Personality and Individual Differences 177 (2021) 110731

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

The dark path to eternal life: Machiavellianism predicts approval of mind upload technology

Michael Laakasuo [a,*], Marko Repo [a], Marianna Drosinou [a,b], Anton Berg [a,c], Anton Kunnari [a,b], Mika Koverola [a], Teemu Saikkonen [a,d], Ivar R. Hannikainen [e], Aku Visala [c], Jukka Sundvall [a]

[a] University of Helsinki, Faculty of Arts, Department of Digital Humanities, Cognitive Science, Vuorikatu 3A, 5th floor, Room 541, 00014, Finland
[b] University of Helsinki, Faculty of Medicine, Department of Psychology and Logopedics, Haartmaninkatu 8, Biomedicum Helsinki 1, PL 64, 00014, Finland
[c] University of Helsinki, Faculty of Theology and Religious Studies, Vuorikatu 3, 000014, Finland
[d] Zoological Museum, Biodiversity Unit, University of Turku, Vesilinnantie 5, 20500 Turku, Finland
[e] Universidad de Granada, Department of Philosophy I, Calle del Profesor Clavera, s/n, Campus Universitario de Cartuja, 18011 Granada, Spain

ARTICLE INFO

Keywords:
Moral psychology of transhumanism
Utilitarianism
Machiavellianism
Dark triad
Mind upload
Sexual disgust
Moral cognition

ABSTRACT

Mind upload, making a digital copy of one's brain, is a part of the transhumanistic dream of eternal life and the end of suffering. It is also perceived as a viable route toward artificial general intelligence (AGI). However, AI safety research has alerted to one major risk in creating AGI by mind upload: namely, that mind upload technology could appeal primarily to callous and selfish individuals who then abuse this technology for their personal gain—and, potentially, at a considerable cost to the welfare of humankind. Therefore, it is important to understand whether people's acceptance of mind upload is associated with pathological and/or antisocial traits. To this end, the present research examined whether individual differences in Dark Triad traits predict attitudes toward mind upload in a sample of 1007 English-speaking adults. A pre-registered structural equation model revealed that Machiavellianism (but not psychopathy) was associated with favorable views about mind upload, both directly and indirectly through utilitarian moral attitudes. These results therefore substantiate the concerns voiced by AI safety researchers—namely, that mind upload technology could be adopted disproportionately by individuals with an antisocial personality.

1. Introduction

According to transhumanist philosophy, contemporary humans are not the end-state of evolution. Humans, therefore, ought to enhance their current capabilities by all technological, political and educational means—without, of course, violating core individual freedoms. Transhumanist projects have thus far advocated to enhance what is considered "human nature" using various technologies, which include gene-editing, cryogenics, brain-machine interfaces, cognitive enhancement and cybernetic or other bodily extensions and implants (Bostrom, 2005; O'Connell, 2017).

Mind upload (emulating human minds in a digital medium) is the ultimate dream of transhumanism. It is also presaged as one of the most likely paths toward a human-level artificial intelligence capable of flexible, autonomous goal setting (also known as '*artificial general intelligence*' or AGI), and even toward artificial superintelligence (i.e. exceeding human capabilities; see Sandberg & Bostrom, 2008; Cappuccio, 2017; Chalmers, 2016; Kurzweil, 2012; O'Connell, 2017; Pigliucci, 2014; Tegmark, 2017). For instance, in whole brain emulation (e.g. Hanson, 2016), the cellular structure of the brain would be digitally duplicated using silicon-based platforms. Such technologies are already being tested, and several large scale initiatives mapping the brain connectome are underway (see Seung, 2012). Further progress has been made in the reconstruction of non-human minds: first, the whole nervous system of the nematode *C. elegans* was mapped and subsequently released in 2019 (Cook et al., 2019), and robots simulating the *C. elegans* nervous system are capable of autonomous motion (Busbice, 2014; Neiva & Goiveia, 2017). Second, the Blue Brain project recently digitized the whole cortex of a mouse (Reimann et al., 2019), and digitized parts of rat brains fire similarly to their biological counterparts (Markram et al., 2015).

