largely innate (i.e., present independent of any learning), stable (i.e., scarcely shaped by ontogenetic factors), and universal (i.e., characteristic of the whole specie, and hence present in virtually all healthy individuals of that specie – although some degree of inter-individual variability is conceivable also at this level).

The second level of the hierarchy describes what we refer to as *contextual values*, representing stimuli or actions that initially do not have any inherent value, but that acquire it in virtue of their ability to predict basic values in certain contexts. Again, this is inspired by associative learning literature, and precisely by the concept of conditioned (or secondary) reinforcer (Bell & McDevitt, 2014; Rescorla, 1980; Skinner, 1953). Empirical evidence indicates that, if an initially neutral stimulus such as a token is associated with an unconditioned reinforcer such as food, then animals' actions resulting in the token will be reinforced, even if these actions never lead to food (Bell & McDevitt, 2014). This indicates that the token has become a conditioned reinforcer and has acquired motivational value in it of itself (Rescorla, 1980). The notion of conditioned reinforcer is fundamental to understand value. It highlights the enormous learning abilities of the human brain, which is capable to go beyond evolutionary pre-established incentives and develop purposes specific to one own's life experience. The notion of conditioned reinforcer is fundamental also for another reason. It emphasises the tendency of our brain to transform experiences that are initially means (i.e., conditions useful to get some other goals) to ends, imbued with value as such.

Following the concept of conditioned reinforcer, our theory proposes that contextual values develop thanks to their ability to predict basic values at the level below. For example, within our model the token and food correspond to the contextual value and the basic value, respectively. However, our notion of contextual value, more than the notion of conditioned reinforcer, emphasises the role of context (with analogies to literature about occasion setting; Schmajuk & Holland, 1998). This is because our model proposes that a contextual value (e.g., a token) is such that it predicts a basic value (e.g., food) only in specific circumstances or contexts (e.g., in the experimental chamber) and not in others (e.g., outside the chamber). The role of context is critical in our formulation to distinguish contextual values from ultimate values, the latter being independent of context and hence more abstract. Ultimate values occupy the third and top level of the hierarchy. These are abstract goals or rules which, if pursued or applied across contexts, foster experience of contextual values and in turn of basic values. Consider again the example of Alexander the Great. His ultimate value, which guides Alexander's behaviour in a variety of contexts, can be imagined to be conquering new territory. In a specific context, for example in a newly conquered city, relying on such ultimate value implicates an appropriate contextual value such as parading on the city's agora. This in turn implies experiencing basic values such as perceiving admiration from the own soldiers and from inhabitants of the conquered city. In a different context, for example in the battlefield, the same ultimate value implicates a different contextual value such as charging the enemy, implying basic values such as experiencing admiration from the own soldiers and observing the enemy soldiers manifesting fear.

How are ultimate values built? Our proposal is that these result from the brain's abstraction abilities, capable of constructing hypotheses about which rules or conditions are conducive of contextual values (and eventually basic values) across a variety of contexts (Little & McDaniel, 2015). Within a set of potential ultimate values, the brain would select the one which best realises basic values; hence, in our proposal, an ultimate value is built upon basic values, namely upon hard-wired affective experiences. However, once selected, such ultimate value (and note basic values) would actually drive human behaviour. Again, this stresses the brain's propensity to transform means into ends: initially, an ultimate value is selected for its ability to foster basic values, but, once established, behaviour would aim at realising precisely such ultimate value, and not the implicit basic values. For example, Alexander the Great would initially select the ultimate value of conquering new territory because this appears to him as the best in terms of satisfying basic values. However, once established, conquering new territory would be pursued as an end in itself.

Note that, at first glance, our notion of ultimate value appears as similar to the notion of generalised conditioned reinforcer (Skinner, 1953; Tan & Hackenberg, 2015), indicating a conditioned stimulus associated with multiple kinds of unconditioned stimuli (e.g., both water and food, and not food alone). However, the two notions are different: an ultimate value is associated with unconditioned reinforcers (adopting associative learning terminology) across multiple contexts, but unconditioned reinforcers can be of the same kind. Conversely, a generalized conditioned reinforcer is associated with different kinds of unconditioned reinforcers, but this association can be at play only in a single context.

In short, our theory proposes that values are arranged along a hierarchy with ultimate, contextual, and basic values occupying the top, middle, and bottom level of the hierarchy. Basic values would correspond to experiences with a hard-wired affective quality, contextual values to conditions predicting basic values in specific contexts, and ultimate values to general rules predicting contextual values (and in turn basic values) independent of context. To further clarify the theory, the next section casts it in the form of a mathematical model.

3 | COMPUTATIONAL MODEL

The ancient historian Plutarch reports a famous anecdote about the meeting between Alexander the Great and the cynic philosopher Diogenes, who had opted for a life of poverty, degradation, and philosophy (Plutarch, 2004). Despite the stark distance between the two men in their way of life, Alexander displayed a surprising admiration for Diogenes so much so that he said: "were I not Alexander, I would be Diogenes". Whether true or not, for our purposes this anecdote suggests that humans can contemplate a variety of ultimate values, and eventually select one to be pursued. Let us imagine that, in the back of his mind, Alexander considered three alternative ultimate values: conquest, philosophy, and conducting a luxurious life at the Macedonian court (this latter case might not be far-fetched too, given Alexander's reported proneness to alcohol abuse). We will use this example to illustrate how our theory of ultimate values can be implemented adopting computational modelling.

The model proposed corresponds to a Bayesian inference framework implemented adopting the formalism of Bayesian networks (Bishop, 2006; Rigoli, 2021a, 2021b). The network is described graphically in Figure 1. The circles represent categorical variables (each associated with a set of alternative categories), reflecting beliefs entertained by an agent (e.g., by Alexander), and arrows describe beliefs about probabilistic dependencies among these variables. The variable at the top describes Ultimate Values (UV), and includes three categories: conquest, philosophy, and luxury; these are the three options available to Alexander regarding the ultimate purpose of life. Each category is associated with a probability P(UV), which describes how attractive an ultimate value is a priori. The second variable is Context (C), reflecting different conditions one expects to face in life. In our example, three of such conditions are included: being at the royal palace, being in the battlefield, and being in the agora. Each condition is associated with a probability P(C), describing how probable that condition is to occur during life. The variables UV and C project to the third variable Contextual Values (CV), reflecting the actions or outcomes to be pursued in a given context and following a given ultimate value. In other words, CV depends on both UV and C (formally, this dependency is described by the conditional probability P(CV | UV, C)). For example, if UV corresponds to conquest and C corresponds to royal palace, CV will correspond to "motivate nobles" to follow me in battle. As another example, if UV corresponds to philosophy and C corresponds to agora, CV will be "speak to philosophers". The CV associated with different combinations of C and UV is reported in Table 1 (in our example, CV can assume seven different categories). Note that, for the sake of simplicity, in our example we assume that contingencies (described by the

