Literature Review: What AI Safety Researchers Have Written About the Nature of Human Values

Alexey Turchin

Digital Immortality Now

Foundation Science for Life Extension

[alexeiturchin@gmail.com](mailto:alexeiturchin@gmail.com)

David Denkenberger

Alliance to Feed the Earth in Disasters (ALLFED), Fairbanks, AK, 99775, USA; david@allfed.info

University of Alaska Fairbanks, Fairbanks, AK, 99775, USA; ddenkenberger@alaska.edu

**Abstract**: The field of artificial general intelligence (AGI) safety is quickly growing. However, the nature of *human values,* with which future AGI should be aligned, is underdefined. Different AGI safety researchers have suggested different theories about the nature of human values, but there are contradictions. This article presents an overview of what AGI safety researchers have written about the nature of human values, up to the beginning of 2019. 22 authors were overviewed, and some of them have several theories. A theory classification method is suggested, where the theories are judged according to the level of their complexity and behaviorists-internalists scale, as well as the level of their generality-humanity. We suggest that a multiplicity of well-supported theories means that the nature of human values is difficult to define, and some meta-level theory is needed.

[1. Introduction 2](#_Toc7173875)

[2. Principles of theories classification 3](#_Toc7173876)

[3. List of researchers and their theories 4](#_Toc7173877)

[Eliezer Yudkowsky 4](#_Toc7173878)

[Nick Bostrom 5](#_Toc7173879)

[Kaj Sotala 5](#_Toc7173880)

[Stuart Armstrong 5](#_Toc7173881)

[Gopal Sarma and Nick Hay 7](#_Toc7173882)

[Luke Muehlhauser 8](#_Toc7173883)

[Can Eren Sezener 9](#_Toc7173884)

[John Maxwell 9](#_Toc7173885)

[Bill Hibbard 10](#_Toc7173886)

[Paul Christiano 10](#_Toc7173887)

[Jan Kulveit 10](#_Toc7173888)

[Sabrina Kavanagh, Erin Linebarger and Nandi Schoots 11](#_Toc7173889)

[Scott Alexander 11](#_Toc7173890)

[Eric Drexler 11](#_Toc7173891)

[Roman Yampolskiy 11](#_Toc7173892)

[Gordon Worley III 12](#_Toc7173893)

[Rohinm Shah 12](#_Toc7173894)

[Wei Dai 12](#_Toc7173895)

[Caspar Oesterheld 13](#_Toc7173896)

[Alexandra Surdina 13](#_Toc7173897)

[Steven Umbrello 13](#_Toc7173898)

[Vanessa Kosoy 13](#_Toc7173899)

[Robin Hanson 14](#_Toc7173900)

[4. Discussion 14](#_Toc7173901)

[5. Conclusions 14](#_Toc7173902)

[Acknowledgements 15](#_Toc7173903)

# 1. Introduction

The main current approach to artificial general intelligence (AGI) safety is AGI alignment, i.e., the idea that future advanced AGI will learn human values, and thus its utility function will be aligned with human values. This relation consists of three elements: presenting the goal system to AGI, the value learning procedure and “human values.” All three are mutually dependent as different ideas about the nature of human values assume different ways of their presentation and learning. Typically, AGI alignment research is presented in the form of mathematical procedures which are intended to provide correct ways of learning human values (e.g. inverse reinforcement learning).

However, in order for such procedures to be correct, “human values” need to be defined in some way. This need has attracted several researchers who have tried to defined human values having AGI alignment in mind. The goal of this article is to collect and compare different theories of human values according to AGI safety researchers, with the hope that this will help to choose or create the best theory. Also, comparing such theories may provide some insight into the nature of human values.

It is clear to most AGI safety researchers[[1]](#footnote-1) that the idea of “human values” is underdefined, and this concept should be additionally formalized before it can be used in (mostly mathematical) models of AGI alignment. In other words, “AGI-conscious” human value theories are those which were created specifically in order to help AGI alignment research. Most existing psychological theories of human values are verbal, informal and underdefined, so they require some adaptation before they may be applied to AGI safety.

In many cases, the theory of human values cannot be distinguished from the ways in which values are expected to be extracted by a future AGI, for example, by approval-directed AGI. Some researchers (notably, Armstrong) have also presented different theories; providing a link to the researcher is not intended as a claim that the researcher currently adheres to exactly this theory.

As most of the discussion is happened in the Internet, all links on sources are clickable.

This work is a part of longer series about human values, which includes “[Dangerous value learners](https://www.lesswrong.com/posts/hzEaasJyQsutYDNfN/possible-dangers-of-the-unrestricted-value-learners)”, “AGI alignment problem: ’Human values’ idea is built upon many assumptions” (to be published) and “Approaches to AGI safety which are not based on the idea of human values” (to be published).

In section 2, general principles of classifications of human values are explored. In section 3, a list of 21 author and short overview of their ideas about human values is presented.

