The neuroscience of ethics as a basis for moral behavior interventions?

An assessment of moral enhancement

Markus Christen,

University of Zurich

"Talk Package Insert"

Will this talk answer the question whether debunking arguments are correct or not?

No

So why are you here?

Two reasons:

- To discuss some normative consequences of a shared assumption of "debunkers" and "moral enhancers" – namely that biological processes (with an evolutionary history) "underlie" moral behavior / judgments / decision making.
- To present data that outline possible cultural adaptations of moral intuitions (that may have a biological foundation).


The Moral Brain


MORALITY
BEHAVIOR ENHANCER
NEW!
EXTRA CONSCIENCE

Overview

Conceptual issues

"Morality", "Enhancement", "Neuroscience of Ethics" and and a definition of moral enhancement.

A deeper look at the problem

Morality in the physical world, moral change, levels and means of interventions.

Pro and Con arguments regarding moral enhancement

Analyzing the current debate and practical of moral enhancement based on brain intervention experiences

Outlining a way for moral enhancement

What we plan to do and where the problems are.

Conceptual issues

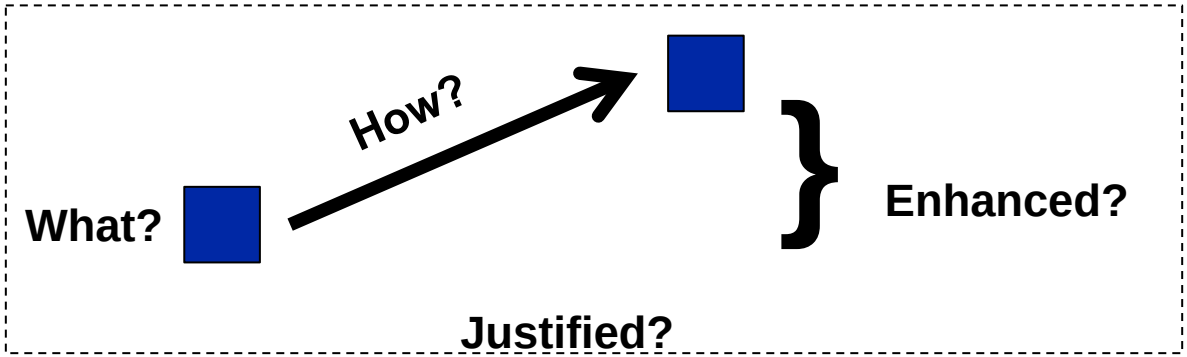
What is “Morality”?

The term “morality” refers to a lot of different entities and aspects:

- Set of rules, codes, practices that differ in degree of explicitness.
- Systems/theories for justifying rules, codes etc. that differ in their degree of plausibility.
- Single beliefs, motives and other “cognitive entities” that differ in their “moral load”.
- Systems/classifications/ontologies of such beliefs, motives etc. that structure the “moral space” in various ways.
- Behavioral dispositions (“virtues”) that lead to manifest behaviors with a certain probability.
- Biological/psychological mechanisms that “underlie” dispositions and actual decisions & behaviors.

What is “Moral Enhancement”?

The idea of moral enhancement (ME) involves the clarification of several different aspects:



Furthermore, one should demonstrate how this approach differs from the rich tradition of “making humans better” in terms of moral education, social reform, setup of legal systems, etc.

What is the “Neuroscience of Ethics”?

A branch of “social neuroscience” that focuses on the “underlying mechanisms” of human moral behavior with the following claims:

- One can relate different “aspects” of morality to partially separable neuro-cognitive “architectures”.
- These “architectures” are shaped by evolution and determine “boundaries” of our moral dispositions.
- Ontogenetic damages to this “architecture” can disrupt the moral competences of agents.
- All aspects of morality involve “affect” – and this may frame competences like “moral sensibility” on a subconscious level.

Methodological and science-sociological aspects are often neglected when debating the plausibility of these claims.

Defining the topic “moral enhancement” (1)

1) **Precondition (claim or demand):** The current state of knowledge and the toolset available makes it likely that in the near future we sufficiently understand the “(neuro)biological underpinning” of human moral behavior.
2) **Focus:** The focus of ME is the individual; the target of the intervention are her biological mechanisms or psychological competences and not aspects that are part of the “mind life” of the agent. This does not exclude that the agent can deliberate on how the intervention may change his “mind life”.
3) **Means:** ME is performed by means that operate directly on the level of biological or psychological processes and that do not translate one-to-one into beliefs etc. that are relevant of the “moral mind life” of the agent.

