AI Alignment Problem:

“Human Values” Idea is Built Upon Many Assumptions

Alexey Turchin

Digital Immortality Now

Foundation Science for Life Extension

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

Abstract. The main current approach to the AI safety is AI alignment, that is, the creation of AI whose preferences are aligned with “human values.” Many AI safety researchers agree that the idea of “human values” as a constant, ordered sets of preferences is at least incomplete. However, the idea that “humans have values” underlies a lot of thinking in the field; it appears again and again, sometimes popping up as an uncritically accepted truth. Thus, it deserves a thorough deconstruction, which I will do by listing and analyzing comprehensively the hidden assumptions of the idea that “humans have values.” This deconstruction of human values will be centered around the following ideas: “Human values” are useful descriptions, but not real objects; “human values” are bad predictors of behavior; the idea of a “human value *system”* has flaws; “human values” are not good by default; and human values cannot be separated from human minds. The method of analysis is listing hidden assumptions on which the idea of “human values” is built. I recommend that either the idea of “human values” should be replaced with something better for the goal of AI safety, or at least be used very cautiously. The approaches to AI safety which don’t use the idea of human values at all may require more attention, like the use of full brain models, boxing, and capability limiting.

[Introduction 3](#_Toc6741413)

[**1. Ontological status and sources of human values** 3](#_Toc6741414)

[1.1. AI Alignment requires actually existing, stable, finite set of predicting data about person’s motivation, which is called “human values” 3](#_Toc6741415)

[1.2. Human values do not actually exist; they are only useful descriptions of human behavior and rationalization 4](#_Toc6741416)

[1.3. Five sources of information of human values: verbalizations, thoughts, emotions, behavior and neurological scans 6](#_Toc6741417)

[1.4. Where do human values come from: evolution, culture, ideologies, logic and personal events 7](#_Toc6741418)

[**2. Critics of the idea of human values as a constant set of personal preferences: it is based on many assumptions** 8](#_Toc6741419)

[2.1. Human preferences are not constant 8](#_Toc6741420)

[2.2. Human choices are not defined by human values 8](#_Toc6741421)

[1. Situation 8](#_Toc6741422)

[2. Internal conflicts 9](#_Toc6741423)

[3. Emotional affect 9](#_Toc6741424)

[4. Peer pressure 9](#_Toc6741425)

[5. Random processes in the brain 10](#_Toc6741426)

[6. Conditional and unconditional reflexes 10](#_Toc6741427)

[7. Somnambulism, the bisected brain, and actions under hypnosis 10](#_Toc6741428)

[8. Actions under influence of drugs; demented people and children 10](#_Toc6741429)

[2.3. “Human values” can’t be easily separated from biases 11](#_Toc6741430)

[2.4. Human values are subject-centered and can’t be separated from the person 11](#_Toc6741431)

[2.5. Many human values are not actionable + the hidden complexity of values 12](#_Toc6741432)

[2.6. Open question: is there any relation between values, consciousness and qualia? 13](#_Toc6741433)

[2.7 Human values as a transient phenomenon: my values are not mine. 13](#_Toc6741434)

[**3. The idea of “human values” maybe not as useful concept as it looks like for AGI Safety** 13](#_Toc6741435)

[3.1. Human values are not safe if scaled, extracted from a human or separated 13](#_Toc6741436)

[3.2. It is wrong to think of values as of property of a single human: values are social phenomena 14](#_Toc6741437)

[1. Not human have values, but values have humans 14](#_Toc6741438)

[2. Values as memes, used for group building 15](#_Toc6741439)

[3. Ideologies 15](#_Toc6741440)

[4. Religion 15](#_Toc6741441)

[6. Hypnosis and unconscious learning 16](#_Toc6741442)

[3.3. Humans don’t “have” values; they are vessels for values, full of different subpersobalities 16](#_Toc6741443)

[1. Values are changeable but identity is preserved 16](#_Toc6741444)

[2. Subpersonalities 16](#_Toc6741445)

[3. The infinite diversity of human values 17](#_Toc6741446)

[4. Normative plurality of values 17](#_Toc6741447)

[3.4. Values, choices, and commands 18](#_Toc6741448)

[Difference between values and desires 18](#_Toc6741449)

[3.5. Meta-values: preference about values 18](#_Toc6741450)

[1. Meta-values as morals 18](#_Toc6741451)

[2. Suppressed values 19](#_Toc6741452)

[3. Subconscious values and sub-personalities 19](#_Toc6741453)

[4. Zero behavior and contradicting values 20](#_Toc6741454)

[5. Preference about other’s preferences 20](#_Toc6741455)

[3.6. Human values cannot be separated from the human mind 20](#_Toc6741456)

[1. Values are not encoded separately in the brain 20](#_Toc6741457)

[2. Human values are dispersed inside “training data” and “trained neural nets” 21](#_Toc6741458)

[3. Amorphous structure of human internal processes and false positives in finding internal parts 21](#_Toc6741459)

[3.7. Human values presentation is biased towards presenting socially accepted claims 21](#_Toc6741460)

[3.8. Human values could be manipulated by the ways and order they are extracted 22](#_Toc6741461)

[3.9 Human values are, in fact, non-human 22](#_Toc6741462)

[Viruses 22](#_Toc6741463)

[3.10. Any human value model has not only epistemological assumptions, but also has axiological (normative) assumptions 22](#_Toc6741464)

[3.11. Values may be not the best route for simple and effective descriptions of human motivation 23](#_Toc6741465)

[3.12. Who are real humans in “human values”? 23](#_Toc6741466)

[3.13. The human reward function is not “human values” 24](#_Toc6741467)

[3.14. Difficult cases for value learning: enlightenment, art, religion, homosexuality and psi 24](#_Toc6741468)

[3.15. Human values excluding each other and Categorical imperative as a meta-value 25](#_Toc6741469)

[Conclusion 26](#_Toc6741470)

[Acknowledgments 27](#_Toc6741471)

[Appendix. Table of assumptions in the idea of human values 27](#_Toc6741472)

# Introduction

The idea of AI which learns human values is a core of the current approach to artificial general intelligence (AGI) safety. However, it is based on some assumptions about the nature of human values, including the assumption that they completely describe human motivation, are non-contradictory, are normatively good, etc. Actual data from psychology provides a rather different picture.

The literature analysis of what other AGI safety researchers have written about the nature of human values is rather large and is presented in another of my texts: [What AGI Safety Researchers Have Written About the Nature of Human Values](https://www.lesswrong.com/posts/GermiEmcS6xuZ2gBh/what-ai-safety-researchers-have-written-about-the-nature-of). A historical overview of the evolution of the idea of human values can be found in Clawson, Vinson, “[Human values: a historical and interdisciplinary analysis](http://www.acrwebsite.org/volumes/9454/volumes/v05/NA-05)” (1978). A list of ideas for achieving AGI safety without the idea of human values will also be published separately.

In Section 1 the ontological status of human values is explored. In section 2 the idea of human values as an ordered set of preferences is criticized. Section 3 explores whether the idea of human values is useful to AGI safety.

# **1. Ontological status and sources of human values**

## 1.1. AI alignment requires an actually existing, stable, finite set of predicting data about peoples’ motivation, which is called “human values”

In an AI alignment framework, future advanced AI will learn human values. So, we don’t need to directly specify human preferences, we just need to create AI capable of learning human values. (From a safety point of view, there is a circularity problem here, as such AI needs to be safe before it starts to learn human values, or it could do it in unsafe and unethical ways, as I describe in detail in [Possible Dangers of the Unrestricted Value Learners](https://www.lesswrong.com/posts/hzEaasJyQsutYDNfN/possible-dangers-of-the-unrestricted-value-learners), but let’s assume for now as it is somehow bypassed—perhaps via a set of preliminary safety measures.)

The idea of AI alignment is based on the idea that there is a finite, stable set of data about a person, which could be used to predict one’s choices, and which is actually morally good. The reasoning behind this basis is because if it is not true, then learning is impossible, useless, or will not converge.

The idea of value learning assumes that while human values are complex, they are much simpler than the information needed for whole brain emulation. Otherwise, full brain emulation will be the best predictive method.

Moreover, the idea of AI alignment suggests that this information could be learned if correct learning procedures are found. (No procedures = no alignment.)

This *actually existing and axiological good, stable, finite set of predicting data about peoples’ motivation is often called “human values,*” and it is assumed that any AI alignment procedure will be able to learn this data. This view on the nature of human values from an AI alignment point of view is rather vague, and it doesn’t say *what* human values are, neither does it show how they are implemented in the human mind. This view doesn’t depend on any psychological theory of the human mind. As a pure abstraction, it could be applied to any agent whose motivational structure we want to learn.

Before the values of a person can be learned, they have to become “human.” That is, they need to be combined with some theory about how values are encoded in the human brain. AI safety researchers have suggested many such theories, and exiting psychological literature suggests even more theories about the nature of human motivation.

In psychology, there is also a “theory of human values,” a set of general motivational preferences which influence choices. This theory should be distinguished from “human values” as an expected output of an AI alignment procedure. For example, some psychological tests may say that Mary’s values are freedom, kindness, and art. However, the output of an AI alignment procedure could be completely different and not even presented in words, but in some set of equations about her reward function. To distinguish human values as they are expected in AI alignment from human values as a part of psychology, we will call the first “human-values-for-AI-alignment.”

The main intuition behind the idea of human values is that, in many cases, we can predict another person’s behavior if we know what he wants. For example: “I want to drink,” or “John wants to earn more money” often clearly translates into agent-like behavior.

As a result, the idea behind AI alignment could be reconstructed as the following: if we have a correct theory of human motivation a priori, and knowledge about the human claims and choices as posteriori data, we could use something like Bayesian logic to reconstruct his actual preferences.