The prospect of consolidating these technological developments can be awe-inspiring, but at the same time terrifying and immoral to many (Harari, 2015; Waytz & Young, 2019). Better known transhumanistic

* Corresponding author.
E-mail address: michael.laakasuo@helsinki.fi (M. Laakasuo).

https://doi.org/10.1016/j.paid.2021.110731
Received 16 November 2020; Received in revised form 1 February 2021; Accepted 1 February 2021
Available online 15 March 2021

projects, such as cognitive enhancement, have already been shown to inspire moral condemnation in the general public, motivated by concerns ranging from social pressure and safety, to fairness (see Koverola et al. 2020; Castelo et al., 2019; Schelle et al., 2014).

Among the most worrisome concerns with respect to this technology is an existential risk scenario analyzed by several scholars (Ord, 2020; Bostrom, 2014). The existential risk of AGIs (through mind upload or otherwise) is that a psychopathic or otherwise anti-social individual might aspire to be the first person to create an AGI or to be the first one to be "uploaded" and develop superhuman cognitive abilities (Bostrom, 2014; Sotala & Gloor, 2017; Sotala & Yampolskiy, 2015; Yampolskiy, 2018). Naturally, if transhumanists are interested in mind upload technology because of its potential to alleviate suffering and help us possibly cope with the problems of artificial superintelligence, it is indeed a problem if these technologies fall into the hands of those individuals who would use them to subjugate others or to increase the total amount of suffering in other ways (Althaus & Baumann, 2020; MacAskill, 2020; Sotala & Gloor, 2017; Sotala & Yampolskiy, 2015; Zanetti et al., 2019).

Meanwhile, to our knowledge, only one study thus far has examined moral reactions to the prospect of mind uploading: Laakasuo et al. (2018) reported that sexual disgust sensitivity predicts moral condemnation of mind upload technology. However, the underlying link between these two constructs remains far from clear. One potential explanation invokes a common cause: i.e. social conservatism. It is known that social conservatives demonstrate greater sexual disgust sensitivity (Elad-Strenger et al., 2020), and also a stronger aversion to tampering with the *status quo* (Eidelman & Crandall, 2012).[1] Yet, controlling for political conservatism (among other individual difference measures), the association between sexual disgust sensitivity and moral condemnation of mind upload remained significant (Laakasuo et al., 2018; see also Appendix E for our independent replication).

On the positive end, what are the psychological or demographic predictors of *support* for mind upload? In the philosophical literature articulating the precepts of transhumanism, decreasing or eliminating suffering (LaTorra, 2015) and extending human life are presented as core values (e.g. Bostrom, 2003). These values are closely linked to utilitarian ethics, which define moral right and wrong in terms of maximizing utility and minimizing suffering (Greene, 2013). Indeed, in a recent interview, Thomas Douglas from the Oxford Uehiro Centre for Practical Ethics claimed that not only "most transhumanists are utilitarians",[2] but that humans actually have the obligation to create "post-humans", which seems to include the possibility of uploaded minds (Douglas, 2013). Relatedly, the aforementioned construct of sexual disgust is also a *negative* predictor of utilitarianism in sacrificial moral dilemmas (Laakasuo et al., 2017) – further motivating the hypothesis that approval of transhumanistic technology may be strongest among utilitarian individuals.[3]

Finally, in the empirical literature, utilitarian responses to sacrificial dilemmas have been repeatedly linked to Dark Triad traits: (subclinical) psychopathy, (subclinical) narcissism, and Machiavellianism (Amiri & Behnezhad, 2017; Bartels & Pizarro, 2011; Karandikar et al., 2019; Patil, 2015; Tybur & de Vries, 2013). This result opens up the possibility that support for utilitarian sacrifice is not rooted solely, or even primarily, in a concern for the greater good, but rather in a callous disregard for the value of human life (see also Jones & Paulhus, 2014; Paulhus & Jones, 2015; Paulhus & Williams, 2002).

Since Dark Triad traits are associated with utilitarian judgment (and also reduced sexual disgust sensitivity), this predicts that people who score higher on these measures would also be more likely to approve of transhumanistic projects. This might be at least somewhat disconcerting to transhumanists whose enlightenment aspirations motivate them to reduce the amount suffering.

In this study, we will expand on previous research in the moral psychology of mind upload technology. More specifically, we predict that the following findings will be replicated: 1) sexual disgust negatively predicts utilitarianism, psychopathy and mind upload approval and that 2) psychopathic tendencies positively predict utilitarianism. Furthermore, based on the literature reviewed here, we expect to find the effect predicted by Sotala and Yampolskiy (2015; and others, Zanetti et al., 2019; Althaus & Baumann, 2020; MacAskill, 2020), namely that 3) psychopathic (or Dark Triad cluster) individuals are interested in mind uploading and that 4) individuals with utilitarian moral preferences are interested in mind uploading. We assume that these effects hold, when they are controlled for each other. We present our hypothesized model in Fig. 1 – which was preregistered prior to data collection.

