Defining Human Values for Value Learners

Kaj Sotala

Machine Intelligence Research Institute, Berkeley, CA 94704, USA;
Theiss Research, La Jolla, CA 92037, USA
kaj.sotala@intelligence.org

Abstract

Hypothetical “value learning” AIs learn human values and then try to act according to those values. The design of such AIs, however, is hampered by the fact that there exists no satisfactory definition of what exactly human values are. After arguing that the standard concept of preference is insufficient as a definition, I draw on reinforcement learning theory, emotion research, and moral psychology to offer an alternative definition. In this definition, human values are conceptualized as mental representations that encode the brain’s value function (in the reinforcement learning sense) by being imbued with a context-sensitive affective gloss. I finish with a discussion of the implications that this hypothesis has on the design of value learners.

1. Introduction

The value learning problem (Dewey 2011, Soares 2014) is the challenge of building AI systems which can first learn human values and then to act in accordance to them. Approaches such as inverse reinforcement learning (Ng & S. Russell 2000) have been suggested for this problem (S. Russell 2015, Sezener 2015), as have more elaborate ones such as attempting to extrapolate the future of humanity’s moral development (Yudkowsky 2004, Tarleton 2010, Muehlhauser & Helm 2012). However, none of these proposals have yet offered a satisfactory definition of what exactly human values are, which is a serious shortcoming for any attempts to build an AI system that was intended to learn those values.

This paper builds on a combination of research into moral psychology, the psychology of emotion, and reinforcement learning theory to offer a preliminary definition of human values, and how that definition might be used to design a value learning agent.

I begin with an argument for the standard concept of preference being insufficient as a definition of value in section 2. Section 3 introduces theoretical background from the field of reinforcement learning and particularly evolutionary reinforcement learning. The background is used in section 4 to offer a preliminary definition of human values as mental representations which encode the brain’s value function (in the reinforcement learning sense, as discussed below) by being imbued with affect. Section 5 elaborates on why affect might be a reasonable candidate for the brain’s way of encoding a value function, and section 6 discusses the connections between emotions, affect, and values. Section 7 discusses the relation of affect and moral judgment in light of the social intuitionist model of morality, and section 8 talks about how this concept of human values could be used for designing value learning agents. Sections 9 and 10 conclude by evaluating the model and comparing it to alternatives.

2. The standard concept of preference is insufficient as a definition of value

The closest existing concept that formalizes something akin to human value is the concept of a utility function, which is widely used in economics and decision theory. Possibly its most well-known problem as a model of value is the empirical finding that humans violate the axioms of utility theory and thus do not have consistent utility functions (Tversky and Kahneman 1981). However, this is far from being the most serious problem.

The von Neumann-Morgenstern utility theorem (von Neumann & Morgenstern 1953) sets up utility functions via preference orderings: of options A and B, either A is preferred to B, B is preferred to A, or both are equally preferred. Essentially, a “preference” is defined as a function that, given the state of the agent and the state of the world in general, outputs an agent’s decision between two or more choices.

A strength of this definition is that it allows treating preferences as black boxes. This has been of great use in economics, as it allows constructing models of behavior based only on observed preferences, without needing to know the reasons for those preferences.

At the same time, ignoring everything that happens inside the preference function is also a weakness for the definition. Preferences are essentially considered atomic units with no internal structure. This leads to a number of problems in trying to use them as a definition of human values, including the below.

The utility function model of value has difficulty dealing with internal conflicts and higher-order

preferences. A drug addict may desire a drug, while also desiring that he not desire it (Frankfurt 1971). “Less Is More” is a measure of executive function in which children may point either to a tray with five treats or to a tray with two treats, while knowing that they will get the treats from the tray which they didn’t point at. Three-year old children frequently fail this task and point at the tray with more treats, despite understanding that this will give them fewer things that they want (Carlson et al. 2005). Although the researchers did not report the children's reaction to their repeated failure, it seems safe to presume that they were not particularly happy, nor would they have liked to have their preference modeled as preferring fewer treats.

The utility function model of value ignores the person’s internal experience. Although “wanting” and “liking” are frequently thought to be the same thing, the two have distinct neural processes: "[l]iking corresponds closely to the concept of palatability; wanting, by comparison, corresponds more closely to appetite or craving” (Berridge 1996). Different interventions may suppress wanting without affecting liking, and vice versa. Intuitively, it seems like behaviors which we both “like” and “want” should be more important than behaviors that we only “want”.

The utility function model of value does not model changing values. As a black box mechanism, classical preference has no model of changing values, preventing us from extrapolating possible development of values.

The utility function model of value does not give a way to generalize from our existing values to new ones. Technological and social change frequently restructures the way that the world works, forcing us to reconsider our attitude towards the changed circumstances.

As a historical example (Lessig 2004), American law traditionally held that a landowner did not only control his land but also everything above it, to “an indefinite extent, upwards”. Upon the invention of this airplane, this raised the question: could landowners forbid airplanes from flying over their land, or was the ownership of the land limited to some specific height, above which the landowners had no control?

The US Congress chose to the latter, designating the airways as public, with the Supreme Court choosing to uphold the decision in a 1946 case. Justice Douglas wrote in the court’s majority that

> The air is a public highway, as Congress has declared. Were that not true, every transcontinental flight would subject the operator to countless trespass suits. Common sense revolts at the idea.

By the decision of Congress and the Supreme Court, the concept of landownership was redefined to only extend a limited, and not an indefinite, amount upwards. Intuitively, one might think that this decision was made because the redefined concept did not substantially weaken the position of landowners, while allowing for entirely new possibilities for travel.