To have this a priori knowledge, we need to know the internal structure of human values and how they are encoded in the human brain. Several over simplified theories about the structure of the human values have been suggested: as a human reward function, as a reward-concept association, as a combination of liking-approving-wanting etc.

However, all these are based on the assumption that human values *exist* at all: That the human motivation could be compressed unequivocally into one—and only one—simple stable model. This, and other assumptions, appear even before the “correct psychological theory” of human values is chosen. In this article, these assumptions will be analyzed.

## 1.2. Human values do not actually exist; they are only useful descriptions of human behavior and rationalization

The idea that human behavior is determined by human values is now deeply incorporated into people’s understanding of the world and is rarely subject to reservations. However, the ontological status of human values is uncertain: do they actually exist, or they are just a useful way of describing human behavior? The idea of human-values-for-AI-alignment requires that some predicting set of data about motivation does actually exist. If it is only a description, in which there could be multiple descriptions in various situations, extrapolating such a description will create problems.

In other words, descriptions are observer-dependent, while actually existing things are observer independent. If we program in an agent with utility function U, this utility function exists independently of any observations and could be unequivocally learned by some procedures. If we have a process with agent-like features, it could be described differently, depending on how complex a model of the process we want to create.

For example, if we have a mountain with several summits, we could describe it as one mountain, two mountains or three mountains depending on the resolution ability of our model:



In the discussion of the ontological status of human values we encounter a very long-standing philosophical problem of the reality of [*universals*](https://en.wikipedia.org/wiki/Problem_of_universals), that is, high-level abstract ideas. The Middle Ages dispute between realists, who thought that universals are real, and nominalists, who thought that only singular things are real, was won by nominalists. From history, we know that “human values” is a relatively new construction which appears only in some psychological theories of motivation.

However, we can’t say that “human values” are just random descriptions, which can be chosen completely arbitrarily, because there are some natural levels where the description matches reality. In case of values, it is a level of one’s claims about his-her preferences. While these claims may not perfectly much any deeper reality of one’s values, they exist unequivocally, at least in the given moment. The main uncertainty in the human values is the unequivocal existence of some deeper level, which creates such claims, which is the level of “true values”.

Human preferences are relatively easy to measure (while measuring libido is not easy). One could ask a person about his/her preferences, and s/he will write that s/he prefers apples over oranges, Democrats over Republicans, etc. Such answers could be statistically consistent in some groups, which could allow the prediction of future answers. But it is often assumed that real human values are something different than explicit preferences. Human values are assumed to be able to generate preferential statements but not to be equal to them.

One could also measure human behavior in these choice experiments, and check if the person actually prefers oranges over apples, and also get consistent results. The obvious problem of such preferential stability is that it is typically measured for psychologically stable people in a stable society, and in a stable situation (a controlled experiment). The resulting stability is still statistical: That is, one who like apples, may sometimes choose an orange, but this atypical choice may be disregarded as noise in the data.

Experiments which deliberately disrupt situational stability consistently show that human preferences play a small role in actual human behavior. For example, changes in social pressure result in consistent changes in behavior, thus contradicting declared and observed values. The most famous example is the [Stanford Prison Experiment](https://en.wikipedia.org/wiki/Stanford_prison_experiment), where students quickly took on abusive rolls.

The only way for human values to actually exist, would be if we could pinpoint some region of the human brain where they are explicitly presented as rules. However, only very simple behavioral patterns, like the swimming reflex, may actually be genetically hardcoded in the brain, and all others are socially defined.

So, there are two main interpretations of the idea of “human values”:

1. *Values actually exist, and each human makes choices based on their own values.* There is one stable source of human claims, actions, emotions and measurable preferences, which completely defines them, is located somewhere in the brain, and could be unequivocally measured.

*2) Values are useful descriptions.* Humans make choices under the influence of many inputs, including situation, learned behavior, mood, unconscious desires, and randomness, and to simplify the description of the situation we use the designation “human values.”More detail on this topic can be found in the book by Lee Ross et al: “[The person and the situation: Perspectives of social psychology](https://www.amazon.com/Person-Situation-Perspectives-Social-Psychology/dp/1905177445).

Humans have a surprisingly big problem when they are asked about their ultimate goals: They just don’t know them! They may create *ad hoc* some socially acceptable list of preferences, like family, friendship, etc., but this will be a poor predictor of their actual behavior.

It is surprising that most humans can live successful lives without explicitly knowing and using their list of goals and preferences. In contrast, a person can generally identify his/her current wishes, a skill obviously necessary for survival, for example s/he can consider thirst and the desire for water.

## 1.3. Five sources of information of human values: verbalizations, thoughts, emotions, behavior and neurological scans

There are several different ways one could learn about someone’s values. There is an underlying assumption that all of these ways converge to the same values, which appears to be false upon closer examination. The main information channels to learn someone’s preferences are:

1. *Verbal claims*. This what a person says about his/her preferences. Such claims try to present a person as better according to expected social norms. Armstrong [suggested](https://www.lesswrong.com/posts/pQz97SLCRMwHs6BzF/using-lying-to-detect-human-values) the examination of facial expressions when a person lies about his/her true values to deduce his/her real values, perhaps by training some AGI to do it. He based this suggestion on the interesting idea that “Humans have a self-model of their own values.” However, it appears that most humans either could live without such a model, or that their model is rationalization made to look good. Such claims could have different subtypes: what a person says to friends, writes in books, etc. Written claims could be more consistent and socially appropriate, as they are better thought out. Claims to close friends could be more orientated toward short-term effect, manipulation, and social situation-dependence. At the same time, claims to friends could also be more sincere, as they have been subjected to less internal censorship. Similarly, claims made under drugs, especially alcohol, could be even less censored, but might not present “true values”. They could present some suppressed “counter-values”, such as the use of an obscene lexicon or some mimetically replicated social cliché, like “I hate all members of social group X.”
2. *Internal thought claims:* the private thoughts which appear in the internal dialog or planning. People may be more honest in their thoughts. However, many people lie to themselves about their own values, or are just unable to fully articulate the complexity of their values.
3. *Behavior.* What a person actually does could represent the sum of all his/her desires, trained models of behavior, random actions, etc. Contradicting values could result in zero behavior, such as a case when one wants to buy a dress, but is afraid to spend too much money on it. Behavior can also take different forms: *choices* between two alternatives, which one might signal in many ways; *verbal behavior*, other than statements of one’s own values; and chains of *physical actions* (e.g. dancing).
4. *Expression of emotions*. Human values could be reconstructed based on emotional reactions to stimuli. A person could prefer to look at some images longer, feel arousal, smile, etc. However, this way of learning values would overestimate suppressed emotions and underestimate rational preferences. For example, a pedophile may be become aroused by some type of images, but on the rational level s/he may fight this type of emotion. Emotions could be presented to the outside in many ways, by facial expressions, tone of voice and content of speech, pose, and even body odor. A person could also suppress expression of emotions or fake them.
5. Non-behavioral, *neurophysiological representations of values.* Most of these are currently unavailable to outside observers, but brain waves, neurotransmitter concentrations or single-neuron activations, as well as some connectome connections, could be directly or indirectly used to gather information about one’s values. AGI with advanced nanotechnology may have full access to the internal states of one’s brain.

## 1.4. Where do human values come from: evolution, culture, ideologies, logic and personal events

If one has something (e.g. a car), it is assumed that one made an act of choice by either buying it or at least keeping it in one’s possession. However, this description is not applicable to values, as one is not making a choice to have a value, but instead makes choices based on the values. Or, if one says that a person makes a choice to hold some value (and this choice was not based on any other values), one assumes the existence of something like “free will”, which is an even more speculative and problematic concept than values [ref]. Obviously, some instrumental values could be derived from terminal values, but that is more like planning, not generation of values.

If one were to define the “source” of human values, it would simplify value learning, as one could derive values directly from the source. There are several views about the genesis of human values:

1. God gives values as rules.
2. “Free will”: some enigmatic ability to create values and choices out of nothing.
3. Genes and evolution: Values are encoded in human genes in the form of some basic drives, and these drives appeared as a result of an evolutionary process.
4. Culture and education: Values are embedded in social structure and learned. There are several subvariants regarding source, e.g. language, religion, parents, social web, social class (Marx), books one read or memes which are currently affecting the person.
5. Significant personal events: These could be trauma or intense pleasure in childhood, e.g. “birth trauma,” or first love in school.
6. Logical values: A set of ideas that a rational mind could use to define values based on some first principles, e.g. Kant’s imperative [ref].
7. Random process: Some internal random process results in choosing the main priorities, probably in childhood [ref].

God and free will are outside of rational discussion. However, all the other ideas have some merit as these six factors could affect the genesis of human values and it is not easy to choose one which is dominating.

# **2. Critics of the idea of human values as a constant set of personal preferences: it is based on many assumptions**

## 2.1. Human preferences are not constant

Personal values evolve from childhood to adulthood. They also change when a person becomes a member of another social group, because of the new and different role, exposure to different peer pressure and different ideology.

Moreover, it is likely that we have *a meta-value about evolving values*: that it is good that someone’s values are changing with age. If a person continues to play with the same toys at 30 he played with at 3 years old, it may be a signal of developmental abnormalities.

Another way to describe human preferences is not as “values”, but as “wishes”. The main difference is that “values” are assumed to be constant, but wishes are assumed to be constantly changing and even chaotic in nature. Also, a wish typically disappears when granted. If I wish for some water and then get some, I will not want any more water for the next few hours. Wishes are also more instrumental and often represent physiological needs or basic drives.

## 2.2. Human choices are not defined by human values

The statement that “humans have values” assumes that these values are most important factor in predicting human behavior. For example, if we know that a chess AI’s terminal goal is to win in chess, we could assume that it *will* try to win in chess. But in the human case, knowing someone values may have surprisingly little predictive power about this person’s actions.