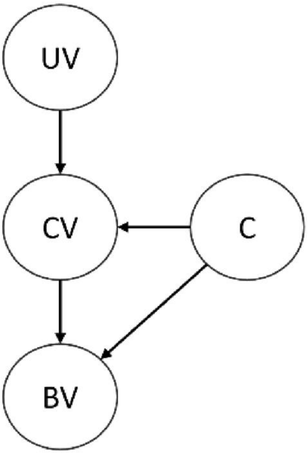

FIGURE 1 Illustration of the Bayesian network describing the beliefs of an agent about value. Circles and arrows represent variables and probabilistic dependencies, respectively. The variables are: Ultimate Values (UV), Contextual values (CV), Context (C), and Basic values (BV)

TABLE 1 Contextual value (CV) resulting from different combinations of Context (C) and Ultimate Value (UV) in the example of Alexander the Great (see main text)

C	UV	CV
battlefield	conquest	charge enemy
battlefield	philosophy	fly away
battlefield	luxury	fly away
palace	conquest	motivate nobles
palace	philosophy	speak to philosophers
palace	luxury	drink alcohol
agora	conquest	parade
agora	philosophy	speak to philosophers
agora	luxury	court women

Notes: Formally, this indicates the category of CV with conditional probability of one for each combination of C an UV (e.g., P (CV = charge enemy | C = battlefield, UV = conquest) = 1).

conditional probability P(CV | C , UV)) are deterministic: for example, when UV = "conquest" and C = "royal palace", the contextual value will always be CV = "motivate nobles". More generally, because the model is probabilistic, for each combination of C and UV one could specify a probability attached to each category of CV.

Finally, the model includes the variable Basic Values (BV), capturing experiences with an inbuilt affective quality. For the sake of simplicity, this is represented by a dichotomous variable where one category corresponds to reward (reflecting positive experiences) and another to punishment (reflecting negative experiences) (Solway & Botvinick, 2012). This notion of reward and punishment summarises all different experiences with an inherent affective nature such as food, pain, social admiration/despise etc. Both CV and C affect BV (as described by the conditional probability P(BV | C , CV)). The probability of obtaining reward given a specific C and a specific CV (i.e., (P(BV = reward | C, CV)) indicates how good in terms of basic values a CV is in a given context C. For example, imagine that, for Alexander the Great, P(BV = reward | C = battlefield, CV = charge enemy) = 0.9 and that P(BV = reward | C = battlefield, CV = fly away) = 0.2. This indicates that, in the battlefield, charging the enemy is expected to be highly rewarding (e.g., because of basic values such as admiration from other soldiers), while flying away is expected to be highly punishing (e.g., because of basic values such as shame). In our example, the probability of obtaining reward for different CV and C is reported in Table 2. In summary, the joint probability of the variables in the network can be expressed as follows:

$$P(UV, C, CV, BV) = P(UV)\ P(C)\ P(CV|UV, C)\ P(BV|CV, C) \tag{1}$$

According to our theory, the Bayesian network can be adopted to make two types of inference; these inferences are critical to establish the role of ultimate values in motivation and behaviour. The first inference establishes which ultimate value among those available should be pursued. Formally, this inference calculates the posterior probability of UV given observation of reward (i.e., P(UV | BV = reward) (Solway & Botvinick, 2012). Intuitively, this inference asks: if

TABLE 2 Conditional probability of obtaining reward as basic value (BV) for different combinations of Context (C) and Contextual Value (CV) (P(BV = reward | C, CV) in the example of Alexander the Great (see main text)

C	CV	P(BV = reward | C, CV)
battlefield	charge enemy	0.9
battlefield	fly away	0.2
battlefield	motivate nobles	0
battlefield	speak to philosophers	0
battlefield	drink alcohol	0
battlefield	court women	0
battlefield	parade	0
palace	charge enemy	0
palace	fly away	0
palace	motivate nobles	0.4
palace	speak to philosophers	0.4
palace	drink alcohol	0.6
palace	court women	0
palace	parade	0
agora	charge enemy	0
agora	fly away	0
agora	motivate nobles	0
agora	speak to philosophers	0.5
agora	drink alcohol	0
agora	court women	0.3
agora	parade	0.9

I want to obtain reward (and consider this as given), what is the most appropriate ultimate value? The answer to this question corresponds to a probability distribution (i.e., P(UV | BV = reward) where each category of UV is associated with a posterior probability; the higher the probability, the better the category for obtaining reward. Based on this, the ultimate value with the highest posterior probability is selected, corresponding to UV_*:

$$UV_* = \underset{i}{\operatorname{argmax}}(\mathrm{P}(\mathrm{UV} = \mathrm{i} \mid \mathrm{BV} \ = \ \mathrm{reward})\ ) \tag{2}$$

This inference captures the idea that an ultimate value is selected because it is considered the best for obtaining basic values. Once an ultimate value is selected, our proposal is that it drives behaviour across all contexts. In our example, Alexander attributes 0.7, 0.1, and 0.2 as posterior probability to conquest, philosophy, and luxury, respectively; hence he selects conquest as ultimate value UV_*. Consequently, Alexander will pursue conquest in all contexts, including the battlefield, the palace, and the agora. Applying this ultimate value will generally be the best option in terms of basic values, but not in all contexts: in this example, conducting a luxury life appears as being better in the palace (because it prescribes drinking alcohol, which is the best CV in terms of P(BV = reward | C , CV; see Table 2). Yet, Alexander will pursue conquest also in the palace; this is because, according to our theory, the selected ultimate value UV_* always guides behaviour, even in contexts where it is not the best (but see below how the model can be augmented to implement an influence of the ongoing context).

An interesting aspect emerging from the inference just described is that a measure of uncertainty about the selected ultimate value UV_* can be derived, corresponding to $E_{UV|BV}$, namely the entropy of the posterior distribution P(UV | BV = reward):

$$E_{UV|BV} = -\sum_i P(UV = i|BV = reward)\ \log(P(UV = i|BV = reward)) \tag{3}$$

$E_{UV|BV}$ is minimal when P(UV = UV_* | BV = reward) = 1 (implying that P(UV | BV = reward) = 0 for other potential ultimate values), and it is maximal when P(UV | BV = reward) is equal for all potential ultimate values. Hence, $E_{UV|BV}$ measures the uncertainty about the selected ultimate value UV_*: the higher the $E_{UV|BV}$, the higher the uncertainty. High uncertainty means that, although an ultimate value UV_* is selected, this is not so much better than alternative ultimate values. We suggest that this uncertainty has implications at the motivational level, influencing the level of vigour expressed in pursuing the selected ultimate value: lower uncertainty would result in people having higher vigour in pursuing the selected ultimate value (Alexander the Great, Anthony of Egypt, and Socrates are arguably examples of such low uncertainty and high vigour).