However, a black-box approach to value, which does not reveal the reasons underlying preferences such as “landownership should extend indefinitely upwards”, would be incapable of making such a judgment.

3. Evolutionary reinforcement learning

A good theory of human psychology, including human value, requires an understanding of the evolutionary functions of the psychological phenomena (Tooby & Cosmides 1995). Before we can develop a good model of what human values are, we need to develop an understanding of their computational role in the kinds of tasks that human brain has needed to perform.

A defining characteristic of human thought is the ability to develop solutions to novel problems in novel environments. Humans are capable of learning a wide variety of behaviors far beyond anything that evolution could have “preprogrammed” into them. Instead, they experience some events (such as tissue damage or hunger) as aversive and learn to avoid things that cause those events, while learning to pursue things that feel rewarding.

The problem of learning a novel environment in order to maximize the amount of rewards is the reinforcement learning problem, which “explicitly considers the whole problem of a goal-directed agent interacting with an uncertain environment” (Sutton & Barto 1998), as opposed to merely considering some isolated subproblems.

As the theory of reinforcement learning is the general answer to the question of how an agent should behave in an uncertain environment and learn from it, we should expect the design of both human and animal minds to be strongly shaped by the principles of reinforcement learning theory. Empirical evidence from a variety of fields, including behavioral trainers (Pryor 1999), studies on habit-formation (Duhigg 2012) as well as neuroscience (Dayan 2011) supports this prediction.

Standard reinforcement learning theory involves learning to maximize a reward signal which the agent can observe. However, evolution selects for traits with the highest inclusive fitness, an abstract measure of a trait’s effect on the production and survival of direct and related offspring. As organisms cannot directly observe the effect of their actions on their lifetime fitness, they cannot maximize this value directly.

Singh et al. (2009, 2010) expand reinforcement learning to cover the evolutionary case, and define an “optimal reward function” as follows. An agent A in an external environment e receives observations and takes actions. It has an internal environment which computes a state based on the observations from the environment. The agent tries to maximize a reward, which is also computed by the internal environment according to a reward function rA, drawn from the space of reward functions RA.

Different agents carry out actions in various environments e drawn from a distribution of environments E. A specific agent A in environment e with reward function rA produces a history h. A fitness function F produces a scalar evaluation F(h) for all histories h. A reward function is optimal if it maximizes the expected fitness of the agent over the distribution of environments.

This formalization mimics an evolutionary environment in that evolution selects for agents which best maximize their fitness, while agents cannot directly optimize for their own fitness as they are unaware of it. Agents can however have a reward function that rewards behaviors which increase the fitness of the agents. The optimal reward function is one which maximizes (in expectation) the fitness of any agents having it. Holding the intelligence of the agents constant, the closer an agent's reward function is to the optimal reward function, the higher their fitness will be. Evolution should thus be expected to select for reward functions that are closest to the optimal reward function. In other words, organisms should be expected to receive rewards for carrying out tasks which have been evolutionarily adaptive in the past.

4. An initial definition of human value

The theory of reinforcement learning distinguishes between reward and value. A value function assigns states of the world a scalar value based on the expectation of future rewards that one may obtain from that state, conditional on some policy of what one would do in each state. Intuitively, a state has a high value if one can reliably move from it to states with a high reward. For reinforcement learning to work effectively, it requires a way to identify states which should be classified as the same or similar, and be assigned the same or a similar value.

We can now consider the relation between the need to identify similar states, and mental concepts. We should expect an evolutionarily successful organism to develop concepts that abstract over situations that are similar with regards to receiving a reward from the optimal reward function. Suppose that a certain action in state s1 gives the organism a reward, and that there are also states s2-s5 in which taking some specific action causes the organism to end up in s1. Then we should expect the organism to develop a common concept for being in the states s2-s5, and we should expect that concept to be "more similar" to the concept of being in state s1 than to the concept of being in some state that was many actions away.

Empirical support for concepts being organized in this kind of a manner comes from possibly the most sophisticated general-purpose AI developed so far, DeepMind's deep reinforcement learning agent (Mnih et al. 2015). This agent managed to "achieve a level comparable to that of a professional human games tester across a set of 49 [Atari 2600] games, using the same algorithm, network architecture and hyperparameters".

This agent developed an internal representation of the different game states of each game that it was playing. An investigation of the agent's representation for the game Space Invaders indicated that representations with similar values were mapped closer to each other in the representation space. Also, some game states which were visually dissimilar to each other, but had a similar value, were mapped to internal representations that were close to each other. Likewise, states that were visually similar but had a differing value were mapped away from each other. We could say that the agent learned a primitive concept space, where the relationships between the concepts (representing game states) depended on their value and the ease of moving from one game state to another.

There is considerable disagreement on what exactly concepts *are*, and various theoreticians use the same term to refer to different things (Machery 2010). For the purposes of this paper, I am loosely defining a "concept" as points or regions within a conceptual space, with concepts having a hierarchical structure so that higher-level concepts are at least partially defined in terms of lower-level ones. Similar assumptions are commonly made in psychology (Gärdenfors 2004) and neuroscience (Wessinger et al. 2001).

Additionally, this definition makes concepts remarkably similar to the representations built up in the machine learning subfield of deep learning. Deep learning models have demonstrated success in a wide range of tasks, including object recognition, speech recognition, signal processing, natural language processing and transfer learning (Bengio 2012, Schmidhuber 2014). They work by building up an internal representation of a domain, where different concepts are arranged in a hierarchical structure, with more abstract concepts at the top.