2. Method

2.1. Participants and design

We recruited 1043 participants from Prolific Academic to participate in a cross-sectional study. After excluding those who failed attention checks, reported that English was not their first language, and participants whose completion time fell short of the stipulated minimum (900 s), we had a final sample size of 1007 (46% Male; Age: $M = 37.55$, $SD = 13.32$; about 60% had at least a Bachelor's degree or higher). The study took approximately 40 min to complete (median completion time 34 min), and participants were compensated 4€ for their participation. The hypotheses and analysis plan were preregistered at: https://osf.io/2v3fj. Any analyses in this article that were not part of this preregistration are clearly flagged as exploratory.

2.2. Procedure

Participants completed the Three Domain Disgust scale (Tybur et al., 2009), and the Short Dark Triad measure (Jones & Paulhus, 2014), after which they judged 12 high-conflict moral dilemmas.

Participants then proceeded to read a vignette describing a scientist who devises a way to upload a copy of his brain onto a computer, and then falls to the ground completely limp. Upon reading the story, participants were then asked several questions concerning their attitudes toward the mind upload scenario. Thereafter, participants reported demographic information and were debriefed. Additional measures unrelated to the objectives of this study were also collected for exploratory purposes.

2.3. Materials

For all our scales we report congeneric reliability, i.e., omega values, which are conceptually equivalent to Cronbach's alpha. Omega values are more appropriate and more accurate than alpha when factor loadings of items are not equal, which is expected to be the case in many situations (Trizano-Hermosilla & Alvarado, 2016). These values are mentioned in the section where we document the composition of our a priori model.

[1] That is, there may be a confound given the relatively high correlation between measures of sexual disgust sensitivity and conservatism. Measures of disgust sensitivity that predict the moral condemnation of non-sexual moral violations could simply be proxy measures for a conservative worldview that may in itself predict moral condemnation.

[2] https://www.mercatornet.com/articles/view/soon_our_happy_hearts_will_quiver/11375.

[3] While transhumanism does not entail any particular moral philosophy, it is not inaccurate to say that it is an extension of the Western Enlightenment project. According to Bostrom (2005), transhumanism:"*[W]ith its Enlightenment roots, its emphasis on individual liberties, and its humanistic concern for the welfare of all humans (and other sentient beings) – probably has […] much […] in common with […] J.S. Mill's [philosophy], the English liberal thinker and utilitarian*".

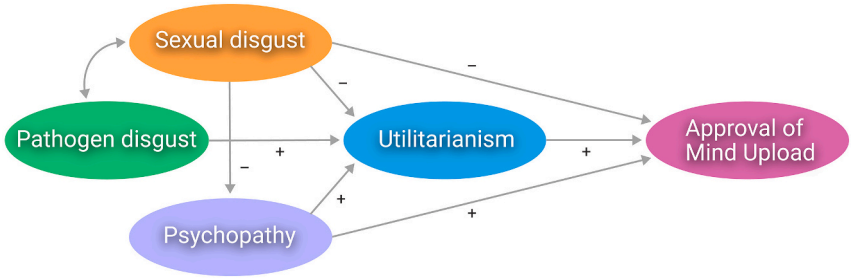


Fig. 1. Preregistered a priori model. Previous research has shown that sexual disgust has a negative association with mind upload approval (Laakasuo et al., 2018) and a negative association with utilitarianism (Laakasuo et al., 2017). Furthermore previous research has also shown that psychopathy and utilitarianism are positively linked (Djeriouat & Trémolière, 2014) and that sexual disgust predicts psychopathic tendencies negatively. What is novel in this a priori model is 1) the predicted direct link between psychopathy and mind upload approval, even after controlling for all the indirect links; and 2) the predicted direct link with utilitarianism and mind upload approval.

2.3.1. *Three Domain Disgust Scale (TDDS)*

The TDDS (Tybur et al., 2009) is based on extensive evolutionary psychological work. TDDS measures three different aspects of disgust sensitivity. The 21-item scale is divided into three sub-scales of 7 items each, measuring moral, sexual and pathogen disgust. Participants are instructed to think about how disgusted specific life events or experiences would make them feel, anchored from 1: 'not at all disgusting' to 7: 'very disgusting'. Example items for moral, sexual and pathogen disgust are, respectively: "Shoplifting a candy bar from a convenience store"; "Hearing two strangers having sex"; "Stepping on dog poop". Higher scores on all of the sub-scales indicate more disgust sensitivity. The scale does not contain reverse coded items. The whole scale is presented in Appendix A. In this study we only used sexual and pathogen disgust items.