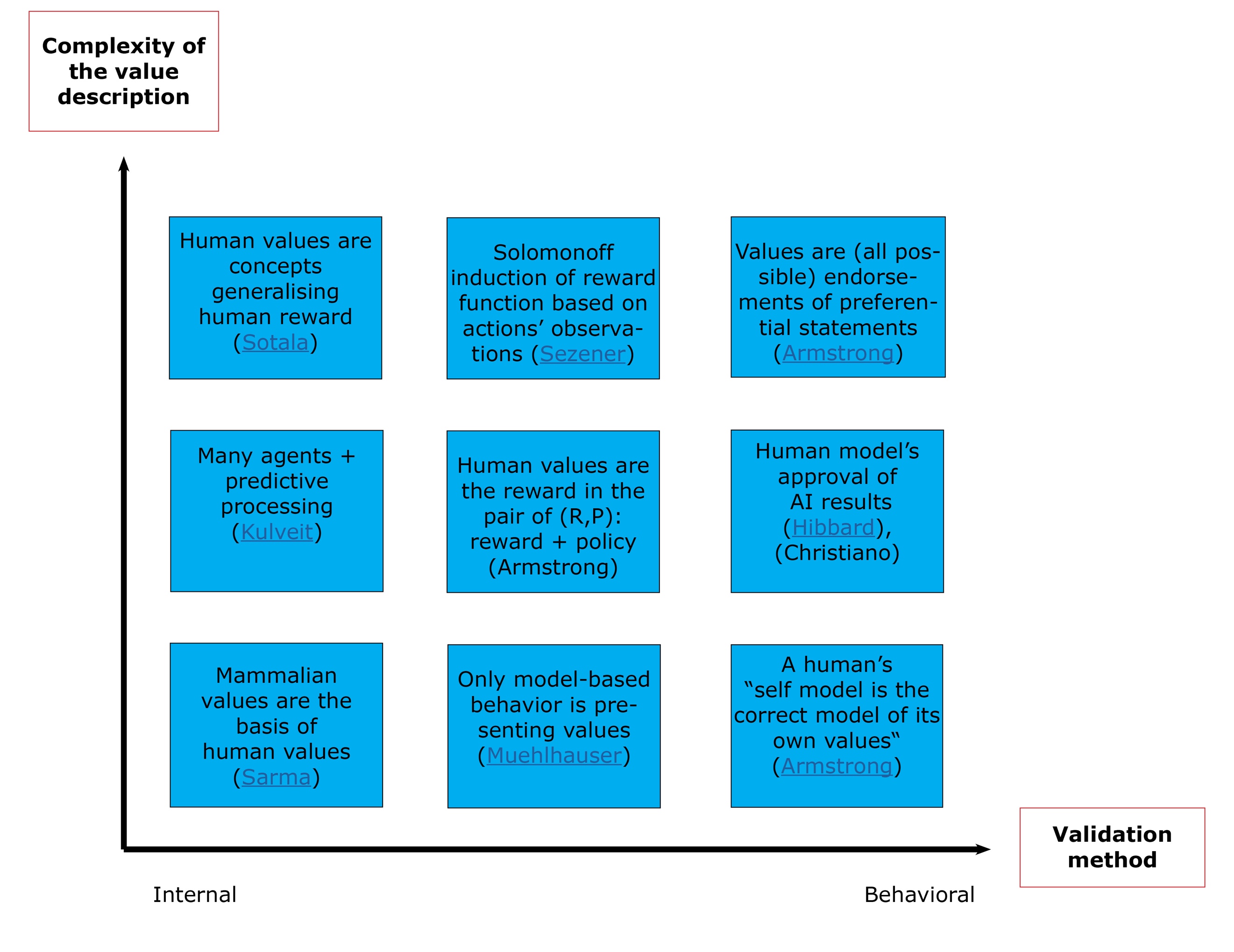
# 2. Principles of theories classification

In Figure 1, all current theories of human values are classified according to two main parameters:

1. *Complexity of the values’ description* (not the complexity of the theory): Some theories of human values assume that human values are very simple, e.g., that there are only two drives: survival and reproduction; or only one desire, that of maximizing pleasure; or only nine basic emotions. Other theories suggest that human values are very complex, e.g. they are a matrix of association between all concepts and rewards.
2. *The level of “behaviorism”*: Most theories are attracted to one of two poles: “internalist” theories, which assume that values actually exist, but are hidden inside the human brain, and “behaviorist” theories, which assume that values are only appear in human behavior (like approval or choice). Behaviorist theories of human values (generally) integrate the values and the method of their extraction (e.g. approval-directed AGI). In contrast, internalist theories typically hold that values exist separately of the ways they are learned.

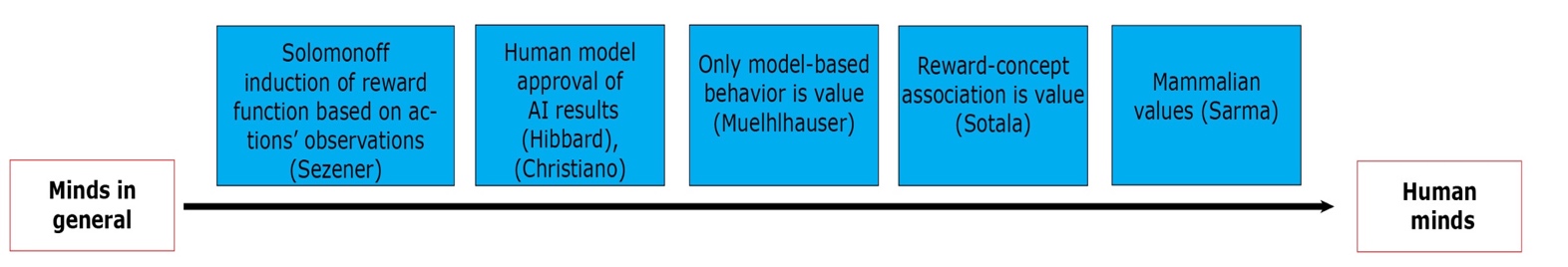
These two parameters are not strict, but serve as a useful instrument to orientate ourselves within the field of all possible theories.

*Figure 1. Different real and possible theories about the nature of human values in AI safety (*[*pdf*](http://immortality-roadmap.com/val.pdf)*)*



Another way to classify theories about human values is according to the level of abstractness: some theories could be applied to any possible mind and thus do not take any inputs from human psychology or neurophysiology. Such theories are computationally complex and may still contain hidden assumptions about some properties of human minds: stability, unity, consistency, etc. Human-centered theories depend on existing theories of human minds. This distinction is illustrated in Figure 2.