These ideas allow us to establish a preliminary definition of value in the "human value" sense. I suggest that human values are concepts which abstract over situations in which we've previously received rewards, making those concepts and the situations associated with them valued for their own sake. A further suggestion is that, as humans tend to naturally find various mental concepts to be associated with affect (the subjective experience of a feeling or emotion, experienced as either positive or negative), the value function might be at least partially encoded in the affect of the various concepts.

In support of this possibility, I next turn to some of the research studying the role of affect in decision-making.

5. Affect as a possible representation for the value function

Affective evaluations of concepts seem to influence people's behavior. For instance, Benthin et al. (1995) found that the experienced affective feel of mental images associated with various health-related behaviors predicted the extent to which high schoolers engaged in those behaviors. Another study (Peters & Slovic 1996) surveyed a representative sample of the US adult population. This study found that both the respondents' general worldview and their affective associations with nuclear power predicted the respondents' support for nuclear power independently of each other.

This kind of a reliance on immediate affective responses to various options in guiding decision-making has been named the affect heuristic, and documented in a number of studies (Slovic et al. 2007).

However, the dissociation between “wanting” and “liking” (Berridge 1996) suggests that the value function may not be completely contained in affective evaluations, as it is possible to “want” things without “liking” them, and vice versa. I am choosing to regardless mainly focus on the affective (“liking”) component. This is due to the intuition that, in the context of looking for a target of value learning, the values that are truly important for us are those that involve a “liking” component, rather than the values with a “wanting” component without a “liking” component. The former seem closer to things that we like and enjoy doing, while the latter might be closer to behaviors such as undesired compulsions. I wish to emphasize, however, that this is only a preliminary conjecture and one which still needs further investigation.

In order to be a good candidate for the representation of a value function, the affect of different concepts should vary based on contextual parameters such as the internal state of the organism, as (for example) a hungry and non-hungry state should yield different behaviors.

Rats who taste intense salt usually both “dislike” and “unwant” it, but when they become salt-deprived, they will start both “wanting” and “liking” the salt, with the “wanting” expressing itself even before they have had the chance to taste the salt in the salt-deprived state and consciously realize that they now enjoy it (Tindell et al. 2009). Thus it seems that both the affective value and “wanting” of something can be recomputed based on context and the organism’s own state, as would be expected for something encoding a value function.

Similarly, a state such as fear causes shifts on our conceptual frames, such as in the example of a person who’s outside alone at night starting to view their environment in terms of “dangerous” and “safe”, and suddenly viewing some of their familiar and comfortable routes as aversive (Cosmides & Tooby 2004). This is another indication of the affect values of different concepts being appropriately context-dependent.

The negative or positive affect associated with a concept may also spread to other related concepts, again as one would expect from something encoding a value function. A person who is assaulted on a particular street may come to feel fear when thinking about walking on that street again. The need to be ready to confront one’s fears and pains is also emphasized in some forms of therapy: if a person with a fear of snakes turns down invitations to go to a zoo out of a fear of seeing snakes there, they may eventually also become anxious about any situation in which they might be invited to a zoo, and then of any situation that might lead to *those* kinds of situations, and so on (Hayes & Smith 2005). Such a gradual spreading of the negative affect from the original source to related situations seems highly similar to a reinforcement learning agent which is updating its value function by propagating the value of a state to other states which precede it.

6. Human values and emotions

Human values are typically strongly related to emotional influences, so a theory which seeks to derive human values from a reinforcement learning framework also needs to integrate emotions with reinforcement learning.

A major strand of emotion research involves appraisal theories (Scherer 1999, Roseman & Smith 2001, Scherer 2009), according to which emotional responses are the result of an individual’s evaluations (appraisals) of various events and situations. For example, a feeling of sadness might be the result of an evaluation that something has been forever lost. The evaluations then trigger various responses that, ideally, orient the organism towards acting in a manner appropriate to the situation. After an evaluation suggests that something important has been forever lost, the resulting sadness may cause passivity and a disengagement from active goal pursuit, an appropriate response if there is nothing that could be done about the situation and attempts to pursue the goal would only lead to wasting resources (Roseman & Smith 2001).

An important property of emotional appraisals is that different situations which might cause the same evaluation may not have any physical features in common with each other:

> Physically dissimilar events (such as the death of a parent and the birth of a child) may produce the same emotion (e.g. sadness) if they are appraised in similar ways (e.g. as involving the loss of something valued). An infinite number of situations can elicit the emotion because any situation that is appraised as specified will evoke the same emotion, including situations that have never before been encountered. Thus, the loss of one’s first love or first cherished possession is likely to elicit sadness; and if people develop the ability to clone copies of themselves, a man who wants this capability but believes that he has lost it will feel sad. (Roseman & Smith 2001)

In other words, emotional responses are a result of appraisals abstracting over situations which are similar on some specific property that has been evolutionarily important. As such, in addition to their direct psychological and physiological effects, they could also be seen as providing a reinforcement learning agent with information about which states are similar and should be treated as similar for learning purposes.

Emotions are also associated with an affect dimension, with the conscious experience of an emotion often being theorized as being the integral blend of its affect (unpleasant-pleasant) and arousal (lethargic-energetic) dimensions (J. Russell 2003).

Combining the above ideas, it is natural to suggest that since emotional appraisals identify evolutionarily important states, the optimal reward function for humans' environment of evolutionary adaptedness (EEA, Tooby & Cosmides 1990) has involved positive rewards for emotions which reflect desirable states, and negative rewards for emotions which reflect undesirable states.