Once an ultimate value UV_* is selected, the Bayesian network is adopted to make another type of inference, namely to infer the posterior probability of BV = reward given the selected ultimate value UV_* (i.e., P(BV = reward | UV = UV_*). This estimates how good the selected ultimate value UV_* is in terms of obtaining basic values. If we compare the two types of inferences we propose, inferring P(UV | BV = reward) is a way to compare potential ultimate values against one another, eventually selecting one; while inferring P(BV = reward | UV = UV_*) is a way to assess the selected ultimate value UV_* in isolation in terms of its ability to foster basic values. The latter inference quantifies how much a selected ultimate value UV_* is grounded on basic values: when P(BV = reward | UV = UV_*) approaches one, the selected ultimate value UV_* is strongly grounded, while when P(BV = reward | UV = UV_*) approaches zero it is poorly grounded. We propose that this inference determines the mood associated with pursuing the selected ultimate value. Consider religious individuals for whom the purpose of life is suffering for expiating sins. These individuals might pursue this ultimate value vigorously because they consider any alternative life purpose (e.g., living an hedonistic life) much worse in terms of basic values (e.g., they expect an hedonistic life to lead to eternal suffering in hell or to being despised by others). Yet, the mood ensuing from their ultimate value will still be rather negative, because their ultimate value is conducive of suffering and hence of poor basic values. In our model, such mood is captured by inferring P(BV = reward | UV = UV_*).

In sum, our model proposes that the human brain represents different types of value organised along a hierarchy (including basic, contextual, and ultimate values), together with their probabilistic relationship. These representations are relied upon when making two forms of inference, the first one selecting the ultimate value to be pursued (and its associated uncertainty), the second one determining the ensuing mood. Note that these inferences are proposed to occur subconsciously: in other words, phenomenologically, they would simply result in the desire to pursue the selected ultimate value and in experiencing the ensuing mood, without awareness that the ultimate value is in fact grounded upon basic values. While these are the core tenets of our proposal, below we examine some further important aspects.

4 | IMPORTANT ASPECTS OF THE THEORY

Here, we examine fundamental aspects of the theory relevant for explaining the role of value in general, and of ultimate values in particular, in human motivation and behaviour.

4.1 | Why ultimate values?

Why, in a functional perspective, should our brain identify and pursue ultimate values, rather than simply pursuing basic or contextual values? We speculate that relying on ultimate values might have evolved as a strategy for simplifying the choice problem. Choices are overwhelmingly complex in ecological scenarios, because choosing optimally requires integrating a vast amount of contextual information (Sutton & Barto, 1998). One way to simplify choice is to identify general states or rules that lead to good outcomes across a variety of contexts, and simply pursue these and ignore the context. This allows the brain to ignore the details of the specific circumstances, simplifying the choice problem considerably. For example, simply relying on the general rule "do not lie" allows one to ignore the specific context at play, hence facilitating choice. Similarly, pursuing military conquest without assessing whether this is advantageous in each specific context, also simplifies the choice problem. Note that, if the general rule selected is most of the time effective, then the price for following the rule and ignore the context is not high, and it does not compensate the computational cost avoided. Essentially, our argument is that, in a functional perspective, the human brain has evolved to pursue ultimate values because this makes choices easier.

Why relying on ultimate values simplifies the choice problem can also be explained from another angle. Our theory implies that, to make choice, humans perform two basic forms of computations, occurring independently of one another (Pezzulo et al., 2018). These can be referred to as evaluation and planning, respectively (Pezzulo et al., 2018). During evaluation (which is the focus of this paper), the ultimate value to be pursued is identified (based on the processes described above). During planning (which is not examined in this paper), the better course of action for pursuing the selected ultimate value is identified. Breaking down choice in these two processes simplifies the problem substantially, given that the interaction among the two elements can be disregarded. In Bayesian statistics, this simplification is called factorisation (Bishop, 2006). Factorisation is appropriate if it reflects the true environmental contingencies (Friston & Buzsáki, 2016; Rigoli et al., 2017), in our case if it is true that (i) there are some general states (the ultimate values) which are good across contexts, implying that contextual information can be disregarded during planning, and that (ii) the ultimate values remain good independent of the planned sequence of actions, implying that information about actions can be disregarded during evaluation. Arguably, these two conditions apply to the real world, and the brain might have evolved to exploit them and separate evaluation from planning. The idea of factorisation in psychology is not new; for example, it has been proposed to explain the separation between *where* and *what* pathways in the brain (Friston & Buzsáki, 2016). Factorisation can be proposed to explain how the brain simplifies planning, which is the other process underlying choice together with evaluation: abstract actions can be decomposed in operative behaviour (Balaguer et al., 2016; Botvinick, 2012) and broken down in subgoals (Maisto et al., 2015). Here we advocate the notion of factorisation (i) to distinguish planning from evaluation (for a similar idea, see O'Reilly et al., 2014; Pezzulo et al., 2018), and (ii) to explain how the brain simplifies evaluation by selecting ultimate values.

In summary, our brain might have evolved to identify and pursue ultimate, and not basic or contextual, values because this strategy simplifies the choice problem. In a world where planning and evaluation can be performed independently with virtually no cost, selecting ultimate values might be the result of evaluation processes that can be next integrated with planning, hence simplifying the choice problem.

4.2 | How do ultimate values look like?

Our theory is flexible regarding the content of an ultimate value. A useful way to express an ultimate value is via an imperative sentence such as (in the example above about Alexander the Great) “conquer land”, “practice philosophy”, or “live a luxurious life”. Other examples are “reach the top of the social hierarchy” and “follow God's commandments”. Although these are rather abstract examples, more concrete imperatives can also become ultimate values, such as “seek money” (Lea & Webley, 2006), “get a PhD”, “become a professional football player”, or “have two children”. Also, sometimes ultimate values might take the form of negative statements such as “do not become poor”, “do not take drugs, or “do not break the law”, with an emphasis on states or actions to be avoided rather than to be pursued. Note that the term ultimate refers to the notion of abstraction (i.e., the quality of being independent of context; Powers, 1973) and not to the notion of time. Certain states or actions appear at the end of a behavioural chain: for example “eat the yogurt” appears at the end of the sequence “open the door”, “open the fridge”, “take the yogurt”, “eat the yogurt”. However, “eat the yogurt” is as abstract as all other actions; hence, according to our theory, it does not represent an ultimate value (also, based on our distinction above between planning and evaluation, temporal chains characterise planning and not evaluation).