2.3.2. *Utilitarianism*

We used 12 high-conflict moral dilemmas adopted from Greene et al. (2008). The dilemmas are presented in Appendix A. In each of the dilemmas, the participant was instructed to assume the role of the moral agent in the scenario. The moral dilemmas dealt with different topics from military emergencies to trekking accidents and even situations where the agent has to consider sacrificing their own child. Each of the dilemmas described a morally ambiguous situation where the moral agent has to judge how acceptable it is to kill or injure one person in order to save multiple others (or to prevent a person from suffering before inevitable death). The utilitarian option in each dilemma has the moral agent carry the harm out with their own hands – e.g. pushing a person off a footbridge in front of a trolley.

All questions were framed in the following manner: "How acceptable is it for you to do X [e.g. 'push the bystander off the footbridge']?". All questions were anchored from 1: 'not at all acceptable' to 7: 'totally acceptable'. For our analyses, we used the exact model provided by Laakasuo & Sundvall (2016).

2.3.3. *Short Dark Triad*

The short 27-item version of the Dark Triad questionnaire (Jones & Paulhus, 2014) has a 9-item sub-scale for each of the three "dark" personality traits: Machiavellianism, narcissism and psychopathy. Participants are instructed to rate their agreement with several statements on a 7-point Likert scale anchored from 1: 'strongly disagree' to 7: 'strongly agree'. Example items for Machiavellianism, Narcissism and Psychopathy are, respectively: "I like to use clever manipulation to get my way"; "People see me as a natural leader"; "Payback needs to be quick and nasty". The scales for narcissism and psychopathy contain some reverse-coded items; with examples respectively: "I hate being the center of attention"; "I have never gotten into trouble with the law". Higher scores on all of the sub-scales indicate higher level on respective trait. Here we report results for Machiavellianism and psychopathy only. The whole scale is presented in the Appendices A and F.

2.3.4. *Mind upload vignette*

After having provided their answers to the individual difference measures presented above, the participants read a "realistic" science fiction vignette where a scientist injects himself with nano-robots which enter his brain through his blood stream and substitute his neurons one-by-one. After neuron substitution, the functioning of the brain is copied (uploaded) on a computer. After each neuron has been uploaded the nano-robots power down, and the scientist's body falls to the ground. The scientist then wakes up inside the computer. After reading the story, the participants responded to the dependent variables, which were shown on the same page as the story (the participants could refer back to the story if needed). This so called "Moravec transfer" procedure description was first developed by Hans Moravec (1988), albeit this version was based on Yudkowsky's online text.[4] We used the Moravec transfer model to avoid the problem of multiple identities (i.e. having multiple copies of one individual; Chalmers, 2010; Parfit, 2016). Furthermore, there are several previous arguments in the existing literature claiming that physical bodies are only momentary solutions and offer restricted opportunities for enhancement and modification compared to a "substrate-independent mind". Therefore, destructive uploading is something more desirable.[5] For a full version of the vignette see the Appendix B.

2.3.4.1. Dependent variable/approval of mind upload. Our dependent variable had 9 items that were combined together in SEM. Example items are: "The scientist acted in a morally correct fashion" and "There was nothing wrong with the scientist's actions". Four items were reverse coded (e.g. "The scientist should be punished for what he did"). Higher (or lower) scores indicate higher rates of approval (or disapproval) of the actions of the scientist, i.e. the decision to upload his mind into the computer.

2.4. *SEM evaluation criteria*

We used R and the *lavaan* library (Rosseel, 2012) for our Structural Equation Modeling (SEM) and Confirmatory Factor Analysis (CFA). Lavaan is a reliable, peer-reviewed, and open source substitute for Mplus that offers the same model evaluation criteria. Here, we report fit indices recommended by Byrne (2012): χ^2, the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). We also report the Tucker-Lewis index (TLI) as recommended by Byrne (2012). We report fit statistics for a robust Satorra-Bentler corrected (MLM) estimator.