Marinier & Laird (2008) experimented with implementing a reinforcement learning-driven agent with simple appraisal-based emotions in a toy environment. They found the agent with emotions to learn faster than a standard reinforcement learning agent, as the emotion-equipped agent received frequent feedback of its progress from its appraisals and thus learned faster than the standard agent, which only received feedback when it reached its goal.

Humans and many animals also enjoy exploration and play from a very early age, in a manner which cannot be explained by those exploratory behaviors having been reinforced by other rewards. Singh et al. (2009) set up a simulated environment in which agents could move about and take actions for the first half of their lifetimes, but could not yet carry out actions that would increase their fitness. In the second half of the agents' lifetimes, actions which increased their fitness became available. The authors found that the optimal reward function for this kind of an environment is one that rewards the agents for learning simple behaviors that can be performed during their "childhood", and which are prerequisites for the more complex fitness-increasing behaviors. Once the more complicated fitness-increasing behaviors become possible during the "adulthood" of the agents, agents with a reward function that has already taught them the simpler forms of the behavior will increase their fitness faster than agents that do not engage in "childhood play" and have to learn the whole behavioral chain from scratch. This and similar examples (Singh et al. 2010) on the value of behaviors such as curiosity and play provide an explanation for why humans would find those behaviors rewarding for their own sake, even when the humans were modeled as reinforcement learners who did not necessarily receive any other rewards from their play.

7. Human values and morality

The discussion so far has suggested that human values are concepts that have come to be associated with rewards, and are thus imbued with a (context-sensitive) level of affect. However, I have said little about morality in particular.

The social intuitionist model of moral psychology (Haidt 2001) proposes that moral judgment is "generally the result of quick, automatic evaluations (intuitions)". It can be contrasted to rationalist models, in which moral judgments are the results of careful moral reasoning. Haidt (2001) begins with a discussion of people's typical reactions to the following vignette:

> Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that? Was it OK for them to make love?

Haidt (2001) notes that most people will have an instant negative reaction to the vignette and say that what the siblings did was wrong (Haidt et al. 2000). Yet the reasons that people offer for the act having been wrong are inconsistent with the story that was presented: for example, people might offer the possibility of birth defects from inbreeding, only to be reminded that the siblings were thorough in using birth control. This is used as an illustration of Haidt's (2001) claim that "moral reasoning is usually a post hoc construction, generated after a judgment has been reached".

In particular, moral judgments are thought to be strongly driven by moral intuitions, which are defined as

> ...the sudden appearance in consciousness of a moral judgment, including an affective valence (good-bad, like-dislike), without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion. Moral intuition is therefore the psychological process that the Scottish philosophers talked about, a process akin to aesthetic judgment: One sees or hears about a social event and one instantly feels approval or disapproval. (Haidt 2001)

I suggest that the social intuitionist model is highly compatible with my framework. Recall that I defined human values as concepts which have become strongly associated with positive or negative affect. Something like a moral intuition for brother-sibling incest being something abhorrent, could be explained if something like the hypothesized Westermarck Effect (Rantala & Marcinkowska 2011) made individuals find the concept of having sex with their siblings to be abhorrent and strongly laden with negative affect. Thus the concept of incest would instantly cause a negative reaction, leading to a moral judgment of condemnation.

The social part of social intuitionist theory emphasizes the impact of culture and the social environment in shaping various moral intuitions. Haidt (2001) suggests at least three kinds of cultural processes which shape intuitions:

1. Selective loss of intuitions is the suggestion that people are from birth capable of developing many different kinds of intuitions, but that intuitions which are not emphasized by the prevailing culture gradually become weaker and less accessible. This is suggest to possibly be analogous to the process in which children lose the ability to distinguish between phonemes which are not distinguished in their native language.

2. Immersion in custom complexes. Various customs that are practiced in a culture are hypothesized to affect the implicit beliefs of people growing up in that culture. For example, the culture in Orissa, India structures spaces and objects by rules of purity and pollution. This involves rules such as dividing temples to zones of increasing purity, with foreigners and dogs being allowed near the entrance, bathed worshippers being allowed into the courtyard, and only the Brahmin priest being allowed into the inner sanctum. It is suggested that after a life of navigating such rules and practices, children internalize a way of thought that makes later intellectual concepts of sacredness, asceticism and transcendence feel natural and self-evident.

It is interesting to note that this suggestion maps fits naturally into my suggested model of the role of concepts. If the function of concepts is to foster the right behavior in the right situations, then a person who is required by their culture to internalize a set of allowed, required, and disallowed behaviors in various high- or low-purity zones needs to develop a set of concepts which link the right behaviors together with the appropriate levels of purity. Once this conceptual network is in place, even if only in an implicit and unconscious level, new concepts which share a similar structure with the previously-learned one may feel easy and intuitive to develop.

3. Peer socialization. Many moral intuitions are learned from the culture in one's peer group; in particular, there might be evidence that immersion within a culture between the ages of 9 and 15 causes permanent internalization of the norms of that culture in a way that causes them to feel obvious and intuitive.

Social intuitionist theory proposes that moral judgments involve a number of moral intuitions, but does not explicitly outline what those intuitions are or where they come from. Other theories building on social intuitionism, such as moral foundations theory (Graham et al. 2012, Haidt 2012) have proposed various foundations from which the intuitions are derived. For example, the care/harm foundation is hypothesized to have its origins in an evolutionary adaptation to care for one's offspring, and to motivate people to care for others and to help them avoid harm. While my model is not committed to any particular set of moral intuitions, theories such as moral foundations are broadly compatible with the model, offering an additional set of sources through which concepts may become imbued with either positive or negative affect.