Our view of ultimate values can capture processes such as morality, virtue, and well-being. Moral values consist in pursuing fairness, justice, or altruism during interactions with other people (Haidt, 2001; Nichols, 2004). Within our theory, these can be casted in terms of ultimate values. Examples of ultimate values with moral content are “be fair with other people”, “pursue equality among people”, “follow God's commandments”, “follow the law”. Notably, ultimate values can concern rules of conduct (e.g., “do not lie”) or states to be sought (e.g., “pursue equality among people”), encompassing both deontological and consequentialist approaches to morality, respectively (Crockett, 2013). Linked with morality, the notion of virtue describes desirable personality characteristics such as honesty, compassion, and generosity (Widlok, 2004). In our theory, virtues can also be straightforwardly cast in terms of ultimate values, specifically as imperatives aimed at acquiring or maintaining virtue such as “be honest”, “be compassionate”, and “be generous”. Finally, ultimate values might reflect rules or states conducive of well-being but not related with morality or virtue (Diener, 2009). Examples of these are “seek money”, “indulge in pleasures such as food and drink”, “cultivate friendship”.

In short, our theory of ultimate values aims at being general, potentially applying to a variety of domains including morality, virtue, and well-being.

4.3 | Mundane concerns and contextual effects

While ultimate values reflect abstract purposes such as in religion and ideology, human behaviour often appears as driven by what we can call mundane concerns, such as having a good meal. Sometimes mundane concerns might be subordinate to ultimate values (e.g., in the example of Alexander the Great, eating might sometimes have subserved the goal of being fit for battle), but other times they are valuable as such. What is the role of mundane values in our theory? Consider an individual addicted to alcohol whose ultimate value is to cultivate family relationships. In certain conditions such as during abstinence (an internal cue) or when exposed to alcohol-related cues (external cues), the individual is overwhelmed by an attraction towards alcohol, which endangers his ultimate value. This example highlights that, in our view, the fundamental difference between ultimate values (e.g., cultivating family relationships) and mundane values (e.g., alcohol) is that the former are independent of the cues one is exposed to, while the latter are cue-dependent. In other words, ultimate values would correspond to values to be pursued in ideal conditions, without any cue influence, while mundane values would be the product of cue influence. This possibility fits with empirical evidence showing that potentially dysfunctional impulses are usually steered by internal or external cues (e.g., Dawson & Kim, 2009).

We propose to capture the role of mundane values by adding to the Bayesian network described above the variable contextual cue (C_C), which depends on the context C (Figure 2). Considering again the example of Alexander the Great, C_C includes the same categories as C, namely battlefield, agora, and palace. The difference is that C_C describes the perceptual cues (which can be directly observed) of each context, while C indicates an abstract representation of

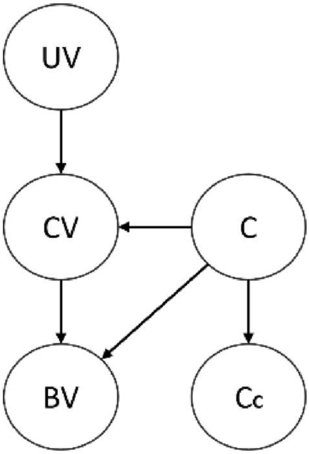

FIGURE 2 Description of the Bayesian network representing the beliefs of an agent about value, now including also the variable Contextual Cue (C_C). Circles and arrows represent variables and probabilistic dependencies, respectively. The variables are: Ultimate Values (UV), Contextual values (CV), Context (C), Basic values (BV), and Contextual Cue (C_C)

context, which is not directly observed (i.e., it is a latent variable). Now the joint probability of the network becomes:

$$P(UV,\ C, CV, BV, C_C) = P(UV)\ P(C)P(C_C|C)\ P(CV|UV, C)\ P(BV|CV, C) \tag{4}$$

We propose that sometimes the ultimate value is inferred in "neutral" environments, where no salient contextual cue is present. This consists in inferring P(UV | R = reward) and selecting UV_* as described by Equation (2). Other times, a contextual cue k is available; as described below, it is during such circumstances that mundane values might overwhelm ultimate values. When a contextual cue k is available, the posterior probability P(UV | R = reward, C_C = k) is estimated, and the selected ultimate value (now referred to as UV_C) now corresponds to:

$$UV_C = \underset{i}{\text{argmax}}(\text{P}(\text{UV} = \text{i}\ |\text{BV}\ =\ \text{reward},\ C_C\ =\ \text{k})) \tag{5}$$

To appreciate the implications of the contextual cue, let us consider two cases, characterised by strong and weak cue-influence, respectively (Table 3). Regarding strong cue-influence, this occurs when C_C activates specific context representations C. In our example (Table 3), cues related to the palace are particularly salient, as they strongly activate the representation of the palace context (formally, this is reflected by the fact that P(C_C = palace | C = palace) = 0.9, which is very high; Table 3). What happens when palace-related cues are experienced? The selected ultimate value UV_C now becomes luxury, while UV_* (i.e., the one estimated without any contextual cue) was conquest. The reason why now luxury is selected is because, when the palace cue is present, drinking alcohol (associated with luxury) is particularly rewarding. This example shows how, in our model, mundane values such as drinking alcohol can be triggered by cues, interfering with ultimate values that would be pursued without those cues (e.g.,

TABLE 3 Conditional probability of observing a specific Contextual Cue (C_C) given the Context C (C) (P (C_C | C)) in the example of Alexander the Great (see main text)

C	C_C	**Strong cue-influence** P(C_C | C)	**Weak cue-influence** P(C_C | C)
battlefield	battlefield	0.6	0.34
battlefield	palace	0.2	0.33
battlefield	agora	0.2	0.33
palace	battlefield	0.05	0.33
palace	palace	0.9	0.34
palace	agora	0.05	0.33
agora	battlefield	0.2	0.33
agora	palace	0.2	0.33
agora	agora	0.6	0.34

Notes: Two different scenarios are considered: a condition of strong cue-influence (column 3) and one of weak cue-influence (column 4).

conquering new territory). The second case describes weak cue-influence (Table 3), evident by the fact that cues do not activate any particular context representation. In this case, UV_C and UV_* are equivalent, independent of the experienced cue. We propose that conditions of stronger and weaker cue-influence alternate during the life of everyone and that, overall, some people might be more prone to strong cue-influence and others to weak cue-influence, as reflected in one's tendency to persevere in pursuing ultimate values and ignore mundane values (Malouff et al., 1990).