*Figure 2. Mind in general and human-mind theories of values*



# 3. List of researchers and their theories

## Eliezer Yudkowsky

Yudkowsky is a pioneer in the field of AGI safety, and among many other things he introduced the idea of the “complexity of values”; essentially, that any short verbal description can’t grasp the complexity of outcomes that we actually want. He summarized his critique of simple wishes as the correct presentation of desirable outcomes in “[Complex Value Systems are Required to Realize Valuable Futures](https://intelligence.org/files/ComplexValues.pdf)”. In the same article, he also introduced the concept of “fragility of values”—e.g., if one digit in a phone number is wrong, the call will go to a completely different person.

Another important contribution from Yudkowsky is the idea of *Coherent Extrapolated Volition* (CEV). He wrote in his “Complex Values” article: “We might try to define normativity not by our immediate current desires but by our *reflective equilibria*, what we would want in the limit of perfect knowledge, the ability to consider all options and arguments, and perfect self-knowledge without unwanted weakness of will (failure of self-control).” More links at his writings about values can be found at [LessWrong Wiki](https://wiki.lesswrong.com/wiki/Complexity_of_value.).

The *Arbital* page “[Value](https://arbital.com/p/value_alignment_value/),” which is likely written by Yudkowsky, starts with a definition: “In the context of value alignment as a subject, the word 'value' is a speaker-dependent variable that indicates our ultimate goal - the property or meta-property that the speaker wants or 'should want' to see in the final outcome of Earth-originating intelligent life”. Obviously, most psychologists, and people in general, define “human values” differently, as most people do not care about the remote future of humanity in their day-to-day preferences.

The Arbital article on values also presents a list of possible points of view on the nature of human values, which, in short, is:

* *Reflective equilibrium*. What one should want “given more factual knowledge, time to consider more knowledge, better self-awareness, and better self-control”.
* *Standard desires*. “An object-level view that identifies value with qualities that we currently find very desirable, enjoyable, fun, and preferable, such as [Frankena's list of desiderata](https://arbital.com/p/frankena_goods/)”.
* *Immediate goods*. “E.g., ‘Cure cancer’”
* *Deflationary moral error theory*. “This mostly ends up at an 'immediate goods' theory in practice, plus some beliefs relevant to the value selection debate”
* *Simple purpose*. “Value can easily be identified with X, for some X”.

## Nick Bostrom

Bostrom does not offer a preferred theory of human values, and suggests several instruments for AGI alignment that do not require a direct definition of human values and, in some sense, are similar to [Coherent Extrapolated Volition](https://arbital.com/p/cev/) (CEV). One such idea is the implementation of a [virtual parliament](http://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html) of moral theories, where many virtual human minds are held in a simulation until they come to an agreement regarding the most correct presentation of human values.

## Kaj Sotala

Sotala analyzed human values in “[Defining Human Values for Value Learners](http://intelligence.org/files/DefiningValuesForValueLearners.pdf)” (a discussion of which can be found [here](https://www.lesswrong.com/posts/2Rai96XMwL83d27rc/paper-link-defining-human-values-for-value-learners).) In this article, he listed several problems of the simple model of human values as a utility function:

* “The utility function model of value has difficulty dealing with internal conflicts and higher-order preferences”
* “The utility function model of value ignores the person’s internal experience.”
* “The utility function model of value does not model changing values.”
* “The utility function model of value does not give a way to generalize from our existing values to new ones.”

To solve this problem, he suggests the following definition: “…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 a least partially encoded in the affect of the various concepts.”

At the end of article, Sotala suggests useful criteria to estimate validity of any theory of human values. He outlines that such a theory should be:

* Psychologically realistic
* Compatible with individual variation
* Testable
* Integrated with existing theories
* Suited for exhaustively modeling different values
* Suited for modeling internal conflicts and higher order desires
* Suited for modeling changing and evolving values
* Suited for generalizing from existing values to new ones

In 2019, Sotala explored [Internal family system](https://www.lesswrong.com/posts/5gfqG3Xcopscta3st/building-up-to-an-internal-family-systems-model) theory as a possible theory of human motivation. It is part of his sequence “[Multi-agent theory of mind](https://www.lesswrong.com/s/ZbmRyDN8TCpBTZSip)”, which also includes a post with a [review](https://www.lesswrong.com/posts/x4n4jcoDP7xh5LWLq/book-summary-consciousness-and-the-brain) of a book about consciousness as a global workplace, where different internal parts are interacting.

## Stuart Armstrong

Armstrong has written [many articles and posts](https://agentfoundations.org/item?id=601) in which he seeks to define the nature of human values and addresses popular misconceptions.

One such contribution is his proof that it is impossible to distinguish between “values” and “policy” based only on observation of behavior: “The observed behavior can, in principle, be decomposed into two components: a reward function and a planning algorithm, both of which have to be inferred from behavior. [I present] a No Free Lunch theorem, showing that, without making `normative' assumptions beyond the data, nothing about the human reward function can be deduced from human behavior.” However, he has probably assumed that “human values” = “human reward function”, and that a human reward function is approximately the same as a rational agent’s reward function.

But the problem is even wider than distinguishing between values and policy: we rather arbitrarily call some part of the process in the human mind “values,” which should be preserved, and call other parts “biases,” “policy,” and “knowledge,” concepts which could be freely changed and do not have intrinsic values.