In this subsection, we will look at different situations in which human choices are not defined by (declared) human values but are affected by some other factors.

### 1. Situation

The idea of “human values” implies that a person acts according to his/her values. This is the central idea of all value theory, because it assumes that if we know choices, we can reconstruct values, and if we know values, we can presumably reconstruct the behavior of the person.

There is also another underlying assumption that the relation between behavior and values is unequivocal, that is, given a set of behavior *B* we could reconstruct one and only one set of values *V* which defines it. But this doesn’t work even from a mathematical point of view, as for any finite *B* there exist infinitely many programs which could create it. Thus, for a universal agent, similar behavior could be created by very different values. Armstrong wrote about this, stating that the behavior of an agent depends not only on values, but on policy, which, in turn, depends on one’s biases, limits of intelligence, and available knowledge.

However, in the human case, the main problem is not that human beings are able to pretend that they have one set of values, but that they actually have different values. Typically, only con artists and psychopaths are lying about their actual intentions. The problem is that human behavior is not defined by human values at all, as demonstrated in numerous psychological experiments. A great description of these results can be found in Ross and Nisbett’s book “The person and the situation: Perspectives of social psychology”.

In a 1973 experiment, Darley and Batson checked if a person would help a man who was lying in their path. “They examined a group of students of the theological seminary who were preparing to utter his first sermon. If the subjects, being afraid of being late for the sermon, hurried, then about 10% of them provided assistance. On the contrary, if they did not hurry, having enough time before it began, the number of students who came to the aid increased to 63%”.

Ross et al wrote that maximum attainable level of prediction of the behavior of a person in a new situation, based either on their personal traits or statistics regarding their previous behavior, has a correlation coefficient of 0.3.

### 2. Internal conflicts

Another important conception described Ross and Nisbett’s “The person and the situation” is that stable behavior can be underpinned by conflicting attitudes, where different forces balance each other. For example, a person wants to have unlimited access to sex, but is also afraid of social repercussions and costs of such desire, and thus uses porn. This may be interpreted as if he has a wish to or values using porn, but that is not so: porn is only a compromise between two forces, and such a balance is rather fragile and could have unpredictable consequences if a person is placed in a different situation. These ideas were explored by Festinger (Ross, p29).

### 3. Emotional affect

It is known that many crimes occur under intense and unexpected emotional affect, for example “road rage,” or murders committed out of jealousy. These emotions are intense reaction of our “atavistic animal brain” to the situation. Such situation may be insignificant in broader context of contemporary civilization, but intense emotions can override our rational judgements and almost take control over our actions.

Note, there is still no conclusion about the nature of emotion in psychological literature, though there is a rational model of emotions as accelerators of learning by increasing appreciation of the situation (as mentioned in Sotala’s [article](http://intelligence.org/files/DefiningValuesForValueLearners.pdf) about values).

*[Umbrello comment: “This is the exact point that Johnson (in the above comment) argues against, the enlightenment era idea of the separation of psychological faculties (i.e., reason vs. imagination). We have to be careful to not fall within this dichotomy since it is not clear what the boundaries of these different states of mind are.”]*

### 4. Peer pressure

Experiments conducted by Ash and Milgram demonstrated that peer pressure can cause people to act against what they perceive or value. Zimbardo’s Stanford Prison Experiment also demonstrated how peer pressure affect people behavior and even believes.

### 5. Random processes in the brain

Some human actions are just random. Neurons can fire randomly, and many random factors affects mood. This randomness creates noise in experiments, but basically, we try to clean the data of noise. We could hide randomness of behavior in some probabilistic predictions about behavior.

Humans can randomly forget or remember something, and this includes their wishes. In other words, declared values could randomly drift.

### 6. Conditional and unconditional reflexes

Some forms of behavior are hardwired in the human brain and even in the primitive hindbrain, and thus are independent of any human values, that is, unconditional reflexes, e.g. the swimming reflex, fight or flight response, etc.

There are also conditional reflexes, i.e. it is possible to train a person to present reaction B if stimuli A is given. If such reflex is trained, such reflex does not present any information about the person’ values. But some desires can be triggered intentionally, an approach which is intensively used in advertising: a person seeing Coca-Cola may start to feel a desire to drink soda. Similarly, a person hearing a loud bang may have a panic attack if he has PTSD.

### 7. Somnambulism, the bisected brain, and actions under hypnosis

It is well-known that humans are capable of performing complex behaviors while completely unconscious. The best example is somnambulism, or sleep walking. Some people are able to perform complex behavior in that state, even drive a car and commit a murder, without any memories of the event (in this way, it differs from actions in dreams, where at least some form of control exists). Surely, a person’s actions in that situation could not be used to extrapolate the person’s preferences.

While somnambulism is an extreme case, many human actions occur mechanically, that is, out of any conscious control, including driving a car and the compulsive behavior of addicts.

Experiments (as often in psychology, questionable) have also demonstrated that humans whose brain hemispheres were separated have two different “consciousnesses” with different preferences (though these results have recently been challenged) [ref].

Another extreme case is hypnosis, where a human is conditioned to act according to another person’s will, sometimes even without knowing it. While extreme cases of hypnosis are rare and speculative, the effectiveness of TV propaganda in “brain washing” demonstrates that some form of suggestion is real and plays an important role in mass behavior. For example, Putin’s autocracy invested a lot to gain control over TV and most of TV-viewers in Russia support Putin politics.

### 8. Actions under influence of drugs; demented people and children

Some drugs, which are part of human culture and value systems, notably alcohol, are known to change behavior and presented values, mostly because self-control is lowered and suppressed instinctive drives become active. Also, the policies to achieve goals become less rational under drugs. It also seems that alcohol and other drugs increase internal mis-alignment between different subpersonalities.

While a person is legally responsible for what he does under influence of drugs, his presenting of his values changes: some hidden or suppressed values may become openly expressed (in vina veritas). Even some cars’ AI can recognize that a person is drunk and prevent him from driving.

For pure theoretical AGI this may be a difficulty, as it is not obvious why sober people are somehow more “value privileged” than drunk people. Why, then, should the AGI ignore this large class of people and their values?

Obviously, “drunk people” is not the only class which should be ignored. Small children, patients in mental hospitals, people with dementia, dream characters, victims of totalitarian brainwashing, etc. – all of these and many more can be regarded as classes of people whose values should be ignored, which could become a basis for some form of discrimination at the end.

Also, presented values depends on the time of the day and physiological conditions. If a person is tired, ill or sleepy, this could affect his-her values centered behavior.

An extreme case of “brainwashing” is feral children risen by animals: and most of their values also should not be regarded as “human values”.

## 2.3. “Human values” can’t be easily separated from biases

The problem of the inconsistency of human behavior was well known to the founders of the rationalists and the AGI safety movement, who described it via the idea of biases. It seems that humans, according to rationalists understanding, have a constant set of values. However, humans act irrationally based on this set of values because they are affected by numerous cognitive biases. By applying different rationalist training and debiasing to a person, we could presumably create a “rational person” who will act consistently and rationally and will effectively reach his-her own positive values. The problem is that such model of purely rational person acting on the set of coherent altruistic values is completely non-human.

*[Umbrello comment: Heuristic tools can be used to de-bias AGI design. I argued this in a paper, and showed a way in which it can be done. See Umbrello, S. (2018) ‘The moral psychology of value sensitive design: the methodological issues of moral intuitions for responsible innovation’, Journal of Responsible Innovation. Taylor & Francis, 5(2), pp. 186–200. doi: 10.1080/23299460.2018.1457401.]*

Another problem is that a lot of humans have different serious psychiatric diseases, including schizophrenia, obsessive-compulsive disorder, mania, and others, which significantly affect their value structure. While extreme cases can be easily recognized, weaker forms may be part of the “psychopathology of ordinary life”, and thus part of “human nature”. We don’t know if a truly healthy human mind exists at all.

Armstrong [suggested](https://www.lesswrong.com/posts/FgRRY3vSwDLx2Hk46/the-low-cost-of-human-preference-incoherence) not to separate biases from preferences, as AGI will find easy ways to overcome biases. But the AGI could, in the same way, find the paths to overcome the preferences.

## 2.4. Human values are subject-centered and can’t be separated from the person

In the idea “humans have values,” the verb “have” assumes the type of relation that could be separated. This implies some form of orthogonality between human mind and human values, as well as a strict border between the mind and values. For example, if I have an mp3 file, I can delete the file. In that case, the statement “I don’t have the file” will be factual. I can give this file to another person and in that case, I can say: “That person now has the file”. But human values can’t be transferred in the same way as a file for two reasons: they are subject-centered, and there is no strict border between values and the other parts of the mind.

Most human values are centered around a particular person (with the exception of some artificially constructed purely altruistic values, like someone who wants to reduce the amount of suffering in the world, but completely ignoring who is sufferings: humans or animals, etc.) One may argue that non-subject values are better, but this is not how human values works. For example, a person attaches a value not to a tasty food, but to the fact that he will consume such food in the future. If healthy food exists without the potential one could consume it, we can’t say that it has value.

From this, it follows that if we copy human values in an AGI, that AGI should describe the same state of the world, but not the same preferences. For example, we don’t want to copy into AGI a desire to have sex with humans, but we want that AGI will help its owner in his/her reproductive success. However, instrumental goals like self-preservation will be still AGI-centered.

The subject of value is more important than a value itself, because if a typical human A has some value X, there is surely someone else on Earth who is getting X, but X doesn’t matter to that person. However, if the same person A gets another valuable thing Y, it is still good for him. Attempts to properly define the subject quickly evolves into the problem of personal identity, which is notoriously difficult and known to be paradoxical. That problem is much more difficult to verbalize, that is, a person may correctly say what he wants, but fails to provide a definition of who he is.