Defining the topic “moral enhancement” (2)

4) **Improvement:** ME should *at least* diminish the likelihood of uncontroversial “bad behaviors” by changing underlying beliefs, motives, dispositions etc. And this improvement should be achieved faster or with more certainty compared to traditional means like moral education.

5) **Justification:** Human moral psychology is not able to cope with the challenges of modern, technological civilizations. Thus, in a same way as we enhance other human capabilities through technology, we should also enhance our moral capabilities.

A deeper look at the problem

Point 1: There are several kinds of "physical representations" of morality

ME assumes that the central physical representations of morality are **agent-internal**, i.e. they consist of the biological processes that underlie the psychological competences of the agent.

But they may be other kinds of "physical representations" of morality:

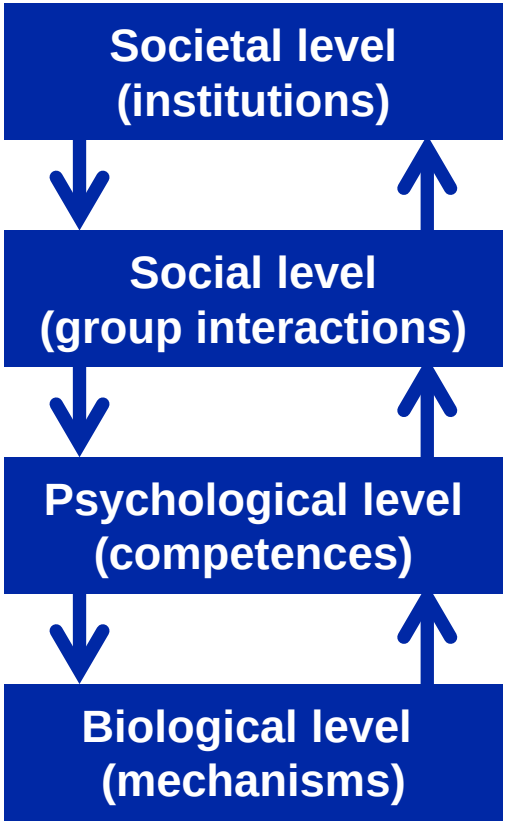
1) **Agent-external:** There is a plethora of stimuli and boundary conditions that shape the way people make moral decisions (priming effects, situationalism, etc.)
2) **Borderline cases (in the sense of the "extended mind hypothesis"):** Tools, instruments etc. agents may regularly use when making moral decisions (e.g., cilice; maybe today smartphone apps)

Point 2: Even a “mechanistic” view on morality has to take into account levels and feedback

There is not a “one way road” from intervening on the biological system all the way up to a desired social change.

Changes on all levels will “feed back” into other levels (this implies difficult questions like downward causation).

What looks like a simple “feed-forward” systems is actually a complex system that required a much more expanded view of the system that is changed by a ME intervention.



Point 3: The interplay between “biological” and “cultural” moral progress/decline is unclear

If intuitive measures for moral progress like the degree of physical violence within a society are taken as empirical markers for moral progress, then most human societies have experienced moral progress in the last few centuries (Pinker 2011).

However, if the development of the human species is taken as relevant time scale, humanity may have experienced a global moral decline.

Thus, two notions of moral progress should be distinguished: a “biological notion” that refers to the inherited capacities that are typical of the evolutionary niche of homo sapiens; and a “cultural notion” that relates moral progress to dealing with an increasing diversity of temptations and possible wrongdoings in a human social world whose complexity accumulates in time.

Point 4: Evaluating the success of moral enhancement involves measurement problems

In order to evaluate whether moral enhancement succeeds, we have to know what has improved.

So, what should we measure?

- Changes in probability distributions of selected behaviors (i.e. the methodology that would be used when developing ME tools)? Here we would have the problem of ecological validity.
- “Macro parameters” like societal degrees of violence? This involves a difficult attribution problem.
- “Meta parameters” like an increased ability to judge the adequateness of certain moral rules for certain contexts?
- And what about measuring the costs of ME?