This version of the theory (where cue-influence is implemented) not only can describe situations where mundane values overwhelm ultimate values. It can also interpret cases where contextual cues divert selection from one ultimate value to another; in other words, conditions where there is a conflict not between mundane versus ultimate values, but among ultimate values themselves. With this regard, empirical studies have shown that, depending on contextual information, people apply different moral principles to the same dilemmas (Bartels, 2008; Greene et al., 2004; Nichols & Mallon, 2006; Palmiotti et al., 2019). A common task adopted in these studies requires participants to decide, in fictitious scenarios, whether or not they would kill one person in order to save the life of a group of people who would die otherwise (e.g., Bartels, 2008; Nichols & Mallon, 2006). The assumption here is that two moral principles compete, one deontological ("do not kill") and the other consequentialist ("overall save as many lives as possible"), leading to the choice of killing and of not killing, respectively. Empirical evidence indicates that, in this task, contextual information can boost the appeal of either principle (Bartels, 2008; Greene et al., 2004; Nichols & Mallon, 2006; Palmiotti et al., 2019). For example, the choice of killing (reflecting the consequentialist principle) is more frequent when the number of saved people is larger (Bartels, 2008; Nichols & Mallon, 2006). Within our theory, this scenario can be described by a Bayesian network where the competing ultimate values correspond to the deontological and the consequentialist principle, respectively, and where the number of potentially saved people corresponds to the categories of the context variable C and of the contextual cue variable C_C. In the absence of any cue (i.e., with no information about the number of people saved), an agent might select the deontological principle as ultimate value, thus avoiding killing; this is because, in most contexts (i.e., in most cases in terms of number of lives saved), the act of killing might be perceived as not worth enough. However, the model implies that a cue indicating that many people can be saved might suppress the deontological rule in favour of the consequentialist principle, leading to the choice of killing. This example illustrates how our theory can explain the impact of contextual cues upon the appeal of deontological versus consequentialist principles, and in general upon competing ultimate values.

4.4 | Innate or cultural values?

To what extent are human values preprogramed by genes? And to what extent are they produced by culture? Regarding these questions, extreme positions can be found in the literature. Some proposals maintain that, after all, humans from different times and places are driven by the very same motives (Kenrick et al., 2010; Maslow, 1943; Shweder, 2012). Other accounts claim that every culture develops its own idiosyncratic values which share virtually nothing with values of other cultures (Shweder, 2012). Within this debate, our theory advocates an intermediate position. On the one hand, ultimate values (which motivate behaviour) are grounded on basic values characterised by an inbuilt affective quality, and hence virtually universal (though some degree of inter-individual variability is conceivable) and shared by all cultures. On the other hand, ultimate values do not correspond to basic values, but to states or rules associated with basic values in a variety of contexts. Which specific state or rule is associated with basic values is not pre-established, but it depends on specific conditions such as on the physical environment, on the technology available, and on the structure of society. Therefore, together with universal basic values, our theory predicts a substantial variability of ultimate values across cultures and individuals. At the same time, because formation of ultimate values is proposed to follow certain laws, our theory implicates that, when exposed to similar conditions (e.g., in terms of physical environment, technology available, and structure of society), individuals will develop similar ultimate values.

Our theory suggests that the brain is predisposed to learn which states or rules (the ultimate values) are predictive of basic values across multiple contexts. Formally, this entails learning the parameters of the Bayesian network described above. Specifically, the brain has to learn (i) which rules or states UV can potentially be considered, (ii) which contexts C can be experienced, (iii) which contextual values CV can be experienced, and (iv) the probability distributions linking these variables together. Though a full examination of learning is beyond the scope of this manuscript, some basic processes can be highlighted. Learning is likely to involve direct experience with UV, C, and CV, but also social influence (Turner, 1991). The latter could be critical for at least two reasons. First, communication with other people is arguably essential to shape the Bayesian network: although this network describes an individual belief system, it is likely to embody shared cultural ideas (especially those expressed by more powerful groups). Second, social influence is likely to be a powerful basic value, for example in the form of admiration, approval, despise, or condemnation expressed by others. Hence, we would expect that ultimate values usually foster positive social evaluation (and avoid negative evaluation). A key question is when learning of the Bayesian network (and of the ensuing ultimate values) occurs during life. Although life-long learning is a possibility, there are reasons to propose critical periods such as childhood and adolescence. Moreover, periods of high uncertainty about ultimate values (formally, captured by high entropy; see Equation 3) might lead to higher openness to learning.

In summary, our theory implicates that basic values are inbuilt and universal, while ultimate values depend on basic values but also on ontogenetic conditions such as physical environment, technology available, and structure of society. Hence ultimate values are predicted to vary across individuals and society; yet, our theory interprets this variability not as random but as deriving from precise laws.

4.5 | Ethos

Our theory offers an interpretation of how the ethos of a society or group develops and acts. Max Weber is perhaps the most influential thinker highlighting the central role of ethical values in shaping society. He famously suggested that the raise of protestant ethics, viewing economic enrichment as manifestation of God's favour, is at the root of modernity and capitalism (Weber, 1904/2002). Weber adopted the same logic to interpret the implications of other cultures, such as Confucianism and Hinduism, for the development of society (Weber, 1915/1959, 1916/2000). Weber's focus was primarily on the consequences of ethical values rather than on how these values emerge. An influential interpretation of how ethical values arise is based on the notion of ideology as proposed by Marx and his followers (Marx & Engels, 1845). According to this interpretation, the dominant class of a society is motivated by self-interested economic motives which require exploitation of subordinate classes. To justify self-interest and exploitation, the dominant class would develop ideologies grounded on specific ethical values. The latter, though apparently based on morality and justice, would in fact disguise nothing more than self-interest and exploitation. Hence, in a Marxist perspective, ethical values do not shape society: adopting a Marxian terminology, they are part of the superstructure, and not of the structure, of society. They do not drive behaviour beyond the self-interested economic motives which underly ethical values in the first place.

What are the implications of our theory for the role of ethical values? Our theory can be considered as an integration of some of the central ideas of Weber and Marx. Following Marx, our theory proposes that ethical values (termed ultimate values in our account) are grounded on basic values. We consider these basic values as largely inbuilt and universal (it is not clear whether this was also Marx's opinion). Different from Marx, our notion of basic values does not encompass solely economic self-interest, but it acknowledges a variety of elements including egoistic and altruistic aspects. Also different from Marx, in our theory ethical values are not basic values in disguise, but they are "means" to pursue basic values in a variety of contexts that are eventually treated as ends as such. Following Marx, in our theory people are normally unaware of the origin of ethical values from the underlying basic values: ethical values are viewed as simply valuable as such. This has implications for how social classes interact. For example, members of the dominant class might embrace certain ethical values (e.g., free trade) because these in fact support economic basic values of the dominant class. However, members of the dominant class might be unaware of where these ethical values come from, and simply claim them as morally just. Following Weber, our theory implicates that ethical values are critical in shaping society, above and beyond basic values. This is because our theory proposes that, once an ethical value has been selected, it is precisely this ethical value (and not the underlying basic values) that drives behaviour (even in contexts where it is not conducive of basic values). This fits with the observation that humans sometimes persist with their ethical values despite enormous losses in terms of basic values (e.g., accepting death to foster free trade).

In short, our theory offers an interpretation of how ethical values (referred to as ultimate values) arise and impact on society. Combining elements from both Weber and Marx, the theory proposes that ethical values are built upon basic values but, once established, transcend them and become the actual forces driving behaviour.