Obviously, there is no easy way to separate values from all underlying facts, neuronal mechanisms, biases and policies – more on that in the next subsection. More about similar problems was said in the post of Joar Skalse “[Two agents can have the same source code and optimise different utility functions](https://www.lesswrong.com/posts/zMHK9gFY6t48Exqup/two-agents-can-have-the-same-source-code-and-optimise).”

Human preferences are self-centered, but if AGI takes human preferences as its own, they will not be AGI-centered, but will be preferences about the state of the world, and this will make them closer to the external rules. In other words, preferences about the well-being of something outside you is an obligation and burden, and AGI will search the ways to overcome such preferences.

## 2.5. Many human values are not actionable + the hidden complexity of values

If someone says that “I like poetry”, it is a clear representation of his/her declarative values, but it is unlikely to predict what he actually does. Is he writing poems every day for an hour, and if so, which type? Or he is reading every week for two hours – and what does he read: Homer, Byron or his girlfriend’s poems? Will he attend a poetry slam?

This could be called the “hidden complexity of values,” but if we start to unknot that complexity, there will be no definite border between values and everything else in the person’s mind. In other words, short textual representations of values are not actionable, and if we try to make a full representation, we will end up reproducing the entire brain.

In Yudkowsky’s example of the complexity of values, about removing one’s aged mother from a burning house, the complexity of values comes from many common-sense details, which are not included into the word “removing”.

## 2.6. Open question: is there any relation between values, consciousness and qualia?

In some models, where preferences dictate choices, where is no need for consciousness. However, many preferences are framed as preferences about future subjective experiences, like pain or pleasure.

There are at least 3 meanings of the idea of “consciousness” and 3 corresponding questions:

1. Consciousness is what I know and can said about it – Should we care about unconscious values?
2. Consciousness is what I feel as pure subjective experience, qualia – Should we solve the problem of qualia to in order to correctly present human preferences about subjective experiences?
3. Consciousness is my reflection about myself and only values which I declare *are* my values should be counted – True or not?

Related: G. Worley on philosophical conservatism: “[Philosophical Conservatism in AI Alignment Research](https://www.lesswrong.com/posts/3r44dhh3uK7s9Pveq/rfc-philosophical-conservatism-in-ai-alignment-research)” and “[Meta-ethical uncertainty in AGI alignment](https://www.lesswrong.com/posts/uR3znuBnaevssYDZY/rfc-meta-ethical-uncertainty-in-agi-alignment),” where he discusses the problems with meta-ethics and the non-existence of moral facts. See also the [post](https://www.lesswrong.com/posts/x4n4jcoDP7xh5LWLq/book-summary-consciousness-and-the-brain) of Sotala about consciousness and the brain.

## 2.7 Human values as a transient phenomenon: my values are not mine

Human values are assumed to be a stable but hidden source of human choices, preferences, emotions and claims about values. However, human values—even assuming that such source of all motivation really exists—are constantly changing on day-to-day basis, as a person is affected by advertising, new books, new friends, changes in hormone levels, and mood.

Interestingly, personal identity is more stable than human values. A person remains the same in his/her own eyes as well as the eyes of other people, despite significant changes of values and preferences

# **3. The idea of “human values” maybe not as useful concept as it looks like for AGI Safety**

## 3.1. Human values are not safe if scaled, extracted from a human or separated

Many human values evolved in the milieu of strong suppression from society, limited availability of needed resources, limits on the ability to consume resources, and pressure from other values, and thus don’t scale safely if they are taken alone, without their external constraints.

A possible example of the problem from animal kingdom: if a fox gets into a henhouse, it will kill all the chickens, because it hasn’t evolved a “stop mechanism”. In the same way, a human could like tasty food, but relies on internal body regulation to decide when to stop, which does not always work.

If one goal or value dominates over all other values in one’s mind, it becomes “[paperclippy](https://wiki.lesswrong.com/wiki/Paperclip_maximizer)”, and turn a person into a dangerous manic. Examples include sexual deviants, hoarders, and money-obsessed corporate managers. In contrast, some values balance one another, like the desire for consumption and the desire to maintain a small ecological footprint. If they are separated, the consumption desire will tile the universe with “[orgasmium](https://wiki.lesswrong.com/wiki/Orgasmium),” and “ecological desire” will end in an attempt to stop existing.

The point here is that values without humans are dangerous. In other words, if I want to get as much X as possible, getting 1000X is maybe not what I want – though my expressed desire can convert my AGI into a paperclipper.

In the idea of “human have values” it is intrinsically assumed that these values are a) good and b) safe. A similar idea has been explored in a post by Wei Dai, “[Three AGI safety related ideas](https://www.lesswrong.com/posts/vbtvgNXkufFRSrx4j/three-ai-safety-related-ideas).”

Historically “human values” were not regarded as something good. Humanity regarded suffering as arising from “original sin” and affecting by all possible dangerous effects: lust, greed, etc. There was no worth in human values for the philosophers of the past, and that is why they tried to create morals or a set of laws, which would be much better than inborn human values. In that case, the state or religion provided the correct set of norms, and human nature was merely a source of sin.

If we take rich, young people at the beginning of 21st century, we may see that they are in general not so “sinister” as humans in the past and that they sincerely support all kinds of nice things. However, humanity’s sadistic nature is still here, we just use socially accepted ways to realize our “desire to kill” by watching “Game of Thrones” or playing “World of Tanks”. If AGI extrapolated our values based on our preferences in games, we could find ourselves in a nightmarish world.

There is also completely “unhuman” ideologies and cultural traditions. First is obviously German national-socialism, and also ancient Maya culture, where upper classes constantly ate human meat. Another example is religion groups practicing collective suicide, ISIS and terrorists. Notably, transhumanist thought states that to be a human means to want to overcome human limitations, including innate values.

**AGI which is learning human values will be not intrinsically safer than AGI with hard coded rules.** We may want to simplify AGI alignment by escaping hand-coded rules and by giving AGI authority to extract our goals and to extrapolate them. But there is no actual simplification: we still have *to* *hand-code a theory of human values* and the ways how to extract them and to extrapolate them. This creates large uncertainty, which is not better than rule coding. Naturally, problems arise regarding the interaction of AGI with “human values”: for example, if a person wants to commit suicide, should AGI help him?

**We don’t need AGI alignment for all possible human tasks**: Most of these tasks can be solved without AGI (by [Drexler’s CAIS](https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf), for example). The only task for which alignment is really needed is “*preventing the creation of other unsafe AGI,*” that is, using AGI as a weapon to stop other AGI projects. Another important and super-complex task which requires superintelligent AGI is reaching human immortality.

## 3.2. It is wrong to think of values as of property of a single human: values are social phenomena

### 1. Not human have values, but values have humans

In the statement “Human have values,” separate human beings are presented as the main subjects of values, i.e. those who have values. But most values are defined by society and describe social behavior. In other words, as recognized by Marx, many values are not personal but social, and help to keep society working according the current economic situation.

Society expends enormous effort to control people’s values via education, advertising, celebrities-as-role-models, books, churches, ideologies, group membership identity, shaming, status signaling and punishment. Social values consist of unconscious repeating of the group behavior + conscious repeating of norms to maintain membership in the group. Much behavior is directed via the unconscious definition of one’s social role, as described by Valentino in the post “[The intelligent social web](https://www.lesswrong.com/posts/AqbWna2S85pFTsHH4/the-intelligent-social-web).”

### 2. Values as memes, used for group building

Very rarely a person could evolve his/her own values without being influenced by anyone else; this often happens against his/her own will. In other words, it is not the case that “humans have values” – a more correct wording would be “values have humans.” This is especially true in the case of ideologies, which could be seen as especially effective combinations of memes, something like a memetic virus consisting of several proteins-memes, often supported by a large “training dataset” of schooling in a culture where this type of behavior seems to be a norm.

### 3. Ideologies

In the case of ideologies, values are not human preferences, but instruments to manipulate and bind people. In ideologies (and religions) values are most articulated, but they play the role of group membership tokens, not actual rules dictating actions. However, most people are unable to follow such sets of rules.

Hanson wrote “X is not about X,” and this is an example. To be a member of the group, a person must vocally agree that his/her main goal is X (e.g. “Love god X”), which is easy verifiable. But if he is actually doing enough for X is much less measurable, and sometimes even unimportant.

For example, Jesus promoted values of “living as a bird,” “poverty” or “turn the other cheek” (“But I say unto you, That ye resist not evil: but whosoever shall smite thee on thy right cheek, turn to him the other also.” Mat 5:39, KJV), but churches are rich organizations and humanity constantly engages in religious wars.

A person could be trained to have any value either by brainwashing or by intensive reward, but still preserve his/her identity.

### 4. Religion

### 4. Religion

In religions, values and ideologies are embedded in more complex mythological context. Most people who ever lived were religious. Religion as a successful combination of memes is something like a genetic code of culture. Religion also requires a whole adherence to all—even smallest rituals, like eating certain types of food and wearing exact forms of hats—not only to a few ideological rules.

There is a theory that religion was needed to compensate the fear of death in early humans, and thus humans are genetically selected to be religious. The idea of God is not a necessary part of religion, as there are religion-like belief systems without a god which, however, have all the structural elements of religion (Buddhism, communism, UFO cults like Raëlian movement).

Moreover, even completely anti-religious and declaratively rational ideologies may still have structural similarities to religion, as was mentioned by Cory Doctorow in “Rapture of the Nerds.” Even the whole idea of a future superintelligent AI cold be seen as a religious view mirrored into the future, in which sins are replaced with “cognitive biases,”, churches with “rationality houses” etc.

In the case of religion, a significant part of “personal values” are not personal, but are defined by religious membership, especially the declarative values. Actual human behavior could significantly deviate from the religious norms because of the combination of affect, situation, and personal traits.