8. Building value learners

In my framework, various sources of reward lead the brain to calculate an approximation of a value function, which then becomes expressed in the affect of various concepts. This preliminary definition of values seems to suggest ways to implement value learning in ways which avoid some of the problems associated with a more structure-free model of preferences.

I have discussed some sources of reward, including classical physical events such as food or physical pain, the affective dimension in various emotional reactions, and moral intuitions. A further candidate for a source of reward might be something like self-determination theory, which posits the existence of the three universal human needs of competence, autonomy and relatedness (Ryan & Deci 2000). I do not expect this list to be comprehensive, and is rather intended as illustrative only.

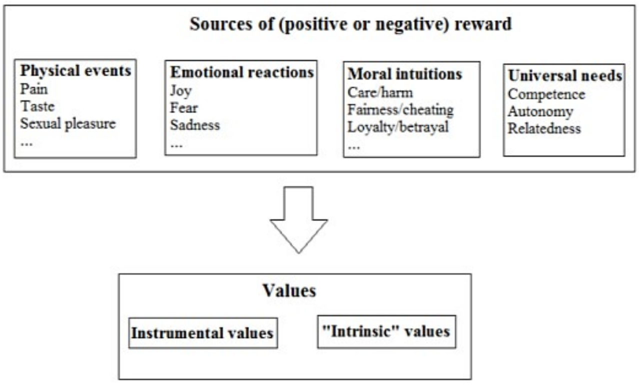

Figure 1. Various source of positive or negative reward may lead to various concepts becoming imbued with reward, giving rise to both "intrinsic" values (which are valued even if the source to the original reward was severed) and instrumental values, which lose their appeal if they cease to be associated with the original source of reward.

A value learning agent attempting to learn the values of a human might first map the sources of reward for that human. Given a suitable definition of sources of reward, discovering the various sources for any given individual seems like the kind of a task that might be plausibly outsourced to the AI system, without all of the sources needing to be exhaustively discovered by human researchers first.

Having this information would assist the value learner in mapping the individual's values, in the form of a map of concepts and their associated affect values. Once the map had been obtained, it could be used to generate preference rankings of different outcomes for the individual and the world, using similar mechanisms as the human brain does when considering the appeal of different outcomes.

When unsure about the desirability of some scenario, the value learner could attempt to model the amount of reward that the human would receive from living in the circumstances implied by that scenario, and use this to evaluate what value the human would then come to place on the scenario.

One issue with this model is that there is no theoretical reason to expect that a person's values would necessarily imply consistent preference orderings, as people tend to have contradictory values. To resolve this issue, I turn back to the social intuitionist models of morality. While social intuitionism places stronger value on moral intuitions than rationalist models do, it must be emphasized that it incorporates, rather than rejects, the possibility of moral reasoning as well. Haidt (2001) writes:

> ...people may sometimes engage in private moral reasoning for themselves, particularly when their initial intuitions conflict. Abortion may feel wrong to many people when they think about the fetus but right when their attention shifts to the woman. When competing intuitions are evenly matched, the judgment system becomes deadlocked [...]. Under such circumstances one may go through repeated cycles of [...] using reasoning and intuition together to break the deadlock. That is, if one consciously examines a dilemma, focusing in turn on each party involved, various intuitions will be triggered [...], leading to various contradictory judgments [...]. Reasoning can then be used to construct a case to support each judgment [...]. If reasoning more successfully builds a case for one of the judgments than for the others, the judgment will begin to feel right and there will be less temptation (and ability) to consider additional points of view. [...] We use conscious reflection to mull over a problem until one side feels right. Then we stop.

I suspect that to the extent that this model of human moral reasoning is correct, it could also be used by the value learner to predict how the individual being modeled might resolve any particular inconsistency. Because the value learner could potentially consider a far larger set of relevant intuitions at once, it could also help the individual in their moral growth by nudging them to consider particular scenarios they might otherwise not have considered.

Drawing partial inspiration from Christiano's (2014) notion of approval-directed agents, this kind of an approach might also help avoid classical problems with seeking to maximize the fulfillment of an individual's preferences, such as the possibility of rewriting a person's mind to maximize the amount of reward the person obtained. This could be accomplished by applying the same criteria for evaluating potential outcomes, into evaluating the value learner's actions. Would the human's values approve the scenario where the value learner attempted to rewrite the human's brain? If the answer is no, that course of action would be prohibited.

Depending on the exact path taken, it seems reasonable to assume that this kind of a moral reasoning process might arrive at several different conclusions. Conceivably, if an examination of different perspectives makes some side in a moral dilemma feel more right than the other, the intuitions that were on the losing side might become weakened as a result. If those intuitions also affected the outcome of some other moral dilemma, then the order in which the dilemmas were attended to might determine the conclusions that the individual reaches. Further history-dependence is introduced by the role of social effects, in which people affect the values of the people around them.

This raises interesting possibilities in introducing flexibility to the value learner. If there is no single correct set of final values that an individual's or a society's values might converge to, the value learner might be allowed to nudge the individual towards such a set of values whose implied preference ordering on the world would be the easiest to fulfill. Depending on the individual's values, they might allow this kind of a nudging, or they might have an aesthetic which preferred developing their values in some other direction, such as some ideal of an objective morality, wanting to have values that lead to doing most good for the world, or simply preferring independence and resolving any value conflicts purely by themselves, without any nudging from the AI. Similarly to an approval-directed agent only changing a person's brain if that was genuinely something that the person consented to, the AI here could use the person's existing values as a base for only nudging the person towards a direction that the current values endorsed being nudged towards (if any).