University of Zurich[UZH]

Institute of Biomedical Ethics

Pro and Con arguments regarding Moral Enhancement

Outlining (valid) pro ME arguments

1. **Duty to explore:** Given the assumption – which is hard to deny – that there are biological/psychological mechanisms that underlie moral behavior, there is a duty to explore these mechanisms in a similar way as we explore other mechanisms of unwanted states (like diseases), i.e. aiming to gain working intervention strategies.
2. **Right to enhance:** At least for people who want to overcome their moral weakness, there is no reason to refrain from means that help them to change their motives in a better way through means of ME. Such an approach actually could increase the freedom of these persons (in a similar way as a depressed person is accessible for behavioral therapy through medication; i.e. counteracting the argument of Harris).

Outlining contra ME arguments (1)

1) **Regarding preconditions:** There is no reason to believe that moral behavior is coupled in any significant way to biological or psychological processes within persons. Moral enhancement through biological means is likely "not real" moral enhancement. Only when someone has been convinced by the right reasons, we can speak of "true enhancement" (Harris).

This is likely a weak argument:

- **The is ample evidence that aberrations in biological processes can lead to unwanted moral behavior.**
- **Biological ME can still be an object of deliberation, in particular regarding the decision to take an enhancer.**

Outlining contra ME arguments (2)

2) **Regarding focus:** Focusing on the biological mechanisms of moral behavior in individuals is simply the wrong approach for inducing changes in moral behavior – in particular for tackling global problems (like tragedy of the commons). The right level of intervention is on the social and institutional level.

This argument has some strength:

- **ME is strategically wrong: Even when we can successfully manipulate the biological mechanisms underlying moral behavior, we do not know how they unfold on the higher levels and what feedback effects will happen.**
- **ME is tactically wrong: We may have not enough time to wait for working ME strategies.**

Outlining contra ME arguments (3)

3) **Regarding means:** There are no safe means that intervene into the mechanisms such that moral changes are reliable and free from severe side effects. All candidates put forward so far (e.g. selective serotonin reuptake inhibitors) are unlikely to succeed (Wiseman).

This argument has some strength:

- **The experiences made with ataractics make it plausible that pharmacological interventions are likely to have side effects and may fail in a considerable number of cases.**
- **However, it is not impossible that other means of interventions and training may have more reliable effects, in particular if they are combined.**

Outlining contra ME arguments (4)

4) **Regarding improvement:** How do democratic societies define about the traits that should be enhanced? Different cultures may enhance different traits (e.g., autonomy and personal responsibility vs. generosity and compassion). Furthermore, ME may reinforce natural variability in moral competences and raise difficult questions for the democratic organization of societies (e.g., should morally enhanced people have more decision making rights?). Furthermore, ME supports dangerous narratives: elites that believe they are superior / sociobiology (Sparrow)

These are indeed the most disquieting arguments. But they mainly come into play when ME is object of a society-wide intervention (which must not be the case)

Outlining contra ME arguments (5)

5) **Regarding justification:** Its not clear whether increased moral competences are actually positive. Negative states, such as emotional overload, exploitation or moral distress could result from, e.g., higher level of moral sensitivity (Weaver). Similarly, a society with many morally sensitive people might be considered a good society, but people living in such a society might not necessarily be happy, because they worry about every immoral and unfair deed (Morioka).

This argument is worth considering for any ME intervention, but depends on the scope and width of the intervention.

Some additional practical issues

1) **How to justify research on means for ME?** Given the current setting of biomedical research, it is unlikely that such research will enter a “clinical phase” unless the problem is conceptualized as a “disease” (which itself poses difficult questions).
2) **When to start with moral enhancement?** It is likely that such interventions may have to be made in children such that they have long-lasting effects. This poses difficult questions regarding informed consent and the like.
3) **What about side effects?** There is a rich set of experiences where brain interventions had side effects that were hard to evaluate and to cope with, e.g. in case of deep brain stimulation.

Summary (so far)