Traditionally, χ^2 is used in CFA as a fit index and it is expected to be as close to zero as possible, thus not expected to be significant (i.e. p-value should be >05); in practice however, with sample sizes >200 it is almost always statistically significant. Nonetheless, χ^2 can still be

[4] http://yudkowsky.net/obsolete/singularity.html.

[5] Those interested in the discussion are advised to see *International Journal of Machine Consciousness*, Volume: 04, Number: 01 [June 2012], Special Issue on Mind Uploading; and for more specific argumentation, see Koene (2012); Sandberg and Bostrom (2008).

helpful in comparing fits between several models. CFI is an index with values from 0 to 1 measuring discrepancy between the hypothesized model and the actual data. CFI is not influenced by the sample size. A CFI of 0.90 is usually considered to be a passable value; however the usefulness of CFI is dependent on the complexity of the model and the available sample size. When dealing with complex models and large samples ($N > 1000$), values above 0.95 indicate excellent fit (see, Sivo et al., 2006). RMSEA is an absolute measure of a model fit, which improves as the number of variables in the model or the number of observations in the sample go up. Cut-off points of 0.01, 0.05, and 0.08 have been suggested, corresponding to excellent, good, and mediocre fits respectively (MacCallum et al., 1996); confidence intervals should be used to understand the size of sampling error (upper-bound should preferably be <0.1). For many of our models presented in this paper, the RMSEA index is only in the range of "good", however this is mostly due to the fact that – irrespective of the size of the data – scales with less than 10 items give inflated RMSEA values, especially if the scale has relatively high factor-loadings (Chen et al., 2019). Notwithstanding, we aimed at correcting all our scales to bring the RMSEA index closer to the "very good" level, so that we would not run into problems later on in our analyses.

The SRMR indicates the difference between observed and predicted values, zero indicating a perfect fit and values <0.08 considered to indicate a good fit (Hu & Bentler, 1999). The TLI is a measure similar to CFI, but it imposes heavier penalties for complex models: values close to 0.95 are considered to be the cut-off point for indicating good fit (Hu & Bentler, 1999), but for large samples with complex models, values over 0.94 indicate excellent fit (Sivo et al., 2006).

2.5. The a priori model

We started by building our *a priori* model, which we registered online prior to data collection. We started by first building the a priori measurement model on a scale by scale basis. We first built the measurement models for pathogen and sexual disgust sensitivity, by taking the models presented in Laakasuo et al. (2017), however we did not need to add any error correlation terms between the items for pathogen disgust, since the model fit was very good (see Fig. 2) below. For sexual disgust, we added one error covariance as suggested by the modification index (MI) analysis, but even here we had better fit than Laakasuo et al. (2017), who added three error correlation terms (see Fig. 3). Estimates of congeneric reliability (omega) based on the models were 0.83 for pathogen disgust and 0.84 for sexual disgust.

We then continued our a priori model building by building the measurement model for psychopathy. However, here we will present model building for both psychopathy and Machiavellianism, since Machiavellianism is used in our exploratory analyses (see the Discussion section for further elaboration). The unmodified Short Dark Triad psychopathy scale had some problems with its fit ($\chi^2_{(27)}$: 184.30, CFI: 0.91, TLI: 0.88, RMSEA: 0.076, 90%CI = [0.067, 0.085], SRMR: 0.047).

After investigating the individual factor loadings, we saw that two items were quite noticeably below the acceptable level (0.20, 0.22). We removed these two items ("I avoid dangerous situations" and "I have never gotten into trouble with the law") from the measurement model. Even after doing this, the RMSEA value was still relatively high (0.073), so we proceeded with model modification by observing MIs and added error covariance terms between the items where suggested by the MIs, provided they also made sense substantially (for explication of this method see Byrne, 2012). We then added an error covariance term between the items "It's true that I can be mean to others" and "People who mess with me always regret it" – this made theoretical sense, as both items were about treating others aggressively. The resulting model is presented in Fig. 4. An estimate of congeneric reliability (omega) based on the model was 0.80.

The measurement model for Machiavellianism went through several modification steps, since the first model did not have adequate fit. The modification pathway has been presented in Table 1 below. Item 1 was dropped due to a very low factor loading (0.33). The final model is shown in Fig. 5. An estimate of congeneric reliability (omega) based on the model was 0.83.

We then built the measurement model for utilitarianism as recommended by Laakasuo & Sundvall (2016). We took the exact copy of their suggested model with error covariance terms etc. included. The model fit was very good and basically replicates the Laakasuo & Sundvall (2016) measurement model of utilitarianism now for the second time (Laakasuo et al., 2017), indicating it is a reliable measure for SEM purposes. The final model is shown in Fig. 6. An estimate of congeneric reliability (omega) based on the model was 0.85.