While this paper is not focused on the question of how to combine the conflicting values of different people, it is also interesting to notice that the flexibility in resolving value conflicts may be useful in trying to find a combination of values that as many people as possible could agree on. If such a set of globally most agreeable values could be found, different AI systems could then coordinate to direct the whole global population towards that set. This might look like something akin the "Coherent Blended Volition" (Goertzel and Pitt 2012) proposal, which seeks to "incorporat[e] the most essential elements of [people's] divergent views into a whole that is overall compact, elegant and harmonious". In this proposal, AI systems would be used to assist people in reaching an agreeable combination of their values. Given the great degree of flexibility allowed by the possibility of social feedback loops, in which directed changes to a culture's customs could potentially cause major changes to the values of that culture, one might be very optimistic about the possibility of reaching such an outcome.

9. Criteria for evaluating the framework

This proposed framework is worth evaluating from two distinct, though overlapping, angles. First, is it correct? Second, to the extent that it is correct, is it actually useful for solving the value learning problem?

The correctness angle suggests the following possible criteria:

1. Psychologically realistic. The proposed model should be compatible with that which we know about current human values. As a bare minimum, it should not make predictions about human behavior which would fail to correctly predict the behavior of most typical test subjects. *Motivation:* an agent seeking to model human values cannot be expected to get it right unless its assumptions are based on a correct model of reality.

2. Compatible with individual variation. The proposed model should be flexible enough to be able to take into account the full range of variation in human psychology. It should be able to adapt it to accurately represent the values (and thus behavior) of groups as differing as Western and non-Western (Henrich et al. 2010), autistic and non-autistic, and so on. *Motivation:*

psychological research often focuses on typical individuals and average tendencies, whereas a value learner should be capable of taking into account the values of everyone.

3. Testable. The proposed model should be specific enough to make clear predictions, which can then be tested. *Motivation:* vague models that do not make specific predictions are useless for practical design work.

4. Integrated with existing theories. The proposed definition model should, to as large an extent possible, fit together with existing knowledge from related fields such as moral psychology, evolutionary psychology, neuroscience, sociology, artificial intelligence, behavioral economics, and so on. *Motivation:* theoretical coherence increases the chances of the model being correct; if a related field contains knowledge about human values which does not fit together with the proposed model, that suggests that the model is missing something important.

A second set of criteria is suggested by the usefulness to the value learning problem. These overlap somewhat with the "correctness" criteria in that a correct model of human value would likely also fulfill the "usefulness" criteria. However, here more emphasis is put on how well-suited the model is for answering these kinds of questions in particular: it would be possible to have a model which was capable of answering these questions in principle, but burdensome and unlikely to find the right answers in practice.

5. Suited for exhaustively modeling different values. Human values are very varied, and include very abstract ones such as a desire for autonomy and an ability to act free from external manipulation. The details of when these values would be considered fulfilled may be highly idiosyncratic and specific to the mind of the person with that value, but the proposed model should still be able of incorporating that value. *Motivation:* again, a value learner should be capable of taking into account the values of everyone.

6. Suited for modeling internal conflicts and higher-order desires. People may be genuinely conflicted between different values, endorsing contradictory sets of them given different situations or thought experiments, and they may struggle to behave in a way in which they would like to behave. The proposed model should be capable of modeling these conflicts, as well as the way that people resolve them. *Motivation:* if a human is conflicted between different values,

7. Suited for modeling changing and evolving values. Human values are constantly evolving. The proposed model should be able to incorporate this, as well as to predict how our values would change given some specific outcomes. *Motivation:* an AI should be capable of noticing cases where we were about to do things that our future selves might predictably regret, and warn us about this possibility. (Yudkowsky 2004) A dynamic model of values also helps prevent "value lock-in", where an AI learns one set of values and then enforces that even after our values have shifted.

8. Suited for generalizing from existing values to new ones. The proposed model should be able to react to circumstances where either our understanding of the world, or the world itself, changes dramatically and forces a reconsideration of how existing values apply to the new situation. *Motivation:* technological and social change is constantly occurring, and often forces these kinds of re-evaluations, as discussed in section 2 earlier.

10. Evaluation

I will now evaluate my proposed framework in light of the above criteria.

1. Psychologically realistic. The framework is motivated by psychological research, particularly in the fields of moral psychology and emotion research. However, not all of the work in the said fields has yet been fully integrated to the framework, which may bias the implications of the framework. In particular, dual-process models of morality (Greene 2007, 2014), which also cover more non-emotional reasoning, are not yet a part of this framework. Dual-process models have made accurate predictions about less emotional people tending to make more "utilitarian" judgments in moral dilemmas, an outcome which the proposed framework would not have predicted. Additionally, while this framework has been developed based on psychological theories, neuroscientific evidence has not yet been considered in depth. With regard to the neuroscientific data, a specific shortcoming of the current framework is the possible existence of several different reinforcement learning mechanisms in the brain (Dayan 2011, Dayan & Berridge 2014), requiring further investigation to identify the extent to which the mechanisms hypothesized here are implemented in one system or several, and how those systems interact when it comes to questions of values and morality.

2. Compatible with individual variation. In the framework, differing values are hypothesized to emerge from individual differences (some of them which are caused by cultural differences) related to things that provide positive or negative rewards. This is compatible with a great degree of individual variation. For example, various differences in moral intuitions (due either to culture or something else) can be modeled as those intuitions causing different combinations of affect in response to the same situation, with this then leading to the same concepts being associated with different levels of affect for different individuals.