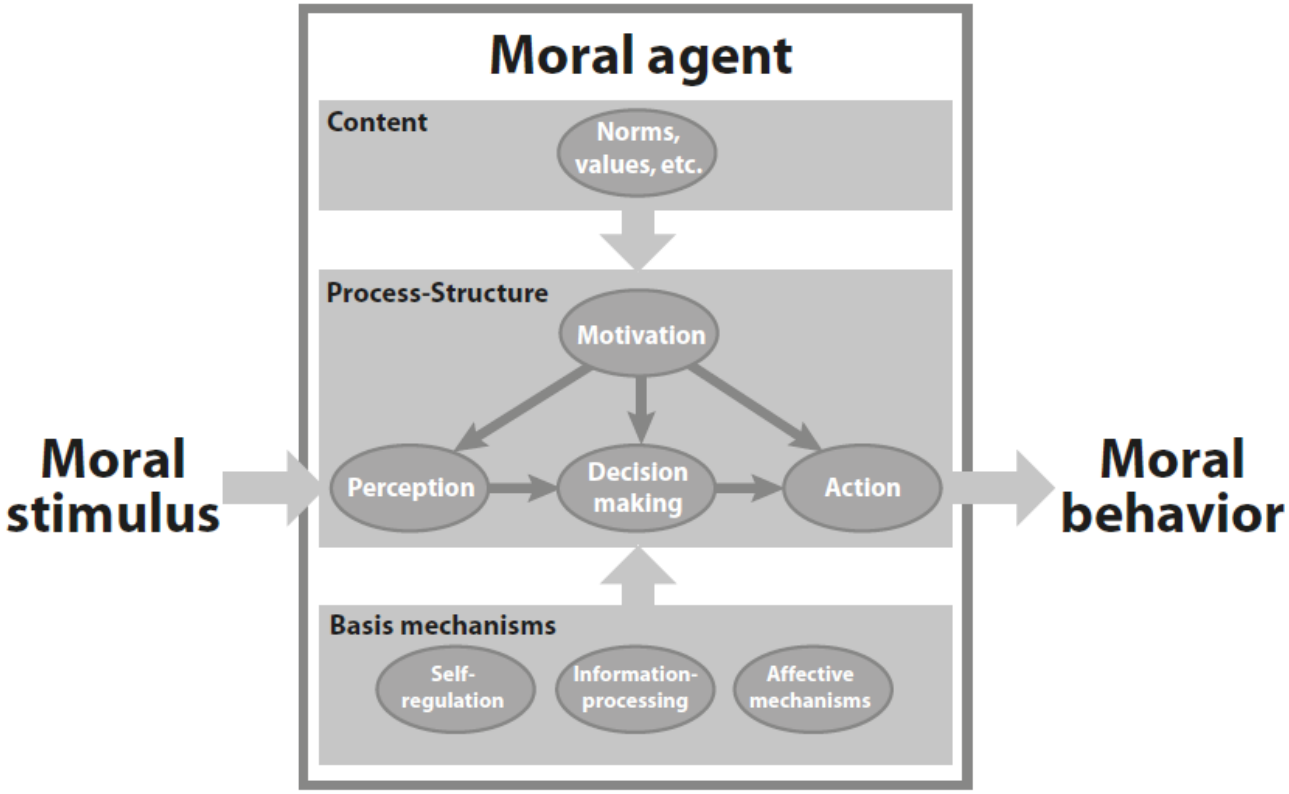
1) It's reasonable to advance research on biological underpinning of moral behavior, in particular in cases of more or less uncontroversial "bad behaviors".
2) In the meantime, working ME interventions are likely to have relevant side effects such that individual risk-benefit assessments will be needed.
3) The overarching goal of "improving society" through moral enhancement involves risks hard to control on various levels.
4) Justifiable moral enhancement should include "personal deliberation" regarding "self-manipulation".

Outlining a feasible way for moral enhancement:

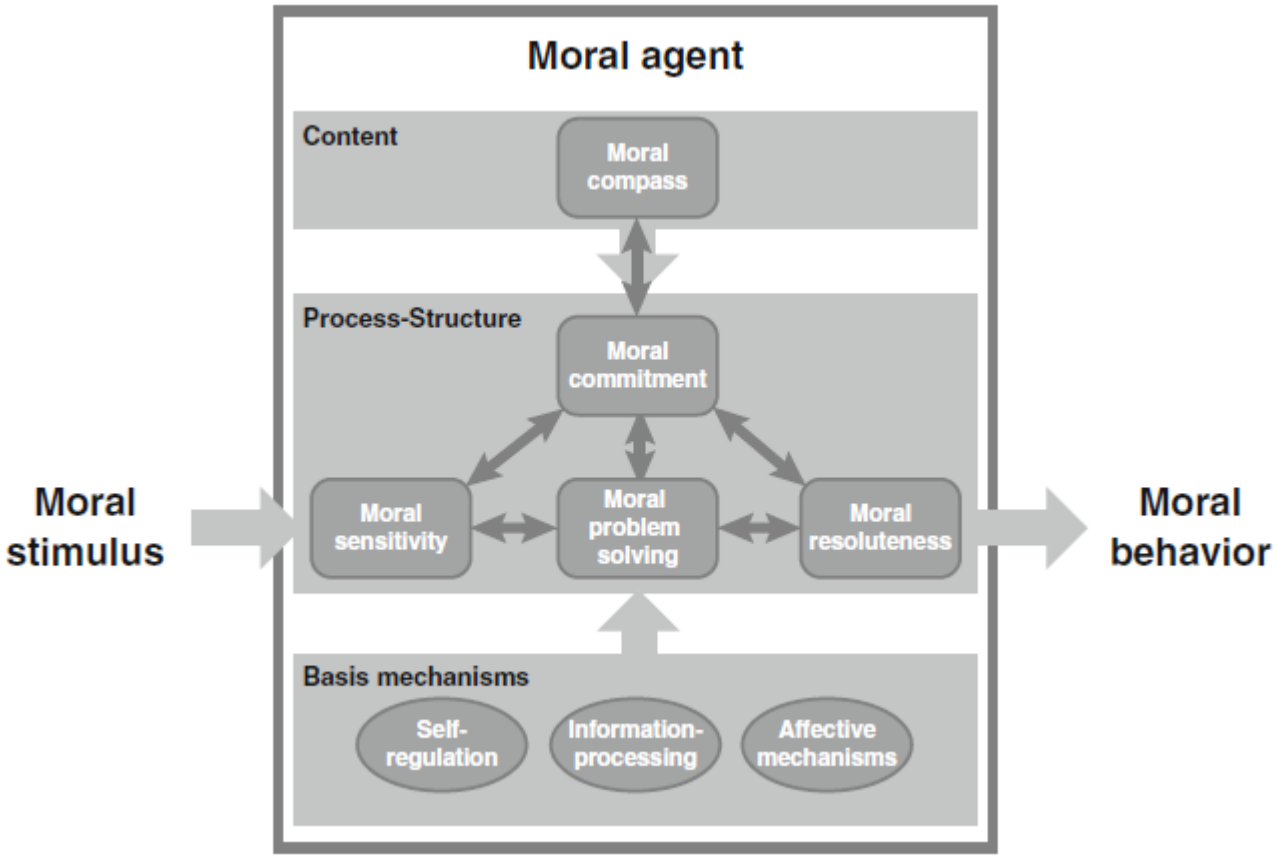
Go up one level (psychology instead of biology)

Our stage model of moral behavior

We work with an adaptation of classical stage models of moral decision making (Rest 1986, Narvaez 2005):



Our working model: Moral Intelligence



Tanner & Christen, 2013

Enhancing moral sensitivity (1)

Moral sensitivity (also referred to as moral awareness, ethical sensitivity/sensibility) is commonly defined as the ability to recognize moral issues when they arise in practice

It includes being responsive to the need of others and envisaging whether a course of action can harm or help others or violates internalized moral standards or codes that govern professional conducts.

In fact, lack of moral sensitivity – also called *moral blindness* – is likely to have far-reaching implications. Without the initial recognition that a moral problem is at stake, no moral problem will exist for the individual and therefore also no need to enter into moral problem-solving

Enhancing moral sensitivity (2)

Researchers found that “morally blind” people can have (morally) best intentions but nonetheless behave in contradiction to their own values and principles, without being aware of it (Bazerman & Tenbrunsel, 2011).

Examples of cognitive mechanisms making people unaware of moral issues are, to name just a few, the slippery slope effect (inability to “see” moral problems when they develop gradually rather than abrupt; Gino & Bazerman, 2009), psycho numbing (loss of compassion when considering a group of victims rather than a single identified victim; Small, Loewenstein & Slovic, 2007), and self-deception or moral disengagement (distortion of reality to maintain a positive self-image and to justify unethical behavior; Detert, Trevino & Sweitzer, 2008).

Enhancing moral sensitivity (3)

Some evidence that moral sensitivity and behavior is positively linked is provided by Tenbrunsel and Messick (1999). They found that participants behaved more cooperatively when they perceived the situation as an ethical rather than a business decision.

Jordan (2009) found that business managers were (compared to academics) less likely to detect and to recall moral-related issues than business-related issues in morally ambiguous vignettes.