### 6. Hypnosis and unconscious learning

At least some humans are susceptible to influence by the beliefs of others, and charismatic people use this ability. For example, I knew about cryonics for a long time, but only started to believe in it after Mike Darwin told me his personal view about it.

The highest (but also the most controversial) from of suggestibility is hypnosis, which has two not-necessarily-simultaneous manifestations: trans induction and suggestion. The second doesn’t necessary require the former.

People also could learn via observation of actions of other people, which is unconscious learning.

## 3.3. Humans don’t “have” values; they are vessels for values, full of different subpersobalities

### 1. Values are changeable but identity is preserved

In this section, we will look more closely at the connection between a person and his/her values. When we say that “person X has value A,” some form of strong connection is implied. But human personal identity is stronger than most of the values the person will have during his/her lifetime. It is assumed that identity is preserved from early childhood; for example, Leo Tolstoy wrote that he felt himself to be the same person from 5 years old until his death. But most human values change during that time. Surely there can be some persistent interests which appear in human childhood, but they will not dominate 100 percent of the time.

Thus, human personal identity is not based around values, and the connection between identity and values is weak. Values can appear and disappear during a lifetime. Moreover, a human can have contradicting values in the same moment.

We could see a person as some vessel where desires appear and disappear. In a normal person, some form of “democracy of values” is happening: he makes choices by comparing the relative power of different values and desires in a given moment, and the act of choice and its practical consequences updates the balance of power of different values. In other words, while values remain the same, the preferential relation between them is changing.

From the idea of the personality as a vessel for values follows two things:

1. Human values could be presented as subagents which “live” in the vessel
2. There are meta-values which preserve the existence of the vessel and regulate the interaction between the values.

### 2. Subpersonalities

Many different psychological theories describe the mind as consisting of two, three, or many parts that can be called subpersonalities. The obvious difficulty of such division is that subpersonalities do not “actually” exist, but instead are useful description. But as descriptions they are not passive; they can actively support any theory and play along in the roles which are expected.

Another difficulty is that different people have different level of schizotypy, or different “decoherence” between subpersonalities: hyper-rational minds can look completely monolithic, fluid minds can create subpersonalities *ad hoc*, and some people can suffer from a strong dissociative disorder and actually possess subpersonalities.

Some interesting literature on subpersonalities (beyond Kulveit’s AGI Safety theory) includes:

[Victor Bogart](http://journals.sagepub.com/doi/abs/10.1177/00221678940342006?journalCode=jhpa), “[Transcending the Dichotomy of Either "Subpersonalities" or "An Integrated Unitary Self](http://journals.sagepub.com/doi/abs/10.1177/00221678940342006?journalCode=jhpa)"

Lester wrote a lot about theory of subpersonalities “[A Subself Theory of Personality](https://link.springer.com/article/10.1007/BF02686812).”

[Encyclopedia of Personality and Individual Differences](https://www.springer.com/gp/book/9783319246109) includes a section by Lester with finding about subpersonalities (p 3691).

Sotala started a new sequence “[Sequence introduction: non-agent and multiagent models of mind](https://www.lesswrong.com/posts/M4w2rdYgCKctbADMn/sequence-introduction-non-agent-and-multiagent-models-of)”

Mihnea Moldoveanu “[The self as a problem: The intra-personal coordination of conflicting desires](https://www.researchgate.net/publication/222697072_The_self_as_a_problem_The_intra-personal_coordination_of_conflicting_desires)”

Minsky in the “[Society of mind](https://www.emcp.com/intro_pc/reading12.htm)” wrote about many too-small agents in the human mind – K-lines, which are much simpler than “personalities.” But current artificial neural nets don’t need them.

### 3. The infinite diversity of human values

The idea that “human have values” assumes that there is a special human subset of all possible values. However, human preferences are very diverse. For any type of objects, there is a person who collects them or likes the YouTube videos about them. Humans can have any possible values, limited only by values’ complexity.

### 4. Normative plurality of values

Most moral theories like utilitarianism tries to search for just one correct overarching value. However, there appear problems like *repugnant conclusion*. Such problems appear if we take the value literary or try to maximize it to extreme levels. The same problems will affect a possible future AGI if it tries to over-maximize its utility function. Even *paperclip maximizer* just wants to be sure that it will create enough paperclips. Because of this, some writers on AGI safety started to suggest that we should escape utility functions in AGI, as they are inherently dangerous. (For example, in the post of Shah “[AGI safety without goal-directed behavior](https://www.alignmentforum.org/posts/tHxXdAn8Yuiy9y2pZ/ai-safety-without-goal-directed-behavior)”.)

The idea that good moral model should be based on existence of many different values – without any overarching value – is presented in the article by Carter “[A plurality of values](https://www.academia.edu/173502/A_plurality_of_values).” However, this claim is self-contradictory, because the norm “there should not be overarching value” is itself overarching value. Carter escape it by suggesting to use “indifference-curves” from microeconomics: a type of utility function which combines two variables.

However, in that case overarching values maybe “content free”. For example, functional democracy provides everybody the right to free speech, but don’t prescribe the content of speech besides a few highly debated topics like hate speech or speech which affected other ability to speak. But exactly this “forbidden” topics as well as level of their restriction becomes the most attractive to discussions very soon.

Bostrom wrote about [Parliamentary Model](http://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html), where different values are presented. But any parliament need speaker and rules.

## 3.4. Values, choices, and commands

A person could have many contradictory values, but an act of choice is the irreversible decision of taking one of several options – and such choice may take the form of a command for a robot or an AGI. The act of making a choice is something like an irreversible collapse (similar to the collapse of a quantum wave function on some basis), and making a choice requires significant psychological energy, as it often means denying the realization of other values, and consequently, feeling frustration and other negative emotions. In other words, making a choice is a complex moral work, not just a simple process of inference from an existing set of values. Many people suffer from an inability to make choices, or an inability to stick with choices they made previously.

A choice is typically not finalized until we take some irreversible action in the chosen direction, like buying a ticket to the country.

In the case of Task AGI (AGI designed to make a task perfectly and then stop), the choice is moment when we give the AGI a command.

In some sense, making choices is moral work of humans, and if AGI automates this work, it will steal one more job from us – and not only a job, but the meaning of life.

### Difference between values and desires

Inside the idea of human values is a hidden assumption that there is a more or less stable set of preferences which people can *consciously* access, so people have some responsibility for having particular values, because they can change them and implement them. An alternative view is “desires”: they appear suddenly and out of nowhere, and the conscious mind is their victim.

For example, let us compare two statements:

“I prefer healthy environmentally friendly food” – this is a conscious preference.

“I had a sudden urge to go outside and meet new people” – this is a desire.

Desires are unpredictable and overwhelming, and they may be useless from the point of the rational mind of the person, but may still be useful from more a general perspective (for example, they may signal that it is time to take a rest).

## 3.5. Meta-values: preference about values

### 1. Meta-values as morals

The idea that “human have values” assumes that values present some unstructured set of things. In the same way, a person could say that he has tomatoes, cucumbers and onions. But the relation between values is more complex, and there are values about values.

For example, a person may have some food preferences, but doesn’t approve these food preferences, as they result in overeating. Negative meta-values encode acts of suppressing the normal-level value, or, alternatively, self-shaming. Positive meta-values encourage a person to do what he already likes to do, or foster a value for some useful thing.

Meta-meta values are also possible, for example, if one wants to be a perfect person, s/he will encourage his/her value for health food. The ability to enforce one’s meta-values over one’s own values is called “willpower”. For example, all fights with procrastination is an attempt to enforce the meta-value of “work” over short-term pleasures.

Meta-values are closer to morals: they are more consciously articulated, but there is always practical difficulty in enforcing them. The reason for it is that low-level values are based on strong, innate human drives and have close connections with short-term rewards; thus, they have more energy to affect practical behavior (e.g. difficulties in dieting).

As meta-values are typically more pleasant sounding and more consciously approved, humans are more likely to present them as their true values if asked in social situations. But it is more difficult to extract meta-values from human behavior than “normal” values.

### 2. Suppressed values

These are values we consciously know that we have, but which we prefer not to have and do not wish to let affect our behavior. Such values could be excess sexual interest. Typically, humans are unable to completely suppress some undesired values, but at least they know about them and have an opinion about them.

### 3. Subconscious values and sub-personalities

The idea that “humans have values” assumes that the person *knows* what he has, but this is not always true. There are hidden values, which exist in the brain but not in conscious mind, and can appear from time to time.

Freud was the first to discover the role of the unconscious in humans. But the field of the unconscious is very amorphous and easily adjusts to attempts to describe it. Thus, any theory which tries to describe it appears to be a self-fulfilling prophecy. Dreams may be full of libido symbols, but at the same time represent Jungian Anima archetypes. The reason is that unconsciousness is not a thing, but a field where different forces combine.

Some people suffer from multiple personality disorder, when they have several personalities which take control over their body from time to time. These personalities have different main traits and preferences. This adds obvious difficulty to the idea of “human values,” as the question arises, which values are real for a human who has many persons in his/her brain? While true “multiple personality disorder” is rare, there is a theory that in any human there are many sub-personalities which constantly interact. Such sub-personalities could be called one by one by a psychotherapeutic method called a “dialog of voices,” created by Stones (Stone & Stone, 2011).

The theory behind sub-personalities claims that they can’t be completely and effectively suppressed, and will appear from time to time in the form of some coalesced behavior like jokes (this idea was presented by Freud in his work “Jokes and Their Relation to the Unconscious”), tone of voices, spontaneous acts (like shop-lifting), dreams, feelings, etc.

### 4. Zero behavior and contradicting values

Humans often have contradictory values. For example, if I want a cake very much, but also have a strong inclination for dieting, I will do nothing. So, I have two values, which exactly compensate for each other and thus have no effect on my behavior. Observing only behavior will not give an observer any clues about these values. More complex examples are possible, where contradictory values create inconsistent behavior, and this is very typical for humans.