3. Testable. The framework is currently insufficiently specific to make novel predictions, which will be addressed in follow-up work.

4. Integrated with existing theories. The framework is currently moderately integrated with theories of reinforcement learning, moral psychology, and emotion research. However, there remains considerable room for further integration, and there are related fields such as sociology, which have not yet been addressed at all.

5. Suited for exhaustively modeling different values. In the framework, any concept that a human may have, either on a conscious or subconscious level, can be a value.

6. Suited for modeling internal conflicts and higher-order desires. Higher-order desires can be modeled to some extent: for example, a drug addict's desire to quit a drug might be modeled as negative affect around the concept of being a drug user, or as positive affect around the concept of quitting. However, the current framework does not fully explain the existence of internal conflicts and people engaging in actions which go against their higher-order desires. For this, further theoretical integration is needed with models such as ones positing separate mental modules optimizing for either short-term or long-term reward (Kurzban 2012). Such integration would help explain addictive behaviors as the modules optimizing for short-term reward "overpowering" the ones optimizing for long-term reward.

7. Suited for modeling changing and evolving values. As values are conceptualized as corresponding to a value function which is constantly updated and recomputed, the framework naturally models changing values.

8. Suited for generalizing from existing values to new ones. Section 8 discussed a possible way for a value learner to use this framework for generalizing existing values into new situations, by simulating a human in different situations and modeling the amount of reward obtained from those situations, as well as using existing values to guide this simulation.

To my knowledge, there have not been proposals for definitions of human values that would be relevant in the context of AI safety engineering. The one proposal that comes the closest is Sezener (2015). This paper takes an inverse reinforcement learning approach, modeling a human as an agent that interacts with its environment in order to maximize a sum of rewards. It then proposes a value learning design where the value learner is an agent that uses Solomonoff's universal prior in order to find the program generating the rewards, based on the human's actions. Basically, a human's values are equivalent to a human's reward function.

As a comparison, an evaluation of Sezener's proposal using the same criteria follows.

1. Psychologically realistic. Sezener's framework models humans as being composed of a reward mechanism and a decision-making mechanism, where both are allowed to be arbitrarily complex, so e.g. the reward mechanism could actually incorporate several distinct mechanisms. Thus a range of models, some of them more realistic than others, could be a part of the framework.

2. Compatible with individual variation. Because both the agent and the reward function are drawn from the space of all possible programs, Sezener's proposal is compatible with a vast range of individual variation.

3. Testable. Sezener's proposal is very general, and insufficiently specific to make novel predictions.

4. Integrated with existing theories. Various existing theories could in principle used to flesh out the internals of the reward function, but currently no such integration is present.

5. Suited for exhaustively modeling different values. Because the reward function is drawn from the space of all possible programs, any value that can be represented in a computational form can in principle be represented. However, because the simplicity of the program is the only prior probability used for weighting different programs, this may not sufficiently constrain the search space towards the values that would be the most plausible on other grounds.

6. Suited for modeling internal conflicts and higher-order desires. No specific mention of this is made in the paper. The assumption of a single reward function that assigns a single reward for every possible observation seems to implicitly exclude the notion of internal conflicts, with the agent always just maximizing a total sum of rewards and being internally united in that goal. It might be possible to represent internal conflict with the right kind of agent model, but again it seems unclear whether the prior probability used sufficiently constrains the search space.

7. Suited for modeling and changing and evolving values. Because the reward function is allowed to map the same observations to different rewards at different times, the framework is in principle capable of representing changing values. However, the same problem of finding the correct function remains an issue.

8. Suited for generalizing from existing values to new ones. There does not seem to be any obvious possibility for this in the model, except if the reward function that the value learner estimates happens to generalize rewards in the same way as a human would.

Overall, although both my and Sezener's frameworks represent useful progress towards a definition of human value, much work clearly remains to be done.

Acknowledgements

The last two sections were added after an anonymous reviewer pointed out the lack of an evaluation section. Daniel Dewey, Eren Sezener, J. Matias Kivikangas, and an anonymous blog commenter provided useful comments on the evaluation criteria.

This research was supported as part of the Future of Life Institute (futureoflife.org) FLI-RFP-AI1 program, grant #2015-144120.

References

Benthin, A., Slovic, P., Moran, P., Severson, H., Mertz, C. K., & Gerrard, M. 1995. Adolescent health-threatening and health-enhancing behaviors: A study of word association and imagery. *Journal of Adolescent Health*, 17(3), 143-152.

Berridge, K. C. 1996. Food reward: brain substrates of wanting and liking. *Neuroscience & Biobehavioral Reviews*, 20(1), 1-25.

Bengio, Y., Courville, A., & Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE Transactions* *on Pattern Analysis and Machine Intelligence*, 35(8), 1798-1828.

Carlson, S. M., Davis, A. C., & Leach, J. G. 2005. Less is more: executive function and symbolic representation in preschool children. *Psychological Science*, 16(8), 609-616.

Christiano, P. 2014. Model-free decisions. https://medium.com/ai-control/model-free-decisions-6e6609f5d99e

Cosmides, L., & Tooby, J. 2004. Evolutionary psychology and the emotions. In *Handbook of Emotions*, 91-115. New York: Guilford.

Dayan, P. 2011. Models of value and choice. In *Neuroscience of preference and choice: Cognitive and neural mechanisms*, 33-52. Amsterdam: Academic Press.