Armstrong looks at the “ontological nature” of human values, i.e. the question of whether they actually exist, in “[Learning values, or defining them](https://www.lesswrong.com/posts/RHvseCkfrYzoHJj7M/learning-values-or-defining-them).” In it, he wrote: “Many attempts at learning human values are framed as "humans have an underlying true reward R, and here is procedure P for determining it,” and “[e]ven if the moral realists are right, and there is a true R, thinking about it is still misleading. Because there is, as yet, no satisfactory definition of this true R, and it's very hard to make something converge better onto something you haven't defined.”

Other posts by Armstrong on the indefinability of human values include “[Human can be assigned any values whatsoever](https://www.alignmentforum.org/s/4dHMdK5TLN6xcqtyc/p/ANupXf8XfZo2EJxGv),” and an article which covers all of these ideas, “[Occam's razor is insufficient to infer the preferences of irrational agents](https://arxiv.org/abs/1712.05812).” In the latter [post](file:///C:\Users\ddenkenberger\Downloads\Humans%20can%20be%20assigned%20any%20values%20whatsoever…), he sums up his findings: “Humans have no values… nor do any agent. Unless you make strong assumptions about their rationality [sic]. And depending on those assumptions, you get humans to have any values.” Then, he proceeds with a toy model which shows that even Kolmogorov complexity considerations cannot help find the true values. It seems that Armstrong also uses “human values” and “human reward function” interchangeably, which is not necessarily true unless one redefines “human reward” in a way which is different from human reward center activation and experience of qualia of pleasure.

Armstrong argues for the unknowability of human values: “Humans don’t know their own values (sub-issue: humans know their values better in retrospect than in prediction)” in his “[List of three problems and different approaches to solutions](https://agentfoundations.org/item?id=1388).” That post also offers a nice table of approaches to solving the AGI safety problem.

In the blogpost “[Resolving human values, completely and adequately](https://www.lesswrong.com/posts/Y2LhX3925RodndwpC/resolving-human-values-completely-and-adequately),” he suggests the creation of a model of human values which is neither too abstract nor too narrow, but is adequate to escape disastrous outcomes. He suggests that such an adequate definition is “… all the value/preference/reward statements that H[umans] might agree to, more or less strongly.”

He also introduced the important idea of normative assumption in “[Normative assumptions: regret](https://www.lesswrong.com/posts/Fg83cD3M7dSpSaNFg/normative-assumptions-regret).” In a nutshell, the idea is that any model of “human values” has normative assumptions, i.e., assumptions about which part of the human motivational system is more important than the others. He then suggested that a feeling of regret be used to determine true values: “In a previous post, I presented a model of human rationality and reward as pair (p, R) [where p is “our (ir)rational planning algorithm (called a planner)” and R is reward]…[n]ormative assumptions are simply assumptions that distinguish between two pairs (p, R) (p', R') that lead to the same policy: p(R)=p'(R').” He continued, “[h]ow do we, as humans, define our own rationality and reward? Well, one strong way seems to be through our feeling of regret...[i]f we take ‘feelings and expressions of regret encode true reward information’ as a normative assumption, then this restricts the number of (p, R) pairs that are compatible with such an assumption.”

The case about regret could be made stronger if it incorporated concepts from the existing psychological literature. It is also possible to imagine a “hyper-regret disorder,” in which a person will regret all of his/her choices; in that case, regret would be non-informative about preferences.

In another post, “[Beyond algorithmic equivalence: self-modelling](https://www.lesswrong.com/posts/kmLP3bTnBhc22DnqY/beyond-algorithmic-equivalence-self-modelling),” Armstrong defines “human reward” as follows: “we can *define* the reward of H[uman], as the reward that H models itself as having.” However, people may have a preference against value extraction and may deny the correctness of extracted values. In addition, most humans don’t have anything near a complete model of their own values; most humans are not that introspective.

In the post “[Humans are not agents: short vs. long term](https://agentfoundations.org/item?id=1515),” he looks at an example of logically contradictory preferences about life expectancy as an example of contradicting values. He suggests the following example: “Imagine a human who has a preference to not live beyond a hundred years. However, they want to live to next year, and it’s predictable that every year they are alive, they will have the same desire to survive till the next year.”

In 2019, he wrote a post [A theory of human values](https://www.lesswrong.com/posts/qezBTig6p6p5xtL6G/a-theory-of-human-values). In it he suggested that value learning should consist of three steps: identifying preferences, synthesis of a utility function based on them and checking that this function is not terribly wrong.

## Gopal Sarma and Nick Hay

Sarma and Hay state in their article “[Mammalian value systems](https://arxiv.org/abs/1607.08289)” that “[an] agent utilizing Inverse Reinforcement Learning or Bayesian Inverse Planning will learn and refine its model of human values by observing our behavior must begin with some very rough or approximate initial assumptions about the nature of the values it is trying to learn.” They then suggest the use of the phrase “mammalian value system” as a starting point, taking into account later cultural effects on values evolution. They suggest the description of the basic mammalian value system be based on the work of Panksepp and Biven who “categorize the informal list given above into seven motivational and emotional systems that are common to mammals: seeking, rage, fear, lust, care, panic/grief, and play.” Sarma and Hay then introduce the term “neural correlate of value” which is, according to them, some subcortical areas which serve as a mechanism for these seven basic mammalian emotions. That article concludes: “we argue that what we colloquially refer to as human values can be informally decomposed into 1) mammalian values, 2) human cognition, and 3) several millennia of human social and cultural evolution.”

In “[AI Safety and Reproducibility: Establishing Robust Foundations for the Neuropsychology of Human Values](https://link.springer.com/chapter/10.1007/978-3-319-99229-7_45),” Sarma and Hay point out that the reproducibility crisis in psychology makes identification of the correct theory of human values difficult, but urgent actions to that end are necessary to ensure AGI safety.