### 5. Preference about other’s preferences

Human could have preferences about preferences of other people. For example: "I want M. to love me" or "I prefer that everybody will be utilitarian".

They are somehow recursive: I need to know the real nature of human preferences in order to be sure that other people actually want what I want. In other words, such preferences about preference have embedded idea about what I think is the "preference": if M. will behave as if she loves me – is it enough? Or it should be her claims of love? Or her emotions? Or coherency of all three?

## 3.6. Human values cannot be separated from the human mind

### 1. Values are not encoded separately in the brain

The idea “human have values” assumes the existence of at least two separated entities: human and values.

There is not any separate neural network or brain region that presents a human value function (limbic system codes emotions, but emotions are only part of human values). While there is a distinctive reward-regulating region, the reward itself is not a human value (as much as we agree that pure wireheading is not good). Most of what we call “human values” are not only about reward (while reward surely plays a role), but include an explanation for what the reward is, i.e. some conceptual level.

Any process in human mind has intentionality. For example, a memory of smell of a rose will affect our feelings about roses. This means that it is not easy to distinguish between fact and values in some’s mind, and orthogonality thesis doesn’t hold for humans.

The orthogonality thesis can’t be applied to humans in most cases, as there is no precise border between human value and some other information or processes in the human mind. The complexity of human values means that a value is deeply rooted in everything I know and feel, and that attempts to present values as a finite set of short rules does not work very well.

Surely, we can use the idea of human set of preferences if we want some method to approximate a prediction of the person’s approval and behavior. It will offer something, like, say, an 80 percent prediction of human choices. This is more than enough in the prediction of the behavior of a consumer, where we could monetize any prediction above random; e.g. if we predict that 80 per cent people would prefer red t-shirts to green ones, we could adjust manufacturing and earn a profit. (Interesting article on the topic: “[Inverse Reinforcement Learning for Marketing](https://arxiv.org/abs/1712.04612).”)

However, a reconstructed set of values is not enough to predict human behavior in edge cases, like “Sophie’s Choice” (a novel about Nazi camp, where a woman has to choose which of her children will be executed), or a real-world trolley problem. But exactly such predictions are important in AGI safety, especially if we want AGI to make pivotal decisions about the future of humanity! Some possible tough questions: should humans be uploaded? Should we care about animals or aliens or unborn possible people? Should a small level of suffering be preserved to avoid eternal boredom?

Interestingly, humans evolved ability to predict each other’s behavior and choices to some extent, partly limited to the same culture, age and situation, as this skill is essential to effective social interaction. We automatically create some “theory of mind”, and there is also a “folk theory of mind”, in which people are presented as simple agents with clear goals which dictate their behavior (like “Max is only interested in money and that’s why he changed jobs.”)

### 2. Human values are dispersed inside “training data” and “trained neural nets”

Not only are values are not located in some place in the brain, they are not learned as “rules.” If we train an artificial neural net on some kind of dataset, like Karpathy’s RNN on texts, it will repeat properties of the texts (such training includes a reward function, but it rather simple and technical and only demonstrates similarity of output to the input). In the same way, a person who grew up in some social environment will repeat its main behavioral habits, like car driving habits or inter-personal relations models. The interesting point is that these traits are not presented explicitly either inside the data nor inside the neural net trained on it. No single neuron is coding the human preference for X, but behavior which could be interpreted as a statistical inclination to X is resulting from collective work of all neurons.

In other words, a statistically large ensemble of neurons trained on a statistically large dataset created a statistically significant inclination to some type of behavior, which could essentially be described as some “rule-like value,” though this is only an approximation.

### 3. Amorphous structure of human internal processes and false positives in finding internal parts

Each neuron basically works as an adding machine of inputs and is triggered when the sum is high enough. The same principle can be found in psychological processes which add up until it triggers action. This creates difficulty in inferring motives from actions, as there is a combination of many different inputs.

This also creates the problem of false positives in human mind modeling, where a human behavior under some fixed conditions and expectations produces the expected types of behavior, statistically confirming the experimenter’s hypothesis.

## 3.7. Human values presentation is biased towards presenting socially accepted claims

The idea of “human values” is biased towards morality. When we think of human values we expect that something good and high-level will be presented, like “equality” or “flourishing,” as humans are under social pressure to present an idealized version of the self. In contemporary society, someone will not be prized if he said that he likes “kill, rape and eat a lot of sugar”. This creates internal censorship, which could be even unconscious (Freudian censorship) [ref]. Humans claim and even believe that they have socially accepted values: that they are nice, positive, etc. This creates an idealized image of self. But humans are unreflective about their suppressed motives and even actions. Thus, they lie to themselves about the actual goals of their behavior: they do A thinking that the goal is X, but their real motive is Y [Hanson].

Societies with strong ideologies will more strongly affect self-representation of values. Idealized and generalized version of values start to look like morals.

In his book “Elephant in the brain,” Hanson presented a model in which the selfish subconsciously tries to maximize personal social status, and consciously create a narrative to explain the person’s actions as altruistic and acceptable.

## 3.8. Human values could be manipulated by the ways and order they are extracted

The idea that “humans have values” assumes that such values exist independently of some third-party observer who can objectively measure them.

However, by using different questions and ordering these questions differently, one can manipulate human answers. One method of such manipulation is the Ericksonian hypnosis, where each question creates certain frames, and also has hidden assumptions.

Another simple but effective marketing manipulative strategy is the technique of “Three Yeses”, where previous questions frame future answers. In other words, by carefully constructing the right questions we could extract from a person almost any value system, which would diminish the usefulness of such extraction.

This could also affect AGI safety, if AGI has some pre-conceptions of what the value system should be, or even if AGI wants to manipulate values, – it could find the ways to do so.

## 3.9 Human values are, in fact, non-human

Human values are formed by forces which are not humans. First of all, it is evolution and natural selection. Human values are also shaped by non-human forces like capitalism or facebook algorithm and targeted advertising. Being born in some culture, being affected by some books or traumatic events is also random processes out of the person choice.

### Viruses

Many viruses could affect human behavior with the goal to make replication easy. Common cold makes people more social. It seems that toxoplasma infection makes people (and affected mice) less risk averse. See e.g. “[Viruses and behavioral changes: a review of clinical and experimental findings](https://www.ncbi.nlm.nih.gov/pubmed/9155866)”.

There are even more outstanding claims that our microbiome controls human behavior, including food choices and reproduction via production of feromons-like chemicals on the skin. It was claimed that fecal transplants can cure autism via changes in gut microbiome.

## 3.10. Any human value model has not only epistemological assumptions, but also has axiological (normative) assumptions

If a psychological model does not just describe human motivation, but also determines what part of this motivational system should be learned by AGI as “true values,” it inevitably includes axiological or normative assumptions about what is good and what is bad. A similar idea was explored by Armstrong in in “[Normative assumptions: regret](https://www.lesswrong.com/posts/Fg83cD3M7dSpSaNFg/normative-assumptions-regret).”

The most obvious such “value assumption” is that someone’s reward function should be valued at all. For example, during interaction between a human and a snail, we expect that the human reward function (if we are not extreme pro-animal rights activists) is the correct one, and the “snail’s values” should be ignored.

Another type of axiological assumption is about what should be more correctly regarded as actual human values: rewards or claims. This is not a factual assumption, but assumption about importance, which could also be presented as a choice between whom an observer should believe: rationality or emotions, rider or elephant, System 2 or System 1, rules or reward.

There are also meta-value assumptions: should I regard “rules about rules” as more important than my primary values. For example, I often say people should ignore the tone of my voice, I only endorse the content of my verbal communication.

Psychological value models are often normative, as they are often connected with psychotherapy, which is based on some ideas what is healthy human mind. For example, Freud’s model not only presents a model of human mind, but also a model of disease of the mind; in Freud’s case, neuroses.

## 3.11. Values may be not the best route for simple and effective descriptions of human motivation

From the point of view of naïve folk psychology, a value system is easily tractable: “Peter values money, Alice values family life” – but the analysis above showed that if we go deeper, the complexity and problems of the idea of human values grows to the point of intractability.

In other words, the idea that “human have values” assumes that “value” is a correct primitive which promises easy and quick description of human behavior but doesn’t fulfil this promise after close examination. Thus, maybe it is the wrong primitive, and some other simple idea will provide better description – one with lower complexity and that is more easily extractable – than values? There are at least two alternatives to values as short descriptors of the human motivational system: “wants” and commands.

Obviously, there is a difference between “values” and “wants”. For example, I could sit on a chair and not want anything, but still have some values, e.g. about personal safety or the well-being of African animals. Moreover, a person with different values may have similar “wants”. Intuitively, a correct understanding of “wants” is the simpler task.

I can reconstruct my cat’s “wants” based on the tone of her meows. She may want to eat, have a door opened, or to be cuddled. However, reconstructing the cat’s values is a much more complex task which must be based on assumptions.

The main difference between wants and values: if you want something, you know it. But if you have a value, you may not know about it. The second difference: wants can be temporarily satisfied, but will reappear, while values are constant. Values generate wants, wants generate commands. Only wants form the basis for commands to AGI.

## 3.12. Who are real humans in “human values”?

In the idea of human values is assumed that we could easily define who are “humans”, that is morally significant beings. This question suffers from the edge cases, which may be not easy to guess by AGI? Such edge cases:

* Are apes humans? Neanderthals?
* Is Hitler human?
* Are coma patients humans?
* What about children, drug-intoxicated people, Alzheimer patients?
* Extraterrestrials?
* Unborn children?
* Feral children?
* Individuals with autism and victims of different genetic disorders?
* Dream characters?

By manipulating the definition of who is “human”, we could manipulate the outcome of a measurement of values.