Studies by Reynolds (2006) revealed that individuals with a deontological predisposition demonstrated a greater capacity in recognizing both harm and behavioral norm violation, while people with a utilitarian predisposition were only sensitive to harm. Furthermore, Reynolds (2008) has shown that moral attentiveness is positively related to (self-reported) moral behavior.

Moral Sensibility

We are currently developing tools for each component of our model of moral intelligence aiming to integrate them later in a unifying game setting adapted to specific social spheres (medicine, finance).

One of them is moral sensibility, the ability to recognize and identify a moral issue. This requires, as a first step, to identify what are common moral and non-moral value orientations in these social spheres and how are they generally rated as being moral or non-moral.

For medicine, we identified 14 value orientations. Exemplars have been rated (by students and professionals, N=317) along 4 dimensions:

1. moral/universal – non-moral/universal
2. community-oriented – self-oriented
3. collaborative – competitive
4. consequentialist – principle-focused

Survey – Step 1

In a first step, we have – based on a literature review, expert interviews, and a small survey among health professionals (N=17) – identified 14 values that are of considerable importance within medicine:

- autonomy (Autonomie)
- care (Fürsorge)
- cost-effectiveness (Wirtschaftlichkeit)
- feasibility (technischer Imperativ)
- honesty (Ehrlichkeit)
- integrity (Integrität)
- justice (Gerechtigkeit)
- loyalty (Loyalität)
- non-maleficence (Nichtschaden)
- performance (Leistung)
- professionalism (Professionalität)
- reputation (Reputation)
- respect (Respekt)
- responsibility (Verantwortung)

The notion of “value” referred to goals individuals and/or institutions consider being positive to achieve, they are not necessarily moral values. For all values we also collected typical, widely shared characterizations.

Survey – step 2

In a second step, we created an online survey (in German) consisting of two parts: In the first part, the participant provided demographic information and information on their work experience in medicine. In the second part, the participants answered for each value two types of questions: They rated each value along four dimensions using a 6-point Likert scale, and they evaluated the accurateness of five statements (one distractor statement) intending to characterize the given value.

Dimension	Description of left endpoint	Description of right endpoint
A: universal-moral – conventional-non-moral	A value is "moral" if it claims to be universally valid within a society and its corresponding actions are judged as right or wrong.	A value is "non-moral" if it is not claimed that the value is universally valid within a society and if corresponding actions are not subject of evaluations as right or wrong.
B: self-oriented – community-oriented	A value is "self-oriented" if it refers to the priority of personal goals, personal interests or the individual.	A value is "community-oriented" if it refers to the goals of a community, common interest or the relation among individuals.
C: cooperative – competitive	A value is "cooperative" if it refers to the collaboration or communication between human beings or institutions.	A value is "competitive" if it refers to the competition or rivalry between human beings or institutions.
D: consequentialist – principle-focused	A value is " consequentialist " if it focuses to the evaluation of consequences of an action when the value is used to valuate actions.	A value is "principle-focused" if it focuses on the legitimacy of the act itself when the value is used to valuate actions.

Methodology (2)

We used Mann-Whitney and Kolmogorov-Smirnov as two complementary nonparametric tests (former has a higher power for refusing the null-hypothesis, latter is more sensitive for the form of the distribution, e.g. bimodality). Based on these tests, we performed two types of classifications for each dimension:

1) We classified two values X and Y as being in the same group, if either the Mann-Whitney- or the Kolmogorov-Smirnov-test does not distinguish them (i.e., $p>0.05$) for a specified dimension.
2) We used superparamagnetic clustering either the p-values of the Mann-Whitney- or the Kolmogorov-Smirnov-test as similarity measure for each dimension.

In this way, two values X and Y could be maximally 12 times (2 measures x 3 dimensions x two classification methods) in the same group. In this way, a count matrix is generated in which each matrix element stands for the number of times the two associated values have been put in the same group

Result (1)