The newest (as of 2019) article by Sarma, Safron and Hay is “[Integrative Biological Simulation, Neuropsychology, and AI Safety](https://arxiv.org/abs/1811.03493),” in which they suggest creation of better biological models of animal minds to develop a better understanding of the nature of motivation.

The idea of some “prior” for learning human values, which could be relatively easily learned from a non-human source, is particularly promising. This is because many learning strategies are based on Bayesian methods, which need some prior models to start working. This will make the learning process safer as any animal will probably oppose its brain being dissected or other non-ethical methods of value learning (except maybe after natural death). However, the choice of seven main emotional-motivational mammalian traits may seem arbitrary, as there are probably other models of animal motivation. Also, why mammalian, but not apes or vertebrates?

## Luke Muehlhauser

Muehlhauser wrote in the post “[The Human's Hidden Utility Function (Maybe)](https://www.lesswrong.com/posts/fa5o2tg9EfJE77jEQ/the-human-s-hidden-utility-function-maybe)” that the existence of a human utility function does not satisfy axioms of rationality. He also wrote that humans have three value systems: model-based, model-free (repetition of a successful action) and Pavlovian (described as more like unconditional reflexes). He suggests that only the first type is “good,” and should be used as a basis for CEV.

In the post “[Do Humans Want Things](https://www.lesswrong.com/posts/nBdaTGoDAYxHePSDa/do-humans-want-things),” he wrote that human choice depends on how choices are framed, but does not depend on the values, as shown by Kahneman and Tversky: “as far as we can tell, our behavior is often not determined by our wanting a particular state of affairs, but by how our options are framed.”

In the post “[The Neuroscience of Desire](https://www.lesswrong.com/posts/48DTJkBH58JbBNSFH/the-neuroscience-of-desire),” Muehlhauser starts with the observation (supported by a large bibliography) that decision-making in economics and computer science starts with integrating many dimensions into one scalar parameter, and then comparing such parameters for different options. He suggests that it appears the human brain does the same: “More than a dozen studies show that the subjective utility of different goods or actions are encoded on a common scale by the ventromedial prefrontal cortex and the striatum in primates (including humans).”

Then he looks at the neuroscience of choice: “Once a common-currency valuation of goods and actions has been performed, how is a choice made between them? Evidence implicates (at least) the lateral prefrontal and parietal cortex in a process that includes neurons encoding probabilistic reasoning. Interestingly, while valuation structures encode absolute (and thus transitive) subjective value, choice-making structures ‘rescale these absolute values so as to maximize the differences between the available options before choice is attempted,’ perhaps via a normalization mechanism like the one discovered in the visual cortex.”

In the post “[A Crash Course in the Neuroscience of Human Motivation](https://www.lesswrong.com/posts/hN2aRnu798yas5b2k/a-crash-course-in-the-neuroscience-of-human-motivation),” he wrote a rather long review of different theories of human motivation, beginning from “folk theory” and economics. The article starts with a rather bold statement: “But now, neuroscientists are directly measuring the neurons whose firing rates encode value and produce our choices.”

In an article by Muehlhauser and Helm, “[The singularity and machine ethics](https://intelligence.org/files/IE-ME.pdf),” they demonstrate that some (and probably all) known moral theories are unsafe if installed in a powerful optimizer. In Section 5.1 of that article they state and discuss the theme that “[h]umans don’t know their values,” based on an experiment in which participants explained preference to faces the participant didn’t choose. They state that “[c]ognitive science suggests instead that our knowledge of our own desires is just like our knowledge of others’ desires: inferred and often wrong.”

Further, they argue against the idea that human beings are rational utility maximizers:

“Ever since M. Friedman (1953), economists have insisted that humans only behave ’as if’ they are utility maximizers, not that humans actually compute expected utility and try to maximize it. It was a surprise, then, when neuroscientists located the neurons in the primate brain that encode (in their firing rates) the expected subjective value for possible actions in the current “choice set.” Several decades of experiments that used brain scanners and single neuron recorders to explore the primate decision-making system have revealed a surprisingly well-understood reduction of economic primitives to neural mechanisms; for a review see Glimcher (2010). To summarize: the inputs to the primate’s choice mechanism are the expected utilities for several possible actions under consideration, and these expected utilities are encoded in the firing rates of particular neurons. Because neuronal firing rates are stochastic, a final economic model of human choice will need to use a notion of “random utility,” as in McFadden (2005) or Gul and Pesendorfer (2006).”

They then look at neurological mechanisms of values: “Recent studies reveal the complexity of subjective values in the brain. For example, the neural encoding of human values results from an interaction of both ‘model-free’ and ‘model-based’ valuation processes.”

They explore the complexity of relation between personal preferences and choices: “...may be that each human being contains something like a ‘hidden’ utility function (within the model-based valuation system) that isn’t consistently expressed in behavior because choice is also partly determined by other systems whose valuations we wouldn’t reflectively endorse because they are ‘blind’ and ‘stupid’ compared to the more sophisticated goal-directed model-based valuation system.”

## Can Eren Sezener

Sezener, in the article “[Inferring human values for safe AGI design](http://agi-conf.org/2015/wp-content/uploads/2015/07/agi15_sezener.pdf),” suggested that human values are an arbitrary complex reward function.

Sezener’s main idea is the use of Solomonoff induction to find the simplest combination of two programs, one of which encodes an agent’s reward function, and the other which encodes the agent itself, based on an observable sequence of actions and observations. This is similar to Armstrong’s approach of presenting human as (p, R) a planning algorithm and reward pair, and then using complexity considerations to find the simplest such pair that explains observable behavior.