## 3.13. The human reward function is not “human values”

Many ideas about learning human values are in fact describing learning based on the “human reward function.” From a neurological point of view and subjective experience, human reward is the activation of some centers in the brain and experiencing qualia of pleasure. But when calculated by analyzing behavior, “human reward function” does not necessarily mean a set of rules for endorphin bursts. Such reward function would mean pure hedonistic utilitarianism, which is not the only possible moral philosophy, or might even mean voluntary wireheading. The existence of high-levels goals, principles and morals means that the qualia of reward is only a part of human motivation system.

Alternatively, a human reward function may be viewed as some abstract concept which describes the set of human preferences in the style of [VNM-rationality](https://en.wikipedia.org/wiki/Von_Neumann%E2%80%93Morgenstern_utility_theorem) (converting set of preferences on to a coherent utility function), but which is unknown to the person.

One assumption about human values is that humans have a constant reward – but the human reward function evolves with age. For example, sexual images and activities become rewarding for teenagers and are meditated by the production of sex hormones. Human rewards also change after we are satisfied by food, water or sex.

Thus, the human reward function is not a stable set of preferences about the world, but changes with age and based on previous achievements. This human reward function is black-boxed from the conscious mind but is controlled by presenting different rewards. Such a black-boxed reward function may be described as a rule-based system.

Possible example of such rules: “If age = 12, turn on sexual reward”. Such a rule generator is unconscious but has power over the conscious mind – and we may think that it is no good! In other words, we could have moral preferences about different types of motivation in humans.

## 3.14. Difficult cases for value learning: enlightenment, art, religion, homosexuality and psi

There are several types of situations or experiments where the existing of a stable set of preferences is clear, like a situation of multiple choices between brands (apple vs. oranges), different forms of the trolley problem, questionnaires, etc. However, there are situations and activities which is not easy to describe in this value language.

*Enlightenment* – many practitioners claim that at some higher meditation states the idea of personal identity and of a unique personal set of preferences, or even of the reality of outside world becomes “obsolete,” seen as wrong or harmful in practice. This may or may not be true factually, but obviously affects preference preferences, when a person appears to have a meta-value of not having a value, and of some form of meaningful non-existence (e.g. nirvana, moksha). How could we align AGI with Buddha?

*Art* – rational thinking often has difficulty understanding art, and many of its interpretations based on outside views are oversimplified. Moreover, a significant part of art is about violence, and we enjoy it – but we don’t want AGI to be violent.

*Religion* – seems to often assign a lot of values to false, or at least uncheckable claims. Religion is one of the strongest memetic producers, but it also includes some theories of motivations, which are not about values, but are based on other basic ideas like “free will” or “god’s will.” Religion also could be seen as an invasive ideology or memetic virus, which overrides personal preferences.

*Psi* – contemporary science denies validity of the parapsychological research, but observations like Jungian synchronicity or Grof’s transpersonal psychology continue to appear and imply a different model of human mind and motivation than traditional neuroscience. Even some AGI researchers, like Ben Goertzel, are interested in psi. In Grof’s psychology, the feelings and values of other human and even animals could influence a person (under LSD) in non-physical ways, and in a more minor form this could happen (if it is possible at all) even in ordinary life.

*Idleness* – and non-goal-oriented states of mind, like random thought streams.

*Nostalgia* – this is an example of a value which has very large factual content. It is not just an idea of pure happiness of the feeling of “returning home.” It is an attraction to the “training dataset”: home country and language, often arising from the subconscious, in dreams, but later taking over the conscious mind.

There are a few other fields, already mentioned, where the idea of values experiences difficulties: dreams, drugs-induced hallucinations, childhood, psychiatric diseases, multiple personality disorder, crime under affect, qualia. And all of this is not just edge cases – it is biggest and most interesting part of what makes us humans.

## 3.15. Human values excluding each other and Categorical imperative as a meta-value

As large part of human values are preferences about other people preferences, they mutually exclude each other. E.g.: {I want “X loves me”, but X don’t want to be influenced by other’s desires}. Such situation is typical in ordinary life, but if such values are scaled and extrapolated, one side should be chosen: either I will win, or X.

To escape such situation, something like Kantian moral low, Categorical Imperative, should be used as a metal-value, which basically regulate how other’s people values relate to each other:

Act only according to that maxim by which you can at the same time will that it should become a universal law.

In other words, Categorical Imperative is something like “updateless decision theory” in which you choose a policy without updating on your local position, so if everybody will use this principle, they will come to the same policy. (See comparison of different decision theories developed by LessWrong community [here](https://www.lesswrong.com/posts/QPhY8Nb7gtT5wvoPH/comparison-of-decision-theories-with-a-focus-on-logical).)

From the Categorical Imperative could be derived some human values like: it is bad to kill other people, as one doesn’t want to be killed. However, the main thing is that such meta-level principle of relation between values of different people can’t be derived just from observation of a single person.

Moreover, most ethical principles are describing interpersonal relations, so they are not about personal values, but about the ways how values of different people should interact. The things like Categorical imperative can’t be learned from observation; but they also can’t be deduced based on pure logic, so they can’t be called “true” or “false”.

In other words, AGI learning human values can’t learn meta-ethical principles like Categorical imperative nor it can’t deduce them from pure math. That is why we should provide AGI with correct decision theory, but it is not clear why “correct theory” should exist at all.

This could also be called meta-ethical normative assumption: some high level ethical principles which can’t be deduced from observations.

# Conclusion

The whole arguments presented above demonstrated that the idea of human values is artificial and not very useful for AGI Safety in its naive form. There are many hidden assumptions in it, and these assumptions may affect AGI aligning process, resulting into unsafe AGI.

In this article, we deconstruct the idea of human values and come to the set of conclusions which could be summarized as following:

“Human values” are useful descriptions, not real objects.

* + - * “Human values” are just a useful instrument for the description of human behavior. There are several other ways of describing human behavior, such as choices, trained behavior, etc. Each of these have their own advantages and limitations.
* Human values cannot be separated from other processes in the human brain (human non-orthogonality).
* There are at least four different ways to learn about a human’s values, which may not converge (thoughts, declarations, behavior, emotions).

“Human values” are poor predictors of behavior

* The idea of “human values” or a “set of preferences” is good at describing only statistical behavior of consumers.
* Human values are weak predictors of human behavior, as behavior is affected by situation, randomness, etc.
* Human values are not stable: they often change with each new choice.
* Large classes of human behavior and claims should be ignored, if one wants to learn an individual’s true values.

The idea of a “human value *system”* has flaws

* In each moment, a person has a contradictory set of values, and his/her actions are a compromise between them.
* Humans do not have one terminal value (unless they are mentally ill).
* Human values are not ordered as a set of preferences. A rational set of preferences is a theoretical model of ordered choices, but human values are constantly fighting each other. The values are biased and underdefined – but this is what makes us humans.
* Humans do not “have” values: Human personal identity is not strongly connected with human values: they are fluid, but identity is preserved.

“Human values” are not good by default.

* Anything could be a human value (e.g. some people may have attraction to rape or violence).
* Some real human values are dangerous, and it would not be good to have them in AGI.
* “Human values” are not “human”: they are similar to the values of other animals and, also, they are social memetic constructs.
* Human values are not necessarily safe if scaled, removed from humans, or separated from each other. AGI with human values may not be safe.

Human values cannot be separated from the human mind.

* Any process in the human mind has intentionality; the orthogonality thesis can not be applied to humans in most cases.
* As the human mind is similar to a neural network trained on a large dataset, human values and behavioral patterns are not explicitly presented in any exact location, but are distributed throughout the brain.
* There is not a simple psychological theory which substantially outperforms other theories when it comes to the full model of human mind, behavior and motivation.
* “Human values” implies that individual values are more important than group values, like family values.
* Not all “human values” are values of the conscious mind.For example, somnambulism, dreams, and multiple personality disorder may look like a human value inside a person’s brain, but is not part of the conscious mind.

We recommend that either the idea of “human values” should be replaced with something better for the goal of AGI Safety, or at least be used very cautiously; the approaches to AI safety which don’t use the idea of human values at all may require more attention, like the use of full brain models, boxing and capability limiting.

# Acknowledgments

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

# Appendix. Table of assumptions in the idea of human values

This table presents all findings of this section is more condensed and structured form. The goal of this overview is to help future scientists to estimate validity of their best model of human values.