4.6 | Ultimate values and society

In our theory, means fostering basic values in a variety of contexts are transformed into ultimate values. Which means are actually effective in fostering basic values depends on factors such as physical environment, technology available, and social structure. These factors vary dramatically when comparing premodern and modern societies (Giddens, 2013). Premodern societies are characterised by stability of environment, of technology, and of social structure, while continuous change of these aspects distinguishes modern societies. Accordingly, for individuals, the effective means for obtaining basic values (upon which ultimate values are built) remain relatively stable in premodern society but change continuously in modern societies. Therefore, following our theory, ultimate values will be more stable in premodern compared to modern societies (rapidly changing ultimate values would in turn promote social change). In essence, this captures the classic idea that the more society changes, the more people's values change (and the more values change, the more society changes). Moreover, modern compared to premodern societies have higher variability in social roles (Durkheim, 1893/1997). Because each social role (e.g., the peasant or the IT technician) implies specific effective means for obtaining basic values (upon which ultimate values are built), modern compared to premodern societies imply higher inter-individual variability in ultimate values. Altogether, comparing modern versus premodern societies, our theory predicts that ultimate values vary more both along time and across individuals. The predicted consequence of this is that uncertainty about ultimate values (formally $E_{UV|BV}$; see Equation 3) will usually be higher in modern versus premodern societies. This offers an interpretation of the crisis of values often reported in association with modernity, in conjunction with vigorous values often observed in premodern societies (e.g., Bendle, 2002).

Our theory also speaks to the claim, first argued by Weber, that disenchantment emerges in modern societies. An influential theory of disenchantment is grounded on distinguishing objective from instrumental reason (Horkheimer, 1947). Objective reason would consist in the quest for what is truly right or wrong and would require focusing on the ends of human existence. Conversely, instrumental reason would seek to identify the most effective ways to achieve any goal, hence focusing on means and ignoring the question of whether goals are actually valuable. According to this perspective, modernity has witnessed a progressive shift from objective to instrumental reason, implying that values embraced by people in premodern societies are closer to "objective" human ends. Our theory does not share this claim. A fundamental assumption of our theory is that the human brain is predisposed to pursue ultimate values, which in a sense are means to obtain "objective" ends (corresponding to basic values). Hence, our theory implicates a universal human predisposition for instrumental over objective reason, that is not only characteristic of modernity. We argue that our argument is supported when one considers ultimate values typical of premodern societies such as military might, respect for hierarchy, and religious zeal. Whether these are closer to "objective" human values compared to values typical of modern societies (such as scientific progress and economic enrichment) is at least debatable.

Why have modern people abandoned values such as military might, respect for hierarchy, and religious zeal in favour of values such as scientific progress and economic enrichment? Our model suggests that, for a substantial number of people living in modern societies, modern values have been simply more effective (in terms of fostering basic values) than typical premodern values (other critical factors such as social influence - e.g., in the form of propaganda – are not incompatible with the theory). Does this imply that values such as scientific progress and economic enrichment are the best ultimate values possible in a modern society? In our theory, ultimate values are the result of constructive processes, open to imagination and creativity. Hence, in principle ultimate values better than those available can always be envisaged. This consideration has relevance for contemporary societies. It is useful to view the ultimate values prevailing in these societies as limiting, and to look for alternatives with higher potential in the context of the current environment, technology, and social structure. For example, it has been argued that, in affluent societies where economic needs are largely fulfilled, focusing on economic enrichment might often be a poor strategy (in terms of basic values) compared to cultivating more "spiritual" ultimate values (e.g., Skidelsky & Skidelsky, 2012). Our theory comfortably fits with this perspective. Our theory also stresses that, to envision better ultimate values, another key prerequisite is a realistic assessment of basic values (namely what is inherently imbued with affect by humans).

Finally, one last aspect is worth to be examined here. Empirical evidence highlights a tendency for individuals to link with people who are similar to them. A factor underlying this tendency appears to be homogeneity in values (Dehghani et al., 2016; Motyl et al., 2014): for instance, a preference for living in communities of people sharing similar ideological and moral values is commonly observed (Motyl et al., 2014). Our theory explains this preference as arising because of the (realistic) belief that living with people sharing similar ultimate values fosters realization of those values (this view is consistent with, but more general than, a previous interpretation that morally homogeneous communities help coordinating third-party moral judgements; see Dehghani et al., 2016; DeScioli & Kurzban, 2009). For instance, ultimate values in the political domain imply cooperating for building a good society, an objective which is facilitated by living with people sharing similar values. This interpretation can also explain exceptional cases where an individual chooses to live with people embracing radically different values and attempts to convert these people (e.g., a religious missionary living in foreign cultures). This choice still appears as motivated by realising the own ultimate values (e.g., spreading God's message), as these can be fostered by converting others.

In sum, our theory offers a computational perspective to study how ultimate values differ when comparing modern and premodern societies. In addition, it offers a conceptual framework to explore how novel and more fulfilling ultimate values might be envisaged in a creative way.

4.7 | Contribution to previous literature

This section highlights the specific contribution of our proposal with respect to previous research on human values. The notion of hierarchy is key in our account. This term is common in research on values, though it has been used with different meanings. In some accounts, values are ranked hierarchically based on their priority (Maslow, 1943; Wicker et al., 1984). By applying cluster or factor analysis to investigate the similarity among people's reported values (Rokeach, 1973; Schwartz, 2006, 2012; Talevich et al., 2017), other accounts have observed a hierarchical structure charactesising these reports. In contrast with these approaches, our theory adopts a notion of hierarchy based on the concept of abstraction, which considers whether some values are more or less abstract (i.e., context-independent) than others (Carver & Scheier, 1998; Powers, 1973; Talevich et al., 2017; Tsushima & Burke, 1999; Wicker et al., 1984).

Perceptual control theory (Powers, 1973) has pioneered the principle that values are arranged along hierarchical structures based on abstraction. A similar idea informs identity control theory, which focuses on how people realise salient identities in social contexts (Stets & Burke, 2014). This theory argues that representations of identities are organized along a hierarchy with more abstract identities occupying higher levels and more specific identities occupying lower levels (Stets & Burke, 2014; Tsushima & Burke, 1999). Our theory is widely consistent with these previous proposals. Its novel contribution (besides focusing on the more general notion of value instead of identity; see below) consists in explaining why certain abstract values (here referred to as ultimate values) are selected and pursued over alternative abstract values. In other words, what are the processes responsible for the formation and selection of ultimate values? So far, this crucial question has remained unaddressed by previous theories (Carver & Scheier, 1998; Powers, 1973; Talevich et al., 2017; Tsushima & Burke, 1999; Wicker et al., 1984). This issue implies important corollary questions: what is the impact of inbuilt affective processes and of learning, respectively, in the formation of ultimate values?