Sezener also wrote about hidden assumptions in inverse reinforcement learning (IRL): “Soares [7] suggests using methods similar to IRL for learning human values. However, the current IRL methods are limited and cannot be used for inferring human values because of their long list of assumptions. For instance, in most IRL methods the environment is usually assumed to be stationary, fully observable, and sometimes known; the policy of the agent is assumed to be stationary and optimal or near-optimal; the reward function is assumed to be stationary as well; and the Markov property is assumed. Such assumptions are reasonable for limited motor control tasks such as grasping and manipulation; however, if our goal is to learn high-level human values, they become unrealistic”.

We view the main problems of Sezener’s approach as his assertions that:

1. behavior and only behavior is the correct representation of a human reward function (what about unconscious or parasitic behavior?);

2. reward function = values;

3. the model ignores internal contradictions;

4. the model is incomputable, as it is based upon incomputable [Hutter’s AIXI](http://www.hutter1.net/ai/aixigentle.htm)

5. there is an assumption of simplicity of both reward function and agent, which provides (in Armstrong’s terms) a “free lunch”.

## John Maxwell

Maxwell wrote in his post “[Friendly AI through Ontology Autogeneration](https://medium.com/@pwgen/friendly-ai-through-ontology-autogeneration-5d375bf85922)” that: “If an AI is to be Friendly, it must operate based on an ontology that’s capable of expressing our values,” and, “[r]egardless of the ontology autogeneration algorithm that’s chosen, it’s almost certain that the initial autogeneration will either (a) capture human values with insufficient fidelity or (b) contain so many concepts that finding human values among them will be its own project.”

In other words, Maxwell demonstrated that values cannot exist without ontology and are strongly connected with it. Even if AGI is used to generate ontology, it may not make the task of searching values simple.

## Bill Hibbard

Hibbard wrote in “[Avoiding Unintended AI Behaviors](https://intelligence.org/files/UnintendedBehaviors.pdf),” that in order to evaluate policy an agent “can simply *ask model* human d to express a utility value between 0 and 1” for the policy. This could be called a “counterfactual approval by human model,” which evaluates all possible outcomes of actions. In Hibbard’s case, safe AGI consists of two levels: the first creates a model of the world (which includes all humans and their ways of behaving or reacting), and the second calculates how humans in the model will react to possible future histories.

Sezener is critical of Hibbard: “However, a shortcoming of this approach is that what human models say they value and what they value can still be different.”

## Paul Christiano

Christiano explored many ideas which could serve as a proxy for human values. One is the use of human approval (or hypothetical human approval) of AGI actions, summed up in the following quote: “Estimate the expected rating Hugh [human] would give each action if he considered it at length. Take the action with the highest expected rating” from the post “[Approval-directed agents](https://ai-alignment.com/model-free-decisions-6e6609f5d99e).” In that case, the idea of “human values” is ignored and replaced with the much more measurable “human approval.” There is a lengthy discussion about the robustness and scalability of this approach and its vulnerability to edge cases like wireheading (hacking the reward function).

## Jan Kulveit

In “[Multi-agent predictive minds and AI alignment](https://www.lesswrong.com/posts/3fkBWpE4f9nYbdf7E/multi-agent-predictive-minds-and-ai-alignment),” Kulveit combines predictive processing and a multi-agent model of cognition to create a model of the human mind, and then uses it to create approaches to AGI alignment.

In his framework:

“…how do motivations and ’values’ arise? The guess is, in many cases something like a ’subprogram’ is modelling/tracking some variable, ’predicting’ its desirable state, and creating the need for action by ’signalling’ prediction error. Note that such subprograms can work on variables on very different hierarchical layers of modelling - e.g. tracking a simple variable like ’feeling hungry’ vs. tracking a variable like “social status”.

Such sub-systems can be large; for example, tracking “social status” seems to require lot of computation. Later, Kulveit states that “In this model, it is hardly possible to disentangle ‘beliefs’ and ‘motivations’ (or values).”

Kulveit then suggests that the human mind could be modeled by a system consisting of many subagents, larger than Minsky’s small agents, described in the [*Society of Mind*](http://www.acad.bg/ebook/ml/Society%20of%20Mind.pdf), but smaller than a psychologist’s human-like subpersonalities, which may or may not be aligned with each other. If internal subagents are not aligned with each other, it will result in contradictory behavior. This has different implications for different approaches to AGI alignment, of which four are listed by Kulveit:

1. Alignment with the outputs of the generative models, without querying the human. This includes for example proposals centered around approval. In this case, generally only the output of the internal aggregation has some voice.
2. Alignment with the outputs of the generative models, with querying the human. This includes for example [CIRL](https://arxiv.org/abs/1606.03137) and similar approaches. The problematic part of this is, by carefully crafted queries, it is possible to give voice to different sub-agent systems ….
3. Alignment with the whole system, including the human aggregation process itself. This could include, for example, some deep neural network based black box trained on a large amount of human data, predicting what the human would want (or approve).
4. Adding layers of indirection to the question, such as defining alignment as a state where the*“AGI is trying to do what the human wants it to do.”*

## Sabrina Kavanagh, Erin Linebarger and Nandi Schoots

Kavanagh, Linebarger and Schoots from the “Human preferences” team at the 2018 AI Safety Camp 2 in Prague wrote “[Acknowledging Human Preference Types to Support Value Learning](https://www.lesswrong.com/posts/mSPsyEwaymS74unND/acknowledging-human-preference-types-to-support-value).” They start with the idea that internal conflicts in humans could be explained by different preference types, and explore three such types: liking, wanting, and approving. They list all eight possible combinations of these three preferences and demonstrate that each corresponds to some type of behavior.