See also an attempt to map 20 main assumptions against 20 main theories of human values as a very large spreadsheet [here](https://docs.google.com/spreadsheets/d/1T5VrH_OwTVHpDBoU5KKE8XWCauYj4eNV4rbqiCY9L44/edit#gid=0).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type of assumption** | **The claim of the assumption or alternative points of view** | **Why it may be wrong** | **How AGI could fail based on it** | **Literature** |
| 1. Epistemic assumptions | Do human values actually exist or they are only useful descriptions? | If values are only useful description, there could be many different models | Different AGIs may converge to different models and start a war with each other | [Armstrong](https://www.lesswrong.com/posts/RHvseCkfrYzoHJj7M/learning-values-or-defining-them) |
|  | Which model of human mind is the correct one (or the best model)? |  | An AGI may never be able to learn human values if they do not actually exist - it will tile the universe with different attempts to learn them. |  |
| 2. Axiological assumptions | Which part of the human motivational system is the “right” one? | E.g.: is libido part of the value system?  Dreams? Ideologies? | Case: An autonomous car crashed after assuming the person behind the wheel was awake when he moved while asleep |  |
|  | Human values are good | Are they actually good? What about values of a rapist? | AGI learns a bad set of human values and puts us in a “Game of Thrones”-like world |  |
|  | Suppressed values should be ignored | E.g. desire to commit a crime |  |  |
|  | Individual values are more important than group values. | The values of families, and all other social groups are ignored | Humans are locked alone in one-person simulations: everybody is a king in a world of non-player characters (NPCs) |  |
|  | Adult values are the only right values | The AGI’s engineer idea of values will be biased toward his/her social group | AGI learns only Bay area students’ values, as they have most rational set of preferences and may be also the ones who will construct AGI |  |
|  | Emotions are true source of information about values | In fact, emotions are part of an irrational “animal” brain | AGI will learn to have a suppressed sub-personality like an “internal killer” | Sotala: yes. |
|  | Some human values are politically incorrect |  |  |  |
|  | Psychological theories define what is normal and illness |  | AGI learns values of a mentally ill person, acts weirdly; or AGI doesn’t find any healthy person to emulate |  |
| 3. Factual assumptions | Unity of mind (one brain = one mind) | Hemispheric separation, subpersonalities, dream characters | AGI ignores morally important part of mind, or learns preferences that are not mine |  |
| 4. Meta-ethical | Is moral realism correct and if so, to what extent? |  |  | [Wei Dai](https://www.lesswrong.com/posts/orhEa4wuRJHPmHFsR/six-plausible-meta-ethical-alternatives) |
|  | Which type of ethics is right: deontological, consequentialists, etc. |  |  |  |
| 5. Normative assumptions | Assumption about relationship between reward and rationality; regret |  |  | [Armstrong](https://www.lesswrong.com/posts/Fg83cD3M7dSpSaNFg/normative-assumptions-regret), + another his [post](https://www.lesswrong.com/posts/GTcwKAZ8kdYxyS5G3/normative-assumptions-answers-emotions-and-narratives) |
| 6. Assumptions about are who is considered human | Who are “real humans”? | What about apes? | AGI may have difficulties distinguishing real humans from not-real humans, e.g. images | L. Linsefors post “[Extensive and Reflexive Personhood Definition](https://www.lesswrong.com/posts/z36E4EdMDB4BorqTk/extensive-and-reflexive-personhood-definition)” |
|  | Which humans are the best representative of our values? | Should we include, children, “bad boys,” mentally ill people? |  |  |
| 7.Assumptions about values stability | Human values remain stable, at least during the period of an AGI’s actions | Human values could evolve (in a natural way) | AGI fails to take into account that I stopped being interested in something |  |
| 8. Assumptions about values’ non-contradictory nature | Human values are similar to a set of ordered preferences | Values often contradict each other |  |  |
| 9. Assumptions about reward function | Are human values equal to the reward function? |  |  |  |
|  | Is human reward = intensity of pleasure? |  |  |  |
| 10. Assumptions about value complexity | Are human values = simple reward –concept relation, or human values = a complex thing with many connotations? |  | AGI fails to grasp the full complexity of human values | Sotala, Yudkowsky |
|  | What is the role of natural language in human values? |  |  |  |
| 11. Assumptions about social role of values | Why the values of a single human, not of a family or nation? |  |  |  |
|  | Memetic nature of values |  |  |  |
|  | Ideologies and model |  |  |  |
|  | Group bonding role of values |  |  |  |
| 12. The idea of human values is useful | Human values are good predictors of behavior | We can’t predict behavior based only on values, as situation and policy affect behavior | Difficulties in real life |  |
|  | Human values can be separated from other processes in the mind; there is an obvious border between values and policy |  |  |  |
|  | Other descriptions of motivation are possible, like fundamental human rights, preferences or wishes |  |  |  |
| 13.  Relation of human values with their representation | Human values are correctly presented by one of the following processes:   * thoughts, * declarations, * behavior, * emotions * reward activation, * choices, * utility function, * free energy | We should arbitrarily choose one channel of communication with human mind as the correct one | AGI either ignores my thoughts or my emotions. |  |
|  | Humans know their values | In most cases, humans are clueless about their preferences and if asked produce gibberish | The approval approach will not work if a human can’t decide whether or not to approve something | Wei Dai |
| 14.  Human values exist as a system | Values are ordered | Ordering of values changes from choice to choice |  |  |
|  | Human terminal and instrumental values are different | Instrumental values become terminal (money) |  |  |
|  | Values define personal identity | Values are transitory phenomena |  |  |
|  | Human values could be separated from each other | Values balance each other and are safe only in combination | Separated values become paperclippy maximizers (e.g. money) |  |
|  | Human values could be safely scaled |  |  | Wei Dai |
| 15. Human values are unique and specific | There are some values which are not human values | Everything could be some person’s value |  |  |
|  | Human values are different from animal values | Any human value could be shown to derive from some animal drives |  |  |
|  | There are mental processes in the which are not value-related (human orthogonality thesis) |  |  | Armstrong, (R, p) model. |
| 16. Human mind could be known | There is a correct theory of human mind which substantially outperforms other theories in all aspects | Theories are instrumental and context dependent; only full brain would be a correct model of mind |  |  |
| 17. Human values could be extracted from human mind | Convergence assumption 1: Human values can be presented in one and only one correct way. |  |  |  |
|  | Convergence assumption 2: Human value system extrapolation will converge to only one extrapolation which solves all contradictions |  |  |  |
|  | Only model-based behavior is based on values |  |  | Muehlhauser |
| 19. Value are well-defined | Values are certain and can be compared |  |  | Armstrong, “[Our values are underdefined, changeable, and manipulable](https://www.lesswrong.com/posts/KCg7NeKQ7MycXWpYd/our-values-are-underdefined-changeable-and-manipulable)”  Wei Dai |
|  | Values don’t depend on the results of an action | But: rationalization;  wishes disappear after they’re fulfilled. |  |  |
|  | My values are mine | But: indoctrination and conformity |  |  |
|  | Values are stable | Values can be manipulated in almost every direction |  | [Armstrong](https://agentfoundations.org/item?id=1489) “Optimisation in manipulating humans: engineered fanatics vs yes-men” |
| 20. Values define human choices | Knowing an individual’s values, we can predict choices | But: Situation has more predictive power than values |  | Book “The person and the situation: Perspectives of social psychology” by Ross and Nisbett. |
| 21. Assumptions about extrapolation of future values based on current values | Future values evolution could be univocally predicted based on current set of values and set of other personal traits. | [Ricraz](https://www.lesswrong.com/posts/suxvE2ddnYMPJN9HD/realism-about-rationality): “And it’s plausible that there’s no canonical way to “scale me up” | Wrong CEV | [Realism about rationality](https://www.lesswrong.com/posts/suxvE2ddnYMPJN9HD/realism-about-rationality) |
|  | Values of different people could be converged in one meta-value |  |  |  |
| 22. Values and other important philosophical ideas | Values and consciousness are not related | Phenomenological consciousness will appear in any complex agent | AGI turns humans into *philosophical zombies* (minds without qualia) | [G. Worley](https://mapandterritory.org/ai-alignment-and-phenomenal-consciousness-2ca23de6aebd) |
|  | There is no free will; values predict actions | The full complexity of mind, situation and randomness is necessary for correct prediction; there are no shortcuts | Humans are standardized and suppressed |  |
|  | There is no soul |  |  |  |
|  | There is no God  and no Buddhist enlightenment |  |  |  |
|  | There is no “quantum mind” or other unknowns in the brain. |  |  |  |
|  | Human values can be presented as a numerical, mathematical function | Values depend on exact language because of different webs of connotation |  | [Armstrong](https://www.lesswrong.com/posts/XApNuXPckPxwp5ZcW/bridging-syntax-and-semantics-with-quine-s-gavagai) |
|  | Simplified models are a useful instrument to come to full solution | All models have many assumptions and could create the illusion of knowledge |  |  |
|  | AGI which learns human values is safe. | AGI is must learn values in a safe way, but it can’t do it before it knows human values. Chicken or egg problem | Dangerous value learners | [Cake or death](https://www.lesswrong.com/posts/6bdb4F6Lif5AanRAd/cake-or-death), Armstrong  [Dangerous value learners](https://www.lesswrong.com/posts/hzEaasJyQsutYDNfN/possible-dangers-of-the-unrestricted-value-learners)  Turchin |
|  | Human values belong to the particular human | A person can be influenced by the values of another person in the same room; hypnosis and suggestion. |  |  |
| 23. Assumption related to utilitarianism | Relation between observer-moments doesn’t matter; only amount of pain in each observer-moment counts | Humans may have non-linear preferences about relationship of short and intense pain vs. long and weaker pain |  |  |
|  | The problem of subjective intensity of pain and pleasure qualia | There is a threshold of unbearable pain, and it can’t be measured only by the electrical intensity of brain activity |  |  |
| 24. Orthogonality thesis | Human orthogonality thesis: human mind could be separated into two parts: motivational and rational | Any thought is intentional;  Capabilities become instrumental goals or affect human choices;  Values are not stored in just one part of the brain |  | Armstrong (R,P) :  reward + policy model |
| 25. Meta-ethics: consequentialism | Consequentialism is the best meta-ethics, and humans should act according to it | From the consequentialist point of view, it is better that most people will be deontologist, as typically humans are unable to calculate consequences of their actions and will screw expected utility in their own interests |  | Nature article by Baez shows that terrorists are more likely to be consequentialists |
|  | Humans are similar, and there is a universal set of human values | Different people are unique and could have a completely individual set of preferences |  |  |
|  | Human values are good because they are human | Nietzsche: “Man is something that shall be overcome.” | AGI prevents evolution of posthumans |  |
|  | Human values could be presented as a perfect mathematical model | Only whole-brain emulation may work as a model of human values, and only probabilistically. |  |  |
|  | Values evolve naturally, based on their own logic | Values evolve under control of external (genetic) program, e.g. children like to play, but stop liking it at some age. |  |  |
|  | Values are located “inside” humans | Values exist outside humans: in advertising, ideologies, training data sets |  | Values are narratives told in our training datasets (advertising, tales, etc.) |
|  | Values cause behavior | Having good values is signaling |  |  |
|  | Any human value assumes some ideas about who its subject is, and thus some theory of mind. |  |  |  |