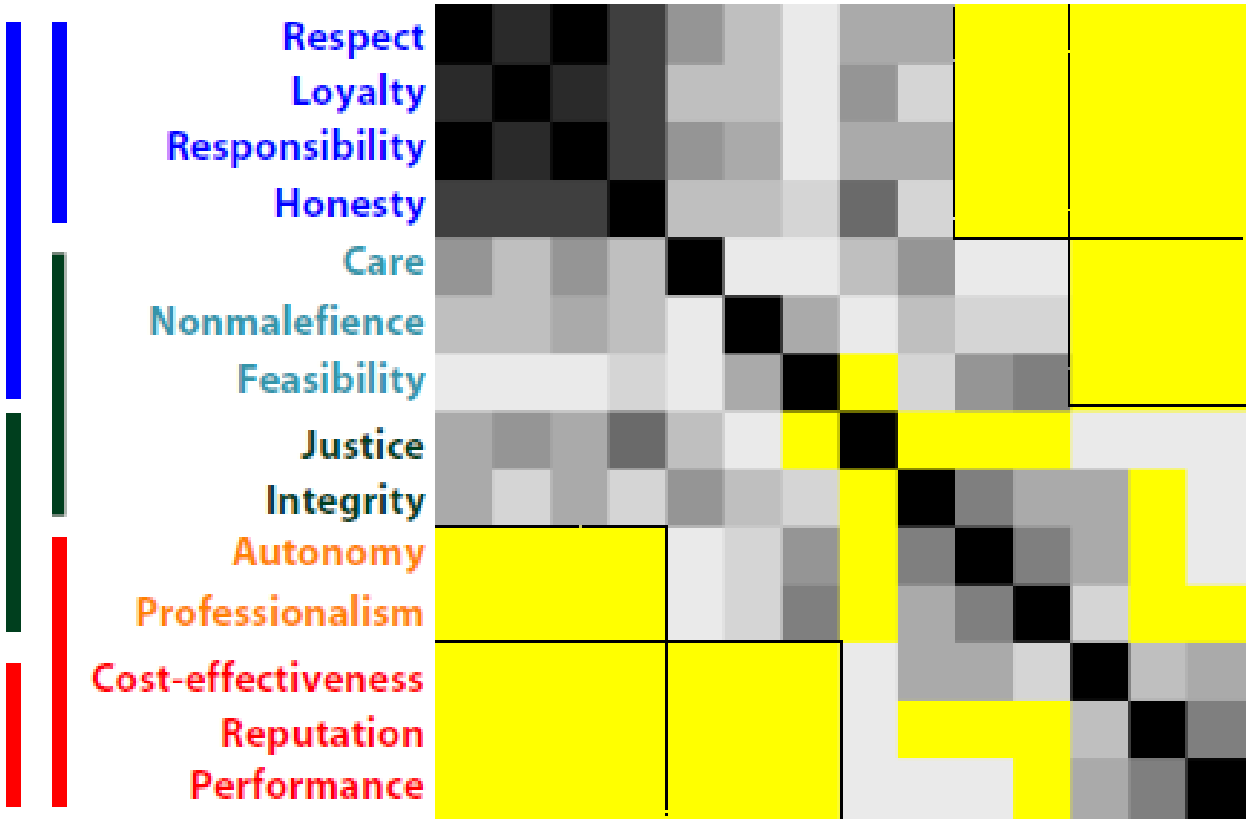
we calculated for each single value using univariate linear regression, whether a pairwise correlation among the four dimensions can be detected:

- The dimensions B and C were significantly correlated ($p<0.05$) for all values (mean estimated slope: -0.32)
- The dimensions A and C were significantly correlated in 13 out of 14 cases (mean estimated slope: 0.27)
- The dimensions A and B were significantly correlated in 10 of 14 cases (mean estimated slope: -0.25).
- The dimension D was in the majority of the cases not correlated to any other dimension (A-D: in 7 cases, B-D: in 9 cases, C-D: in 11 cases).

We thus can conclude that the dimensions A, B and C describe the distinction between “moral” (universal with reference to right and wrong, community-oriented, and cooperative) and “non-moral” (non-universal and no reference to right and wrong, self-oriented, competitive) values as predicted, whereas D is (as expected) not attributable to this distinction.