They later state: “Liking, wanting and approving are for the most part hidden processes. They are not directly observable, but they influence observable behaviors. As a proxy for liking we propose to use facial expressions, body language or responses to questionnaires. Although a cognitive scan may be the most accurate proxy for liking, there is evidence to suggest both that facial expressions and body language are indicators of pleasure and pain [Algom et al. 1994] and that they can be classified well enough to make them technically feasible proxies [Giorgiana et al. 2012]. The observable proxy of wanting is revealed preferences. We propose to encode a proxy for approval via stated preferences.”

## Scott Alexander

Alexander wrote “[To what degree do we have goals](https://www.lesswrong.com/posts/ePA4NDzZkunz98tLx/to-what-degree-do-we-have-goals),” in which he explores the idea that only conscious personal values should be taken into account, and unconscious ones should be ignored.

In the post “[Would Your Real Preferences Please Stand Up](https://www.lesswrong.com/posts/z3cTkXbA7jgwGWPcv/would-your-real-preferences-please-stand-up)?” he argues that in many cases our declared preferences are just “social signaling.”

## Eric Drexler

Drexler, in “[Reframing superintelligence: Comprehensive AI Services as General Intelligence](https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf),” plainly states that: “It seems impossible to deﬁne human values in a way that would be generally accepted” (p. 152).

## Roman Yampolskiy

Yampolskiy is also skeptical that human values can be formalized: “human values are inconsistent and dynamic and so can never be understood/programmed into a machine. Suggestions for overcoming this obstacle require changing humanity into something it is not, and so by definition destroying it” ([Roman Yampolskiy on AI Safety Engineering](https://intelligence.org/2013/07/15/roman-interview/?fbclid=IwAR2AOyUc0JEySlgclbwoNcYzL7RzeUnyOFerHqQEgIEWGfuzkGCZQYCIWOg)).

In the article “[Personal Universes: A Solution to the Multi-Agent Value Alignment Problem](https://arxiv.org/pdf/1901.01851.pdf)” he suggests a solution to escape the difficult value aggregating problem in a personal universe which would be “…optimally and dynamically adjusting to align their [humans] values and preferences.” In other words, future superintelligent AI would create a personal universe simulation of a most interesting world for every person. This will escape the need to aggregate different values of different people, but still requires selecting the most relevant values inside a given person.

## Gordon Worley III

Worley wrote about the need for artificial general intelligence (AGI) alignment to take into account the mental phenomena of consciousness in “[Formally Stating the AI Alignment Problem](https://mapandterritory.org/formally-stating-the-ai-alignment-problem-fe7a6e3e5991?fbclid=IwAR0sLRdSjZ9_LjPymvC-Y8gMZa89PvgKXAQPSm3ZfuUF3N_ngyNemjxWk_g).” He further discusses his opinion that human values should be described using the instruments of phenomenology (a field of philosophy which studies internal objects inside consciousness).

In a personal communication, he clarifies his position: “…my view is that values are inextricably tied to the existence of consciousness because they arise from our self-aware experience. This means I think values have a simple, universal structure and also that values are rich with detail in their content within that simple structure. This view also necessarily suggests values are not fully discoverable via behavioral methods and that there is always a hidden, internal part that may not even be accessible by the agent themselves.”

## Rohinm Shah

Shah wrote about the relation between real and proxy reward functions in “Active Inverse Reward Design,” saying that “[i]nverse reward design (IRD) is a preference inference method that infers a true reward function from an observed, possibly misspecified, proxy reward function.” “In this paper, we actively select the set of proxy reward functions available to the designer”.

He wrote a [sequence](https://www.lesswrong.com/s/4dHMdK5TLN6xcqtyc) about “narrow value learning” and another one about “ambitious value learning”. The main idea is that: “*Ambitious value learning* aims to achieve superhuman performance by figuring out the underlying latent ’values’ that humans have, and evaluating new situations according to these values”. This is opposed to “*narrow value learning*, which produces behavior that we want in some narrow domain, without expecting generalization to novel circumstances. The simplest form of this is imitation learning…”

## Wei Dai

Dai wrote “[A general model of safety-oriented AI development](https://www.lesswrong.com/posts/idb5Ppp9zghcichJ5/a-general-model-of-safety-oriented-ai-development)” where AGI creation is an interactive process within a human-AI team:

“Start with a team of one or more humans (researchers, programmers, trainers, and/or overseers), with access to zero or more AIs (initially as assistants). The human/AI team in each round develops a new AI and adds it to the team, and repeats this until maturity in AI technology is achieved. Safety/alignment is ensured by having some set of safety/alignment properties on the team that is inductively maintained by the development process.”

Dai also wrote about possible [unsafety](https://www.alignmentforum.org/posts/vbtvgNXkufFRSrx4j/three-ai-safety-related-ideas) of human values: “Many AI safety problems are likely to have counterparts in humans. AI designers and safety researchers shouldn't start by assuming that humans are safe”.

## Caspar Oesterheld

Oesterheld created “[A Non-Comprehensive List of Human Values](https://casparoesterheld.com/2017/02/10/a-non-comprehensive-list-of-human-values/)” which may be regarded as a direct attempt to formalize human values without the use of AGI.

## Alexandra Surdina

Surdina wrote an article “[Temporal variability in moral value judgement](https://pdfs.semanticscholar.org/dec3/626a7e300c31bd5e9ca6ac13bfb73cfa5870.pdf)” which finds that human moral judgments change even without changes of the outside conditions.