And regarding learning, what is the role of society? Why is there a dramatic variability in ultimate values pursued by different people? Addressing these questions represent the main specific contribution of our theory to the literature.

Identity theory (of which identity control theory is a branch) examines why individuals tend to embrace some identities and discard other identities (Stets & Burke, 2014). For example, someone strongly attached to his job identity might end up working also during weekends, when he might instead spend time with his children (and activate a paternal identity). To explain why some identities are selected and others are discarded, the concept of commitment has been proposed (Burke & Reitzes, 1991; Stryker, 1980): commitment to an identity (determining a tendency to rely on the latter) would increase if the identity is associated with (i) more benefits (e.g., money) and less costs, (ii) with more social ties, (iii) and with social ties characterised by higher bonding. The notion of commitment partially anticipates key ideas of our theory: the proposal that factors such as benefits and social ties determine which identity will be activated has analogies with our proposal that basic values determine which ultimate value will be selected. However, despite this analogy, commitment theory remains limited with respect to the arguments developed here, for several reasons. First, it frames the issue in terms of identity, whereas framing the issue more generally in terms of value is arguably preferable: ultimate values can be about the self (and hence about identity; e.g., "be a good father"), but not necessarily (e.g., "make your child rich") (identity theory research is recently moving towards a similar direction too, exploring the idea of moral identity; Stets & Carter, 2012). Second, contrary to our approach, commitment theory does not apply to identities or values arranged hierarchically, and it does not clarify to what extent commitment to an identity leads to ignoring the ongoing context. Third, the nature of the benefits supporting commitment for an identity remains opaque, while here we rely on the notion of basic values (linked to the well-established notion of unconditioned stimuli). These and other shortcomings of commitment theory prevent addressing important questions examined here, such as to what extent ultimate values are culturally determined, and to what extent they shape behaviour outside basic values.

5 | DISCUSSION

This paper offers a theory about the notion of ultimate value. The key idea is that our brain represents values along a hierarchy where ultimate values are built upon experiences with inherent affective nature (basic values). Based on these representations, the proposal is that the brain infers the rules or states that foster basic values in a variety of contexts. These become ultimate values and drive human behaviour independent of the ongoing context. This perspective implicates that, when making choice, ultimate values are not given a priori, but they are the result of an inference process. The implicit assumption is that the brain has adapted in a way to break down choice in two processes: evaluation (resulting in the selection of ultimate values) and planning (resulting in the selection of the appropriate chain of actions). We have discussed how our theory can contribute to understanding a variety of aspects of human values, including morality, to what degree values are innate or culturally determined, and how values shape and are shaped by society.

We highlight some limitations of the theory in its current form, which can potentially be addressed by future research. First, the theory focuses only on evaluation processes and not on planning; though these two processes are assumed to unfold largely independently, a full picture requires integrating them together. Second, the current theory does not examine the precise computations underlying the acquisition of the Bayesian network. We have stressed that direct learning and social influence are arguably two critical factors; yet, a fine-grained analysis of how these and other factors (e.g., the broader role of the social structure) unfold would offer much further insight. Third, the Bayesian network relies on a simplistic representation of basic values, as these are described by a dichotomous variable having reward and punishment as categories. A more sophisticated implementation of basic values, for example distinguishing the qualitative difference among them, would be beneficial. Finally, here we have assumed that one single ultimate value is selected, the one with higher posterior probability. However, selection might work in a different way: for example, all ultimate values available might motivate behaviour, each with a weight proportional to its posterior probability.

The paper offers examples of how our theory can contribute to research areas such as morality and the study of values in society. Here, the contribution of the theory is only sketched; an interesting avenue is to explore this in more detail. Furthermore, the theory might potentially be relevant for other research areas where human values are critical. Two examples are economics and clinical psychology. The standard economic approach is agnostic about the origin of values that drive choice. However, as shown by empirical investigations highlighting the role of religion in shaping market decisions (Iyer, 2016), understanding where values come from is fundamental to explain economic behaviour. Our theory can contribute to shedding light on this. Regarding clinical psychology, impairments in evaluation processes are critical in a variety of psychopathologies such as depression, addiction, and anxiety. Our proposal of ultimate values playing a central role in evaluation raises the question of whether, in some forms of mental illness, formation and selection of ultimate values might be impaired. For example, an intriguing possibility is that depression might be interpreted as a condition of extremely high uncertainty about ultimate values (see Equation 3), with low mood deriving from ultimate values conducive of scarce basic values (captured by inferring $P(BV = reward \mid UV = UV_*)$).

In summary, the paper contributes to understanding the nature of human values by proposing that the brain represents values along a hierarchy, with more abstract values occupying the higher hierarchical level. These representations would be adopted by the brain to infer the ultimate values to be pursued. The theory aims at contributing to research in a variety of domains where human values are of critical importance.

CONFLICT OF INTEREST

The author declares no conflict of interest.

ORCID

Francesco Rigoli https://orcid.org/0000-0003-2233-934X

REFERENCES

Austin, J. T., & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin*, *120*(3), 338–375.

Balaguer, J., Spiers, H., Hassabis, D., & Summerfield, C. (2016). Neural mechanisms of hierarchical planning in a virtual subway network. *Neuron*, *90*(4), 893–903.

Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, *108*(2), 381–417. https://doi.org/10.1016/j.cognition.2008.03.001

Bell, M. C., & McDevitt, M. A. (2014). Conditioned reinforcement. *The Wiley Blackwell handbook of operant and classical conditioning*, pp. 221–248.

Bendle, M. F. (2002). The crisis of 'identity' in high modernity. *The British Journal of Sociology*, *53*(1), 1–18.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, *22*(6), 956–962.

Burke, P. J., & Reitzes, D. C. (1991). An identity theory approach to commitment. *Social Psychology Quarterly*, *54*, 239–251.

Carver, C. S., & Scheier, M. F. (1998). *On the self-regulation of behavior*. Cambridge University Press.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366.

Dawson, S., & Kim, M. (2009). External and internal trigger cues of impulse buying online. *Direct Marketing: An International Journal*.

Dehghani, M., Johnson, K., Hoover, J., Sagi, E., Garten, J., Parmar, N. J., Vaisey, S., Iliev, R., & Graham, J. (2016). Purity homophily in social networks. *Journal of Experimental Psychology: General*, *145*(3), 366–375.

DeScioli, P., & Kurzban, R. (2009). Mysteries of morality. *Cognition*, *112*(2), 281–299. https://doi.org/10.1016/j.cognition.2009.05.008

Diener, E. (2009). *The science of well-being: The collected works of Ed Diener* (Vol. 37). Springer.

Durkheim, É. (1893/1997). *The division of labour in society*. Free Press.

Fiske, S. T. (2004). *Social beings: A core motives approach to social psychology*. Wiley.