Dayan, P. & Berridge, K.C. 2014. Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective & Behavioral Neuroscience*, 14(2), 473-492.

Dewey, D. 2011. Learning what to value. In *Artificial General Intelligence,* 309-314. Springer Berlin Heidelberg.

Duhigg, C. 2012. *The power of habit: Why we do what we do in life and business*. New York: Random House.

Frankfurt, H. G. 1971. Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, 68(1), 5-20.

Gärdenfors, P. 2004. *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT press.

Goertzel, B. & Pitt, J. 2012. Nine Ways to Bias Open-Source AGI Toward Friendliness. *Journal of Evolution and Technology* 22(1), 116–131.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. 2012. Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology,* 47, 55-130.

Greene, J.D. 2007.Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, 11(8), 322-323.

Greene, J.D. 2014. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. New York: Penguin Books.

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

Haidt, J. 2012. *The righteous mind: Why good people are divided by politics and religion*. New York: Pantheon.

Hayes, S. C., & Smith, S. 2005. *Get out of your mind and into your life: The new acceptance and commitment therapy*. Oakland: New Harbinger Publications.

Henrich, J., Heine, S.J., & Norenzayan, A. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61-83.

Kurzban, R. 2012. *Why everyone (else) is a hypocrite: Evolution and the modular mind*. Princeton, NJ: Princeton University Press.

Lessig, L. 2004. *Free culture: How big media uses technology and the law to lock down culture and control creativity*. New York: Penguin.

Machery, E. 2010. Precis of doing without concepts. *Behavioral and Brain Sciences*, 33(2-3), 195-206.

Marinier, R., & Laird, J. E. 2008. Emotion-driven reinforcement learning. *Cognitive Science*, 115-120.

Muehlhauser, L., & Helm, L. 2012. The singularity and machine ethics. In *Singularity Hypotheses,* 101-126. Berlin: Springer.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G. Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Dharsan, K., Wiestra, D., Legg, S., & Hassabis, D. 2015. *Human-level control through deep reinforcement learning*. Nature, 518(7540), 529533.

Ng, A. Y., & Russell, S. J. 2000, June. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 663-670. San Fransisco, CA: Morgan Kaufmann Publishers.

Peters, E., & Slovic, P. 1996. The role of affect and worldviews as orienting dispositions in the perception and acceptance of nuclear Power. *Journal of Applied Social Psychology*, 26(16), 1427-1453.

Pryor, K. 1999. *Don't Shoot the Dog: The New Art of Teaching and Training*. New York: Bantam.

Rantala, M. J., & Marcinkowska, U. M. 2011. The role of sexual imprinting and the Westermarck effect in mate choice in humans. *Behavioral Ecology and Sociobiology*, 65(5), 859-873.

Roseman, I. J., & Smith, C. A. 2001. Appraisal theory: Overview, assumptions, varieties, controversies. In *Appraisal processes in emotion: Theory, methods, research. Series in affective science*, 3-19. New York, NY: Oxford University Press.

Russell, J. A. 2003. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145.

Russell, S. 2015. Will They Make Us Better People? http://edge.org/response-detail/26157

Ryan, R. M., & Deci, E. L. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68.

Scherer, K. R. 1999. Appraisal theory. *Handbook of cognition and emotion*, 637-663. Chichester, England: Wiley.

Scherer, K. R. 2009. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, 23(7), 1307-1351.

Schmidhuber, J. 2015. Deep learning in neural networks: An overview. *Neural Networks*, 61, 85-117.

Singh, S., Lewis, R. L., & Barto, A. G. 2009. Where do rewards come from? In *Proceedings of the Annual Conference of the Cognitive Science Society*, 2601-2606.

Singh, S., Lewis, R. L., Barto, A. G., & Sorg, J. 2010. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development* 2(2), 70-82.

Sezener, C.E. 2015. Inferring human values for safe AGI design. In *Proceedings of the 8th International Conference on Artificial General Intelligence, Berlin, Germany*, 152-155. Springer International Publishing.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. 2007. The affect heuristic. *European Journal of Operational Research*, 177(3), 1333-1352.

Soares, N. 2014. The Value Learning Problem. Tech. rep. Machine Intelligence Research Institute, 2014. URL: https://intelligence.org/files/ValueLearningProblem.pdf.

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

Tarleton, N. 2010. Coherent extrapolated volition: A meta-level approach to machine ethics. Berkeley, CA: Machine Intelligence Research Institute. https://intelligence.org/files/CEV-MachineEthics.pdf

Tindell, A. J., Smith, K. S., Berridge, K. C., & Aldridge, J. W. 2009. Dynamic computation of incentive salience: "wanting" what was never "liked". *The Journal of Neuroscience*, 29(39), 12220-12228.

Tooby, J., & Cosmides, L. 1990. The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology*, 11(4), 375-424.

Tooby, J., & Cosmides, L. 1995. The psychological foundations of culture. *The adapted mind: Evolutionary psychology and the generation of culture*, 19-136. Oxford University Press.

Tversky, A., & Kahneman, D. 1981. The framing of decisions and the psychology of choice. *Science*, 211(4481), 453-458.

Von Neumann, J., & Morgenstern, O. 1953. *Theory of games and economic behavior.* Princeton, NJ: Princeton University Press.

Wessinger, C. M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., & Rauschecker, J. P. 2001. Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *Journal of Cognitive Neuroscience*, 13(1), 1-7.

Yudkowsky, E. 2004. *Coherent extrapolated volition.* The Singularity Institute. https://intelligence.org/files/CEV.pdf