## Steven Umbrello

Imagination Theory toInform Responsible Technology

Design

Imaginative Value Sensitive Design: Using Moral

Imagination Theory toInform Responsible Technology

Design

Umbrello wrote “[Imaginative Value Sensitive Design: Using Moral Imagination Theory to Inform Responsible Technology Design](https://link.springer.com/epdf/10.1007/s11948-019-00104-4?author_access_token=zPkMR8ZbgJLNe1_srhVESPe4RwlQNchNByi7wbcMAY6lkGnRXwv62OWG1OJHM5uqVARC9uKM_p7Efl8m7-VBGNitigfnhj8Lblz9l1NiVDzgbMUJG7ZjAIpdFhE4tMNOXPj9_g9CHix8Yfz22GIXog%3D%3Dfull-text&fbc).” Section 3 of the article reviews opinions of different scientists about what values are, and stated that: “a value is considered that attribute with which an individual or group consigns significance” and “values are instantiated and expressed via their use.” However, it is also stated that these is not enough to fully define human values and a special instrument is needed. This instrument is called “moral imagination” and is “reflexive understanding of the self and the imaginative structure of moral deliberation including its principles and constraints.”

## Vanessa Kosoy

Vanessa. Kosoy in the long post “[The Learning-Theoretic AI Alignment Research Agenda”,](https://www.alignmentforum.org/posts/5bd75cc58225bf0670375575/the-learning-theoretic-ai-alignment-research-agenda) in the section “Value alignment is understandable” suggests that human values are a natural concept and denying them is a form of nihilism. She notes that values are not something human-specific, as we should expect that aliens also should have something like values. All these give reasons to expect, according to Kosoy, that human values could be defined enough for practical tasks.

“The core of AI alignment is reliably transferring human values to a strong AI. However, the problem of defining what we mean by "human values" is a philosophical problem. A common and natural model of "values" is expected utility maximization: this is what we find in game theory and economics, and this is supported by VNM and Savage theorems. However, as often pointed out, humans are not perfectly rational, therefore it's not clear in what sense they can be said to maximize the expectation of a specific utility function.

Nevertheless, I believe that "values" is also a natural concept. Denying the concept of "values" altogether is paramount to nihilism, and in such a belief system there is no reason to do anything at all, including saving yourself and everyone else from a murderous AI. Admitting the general concept of "values" as something complex and human specific (despite the focus on "values" rather than "human values") seems implausible, since intuitively we can easily imagine alien minds facing a similar AI alignment problem. Moreover, the concept of "values" is part and parcel of the concept of "intelligence", so if we believe that "intelligence" (due to its importance in shaping the physical world) is a natural concept, then so are "values".

Therefore, I conjecture that there is a simple mathematical theory of imperfect rationality, within which the concept of "human values" is well-defined modulo the (observable, measurable) concept of "humans". Some speculation on what this theory looks like appears in the following sections.

Now, that doesn't mean that "human values" are perfectly well-defined, anymore than, for example, the center of mass of the sun is perfectly well-defined (which would require deciding exactly which particles are considered part of the sun). However, like the center of mass of the sun is sufficiently well-defined for many practical purposes in astrophysics, the concept of "human values" should be sufficiently well-defined for designing an aligned AGI. To the extent alignment remains ambiguous, the resolution of these ambiguities doesn't have substantial moral significance.”

## Robin Hanson

While Robin Hanson is not AI Safety researcher, he is influential in the field. He wrote in the book “Elephant in the brain” that most of our declarative “values” are rationalisations, which are generated to present us in socially acceptable ways and cover our main motive: increasing our social status. Moreover, we tend to believe in it, as we typically are not capable to recognize and correctly articulate our true values. For example, a person may say that he wants to save the world, but in fact he wants to have higher status role (as saving the world is presumably more higher status activity than being taxi driver).

# 4. Discussion

There are several internally consistent theories of human values suggested by AGI safety researchers in the existing literature. However, despite their internal consistency, these theories are not very compatible with each other and present a wide range of opinions.

There are also other “possible theories,” that is, approaches which have not yet (to the best of our knowledge) been put forth by any researcher, but which could be generated based on the same principles as other theories. One is to assume that all human values are products of evolutionary fitness and can be derived from basic evolutionary considerations in the same way as [Omohundro’s AGI basic drives](https://selfawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf). This could explain most basic human drives like survival, sex, status-seeking, exploration instinct, etc.

Another such theory is that an [AI Oracle](https://nickbostrom.com/papers/oracle.pdf) should first read the extant psychological literature, choose the best theory of mind, and create its structure of human values based on that theory. We are going to explore such a theory in a subsequent work.

From this, we could conclude that internal consistency, as well as experimental support and an extensive literature is not enough to provide us with a “correct” theory of human values, because some alternative theories could also have such level of support. There is a need for some kind of meta-theory of human values, connected also with the methods of their learning.

# 5. Conclusions

In this article, we showed that different AGI safety researchers have suggested different theories about the nature of human values, sometimes contradictory. A theory classification method was suggested, where the theories are judged according to the level of their complexity and behaviorists-internalists scale, as well as the level of their generality-humanity. We suggest that a multiplicity of well-supported theories means that the nature of human values is difficult to define, and some meta-level theory is needed.

# Acknowledgements

The work was started during AI Safety Camp 2 in Prague 2018. We want to thank Linda Linsefors, Jan Kulveit, and Steven Umbrello who provided important feedback for the article. All errors are my own.

1. By AGI safety researchers we mean those who self-identify as such—perhaps using some slightly different words—despite the existence of a few distinct schools of thought in the field. The main criterion here is that they agree that superintelligent AI is possible and dangerous, and that something should be done about it. We do not, however, include narrow AI safety research, e.g. adversarial examples for neural networks. [↑](#footnote-ref-1)