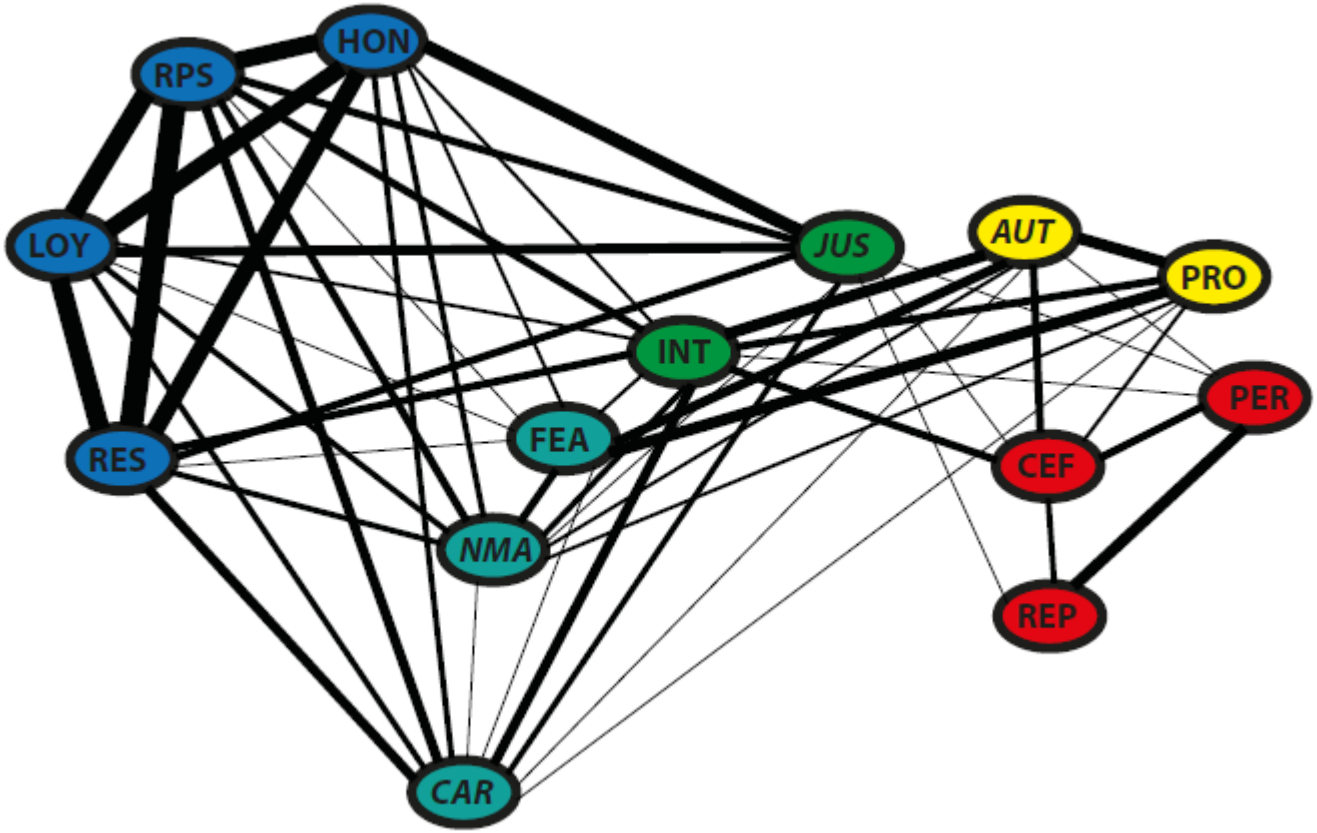
Result (2)

Count matrix (ordered)



Result (3)

Network representation of count Matrix



Interpretation (1)

The fact that care, justice, non-maleficence, and (in particular) autonomy fall off in their “morality rating” by health professionals and students may be surprising. One may consider this as an indication of a failure to convey desired normativity of values to professionals who should work with them.

This however, together with the fact that other values perceive higher “morality ratings”, indicates a more important consequence, namely that the principles are not in the same way grounded in human moral psychology as other moral values.

This raises two questions:

1) How can principles, which are inherently not as moral-laden as assumed, guide health care providers in conflict situations to find a helpful – and for their part “moral” orientation – that would render action guidance?
2) Why do values like ‘care’ or ‘justice’, whose grounding in evolved human moral psychology is likely given the current state of knowledge, receive lower “morality ratings”?

Interpretation (2)

Our findings indicate that the cultural practice of dealing with the principles in a specific way in biomedical ethics (namely as instruments to teach ethics to students and health professionals) weakens their initial appeal to serve as moral guidance, i.e. our evolved moral psychology is more flexible than expected. But on the other hand, this also erodes to some degree the foundation of the principles in common morality.

A prediction from our findings is that medical professionals will identify, e.g., violations of honesty or respect in specific practical clinical problems faster and more reliable compared to violations of autonomy in particular The multiple facets of autonomy in medical decision problems (Schwab & Benaroyo, 2009) make such a prediction plausible.