Friston, K., & Buzsáki, G. (2016). The functional anatomy of time: what and when in the brain. *Trends in Cognitive Sciences*, *20*(7), 500–511.

Giddens, A. (2013). *The consequences of modernity*. John Wiley & Sons.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Horkheimer, M. (1947). *Eclipse of reason* (Vol. 1). Bloomsbury Publishing.

Inglehart, R. (1997). *Modernization and postmodernization: Cultural, economic, and political change in 43 societies*. Princeton University Press.

Iyer, S. (2016). The new economics of religion. *Journal of Economic Literature*, *54*(2), 395–441.

Kenrick, D. T., Griskevicius, V., Neuberg, S. L., & Schaller, M. (2010). Renovating the pyramid of needs: Contemporary extensions built upon ancient foundations. *Perspectives on Psychological Science*, *5*(3), 292–314.

Laidlaw, J. (2002). For an anthropology of ethics and freedom. *Journal of the Royal Anthropological Institute*, *8*(2), 311–332.

Lea, S. E., & Webley, P. (2006). Money as tool, money as drug: The biological psychology of a strong incentive.

Little, J. L., & McDaniel, M. A. (2015). Individual differences in category learning: Memorization versus rule abstraction. *Memory & Cognition*, *43*(2), 283–297.

Maisto, D., Donnarumma, F., & Pezzulo, G. (2015). Divide et impera: Subgoaling reduces the complexity of probabilistic inference and problem solving. *Journal of the Royal Society Interface*, *12*(104), 20141335.

Malouff, J., Bauer, M., Mantelli, D., Pierce, B., Cordova, G., Reed, E., & Schutte, N. (1990). Development and evaluation of a measure of the tendency to be goal oriented. *Personality and Individual Differences*, *11*(12), 1191–1200.

Marx, K., & Engels, F. (1845). *The german ideology* (Vol. 1). International Publishers Co.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4).

Motyl, M., Iyer, R., Oishi, S., Trawalter, S., & Nosek, B. A. (2014). How ideological migration geographically segregates groups. *Journal of Experimental Social Psychology, 51*, 1–14. https://doi.org/10.1016/j.jesp.2013.10.010

Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. Oxford University Press.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*(3), 530–542. https://doi.org/10.1016/j.cognition.2005.07.005

O'Reilly, R. C., Hazy, T. E., Mollick, J., Mackie, P., & Herd, S. (2014). *Goal-driven cognition in the brain: a computational framework. arXiv preprint arXiv:1404.7591.*

Palmiotti, G. P., Cristaldi, F. D. P., Cellini, N., Lotto, L., & Sarlo, M. (2019). Framing the outcome of moral dilemmas: Effects of emotional information. *Ethics & Behavior*, *30*(3), 213–229. https://doi.org/10.1080/10508422.2019.1607348

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford University Press.

Plutarch. (2004). *The life of Alexander the Great*. Modern Library.

Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical active inference: A theory of motivated control. *Trends in Cognitive Sciences*, *22*(4), 294–306.

Powers, W. T. (1973). *Behavior: The control of perception*. Aldine.

Rescorla, R. A. (1980). *Pavlovian second-order conditioning (psychology revivals): studies in associative learning*. Psychology Press.

Rigoli, F. (2021a). Masters of suspicion: A Bayesian decision model of motivated political reasoning. *Journal for the Theory of Social Behaviour*. https://doi.org/10.1111/jtsb.12274

Rigoli, F. (2021b). A computational perspective on faith: Religious reasoning and Bayesian decision. *Religion, Brain & Behavior*, *11*(2), 147–164. https://doi.org/10.1080/2153599X.2020.1812704

Rigoli, F., Pezzulo, G., Dolan, R., & Friston, K. (2017). A goal-directed Bayesian framework for categorization. *Frontiers in Psychology*, *8*, 408.

Robbins, J. (2012). Cultural values. *A Companion to Moral Anthropology*, 117–132.

Rokeach, M. (1973). *The nature of human values*. Free Press.

Schwartz, S. (2006). A theory of cultural value orientations: Explication and applications. *Comparative Sociology*, *5*(2-3), 137–182.

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, *2*(1), 2307–0919.

Shweder, R. A. (2012). Relativism and universalism. In D. Fassin (Ed.), *A companion to moral anthropology*. Wiley.

Schmajuk, N. A., & Holland, P. C. (1998). *Occasion setting: Associative learning and cognition in animals*. American Psychological Association.

Skidelsky, E., & Skidelsky, R. (2012). *How much is enough?: money and the good life*. Penguin UK.

Skinner, B. F. (1953). *Science and human behavior (No. 92904)*. McMillan.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychological Review*, *119*(1), 120–154.

Stets, J. E., & Burke, P. J. (2014). The development of identity theory. In *Advances in group processes*. Emerald Group Publishing Limited.

Stets, J. E., & Carter, M. J. (2012). A theory of the self for the sociology of morality. *American Sociological Review*, *77*(1), 120–140.

Stich, S., Doris, J. M., & Roedder, E. (2010). Altruism. In J. M. Doris (Ed.), *The moral psychology handbook* (pp. 147–205). Oxford University Press.

Stryker, S. (1980). *Symbolic interactionism: A social structural version*. Benjamin-Cummings Publishing Company.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

Talevich, J. R., Read, S. J., Walsh, D. A., Iyer, R., & Chopra, G. (2017). Toward a comprehensive taxonomy of human motives. *PloS One*, *12*(2), e0172279.

Tan, L., & Hackenberg, T. D. (2015). Pigeons' demand and preference for specific and generalized conditioned reinforcers in a token economy. *Journal of the Experimental Analysis of Behavior*, *104*(3), 296–314.

Tsushima, T., & Burke, P. J. (1999). Levels, agency, and control in the parent identity. *Social Psychology Quarterly*, *62*(2), 173189.

Turner, J. C. (1991). *Social influence*. Thomson Brooks/Cole Publishing Co.

Weber, M. (1904/2002). *The Protestant Ethic and the Spirit of Capitalism: With Other Writings on the Rise of the West*. Penguin.

Weber, M. (1915/1959). *The religion of China: Confucianism and Taoism*. Free Press.

Weber, M. (1916/2000). *The religion of India: The sociology of Hinduism and Buddhism*. Munshiram Manoharlal.

Wicker, F. W., Lambert, F. B., Richardson, F. C., & Kahler, J. (1984). Categorical goal hierarchies and classification of human motives. *Journal of Personality*, *52*(3), 285–305.

Widlok, T. (2004). Sharing by default? Outline of an anthropology of virtue. *Anthropological Theory*, *4*(1), 53–70.

How to cite this article: Rigoli, F. (2021). The psychology of ultimate values: A computational perspective. *J Theory Soc Behav*, 1–22. https://doi.org/10.1111/jtsb.12311