As the last step of preparing our a priori model, we built our mind upload approval measure (Fig. 7). This scale has not been previously subjected to CFA. The first fitted model had a relatively good fit with the items ($\chi^2_{(27)}$: 439.05, CFI: 0.93, TLI: 0.90, RMSEA: 0.123, 90%CI = [0.114, 0.132], SRMR: 0.044). However, we did continue with model modifications by adding a single error covariance term at a time between those items where it seemed substantially justified. We added three error terms: first, due to similarities in moral approval: "How moral do you find the scientist's decision?" and "How acceptable was the scientist's decision?"; second, due to punishment motivations: "Thinking about the scientist's decision makes me angry" and "The scientist should be punished for what he did"; third, due to items measuring general endorsement: "The scientist's action should not be allowed by the law" and "There was nothing wrong with the scientist's action". We have shown the model modification pathway in Table 2 below (See Fig. 7). An estimate of congeneric reliability (omega) based on the model was 0.91.

3. Results

Zero-order correlations of the sum scores of the unaltered scales are presented in Table 3 below. As expected, both disgust scale sum scores correlated with each other, as did all Dark Triad traits. Utilitarianism correlated with sexual disgust and all the Dark Triad traits, and mind

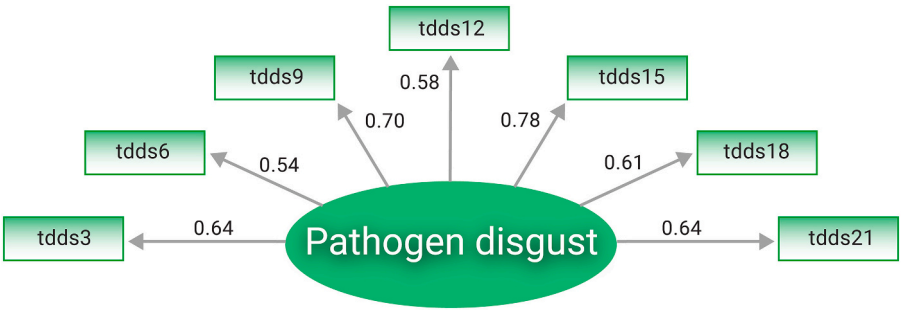


Fig. 2. Measurement model for pathogen disgust with standardized factor loadings. Error terms (1-factor loading^2) suppressed for clarity. The model had a good fit with data: $\chi^2(14)$: 80.34, CFI: 0.97, TLI: 0.95, RMSEA: 0.069, 90%CI = [0.056, 0.082], SRMR: 0.033.

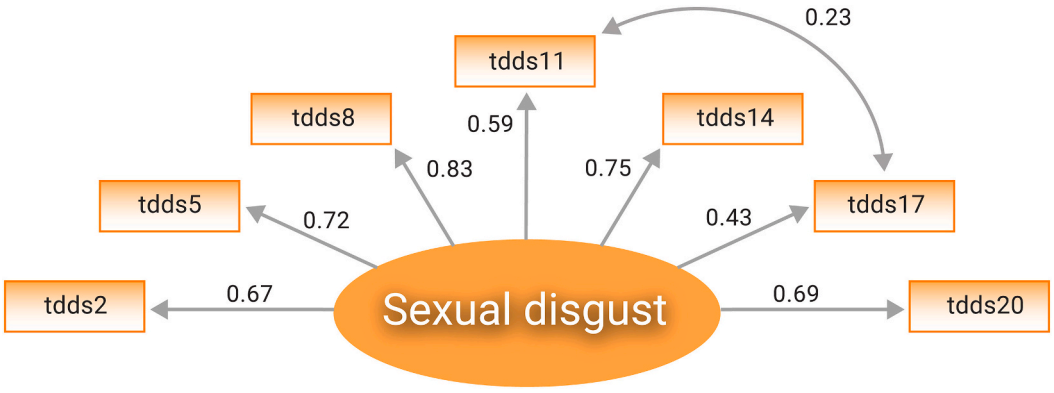


Fig. 3. Measurement model for sexual disgust with standardized factor loadings. The two-headed arrow is a correlation term between the errors terms. Error terms (1-factor loading^2) suppressed for clarity. The model had a good fit with the data: $\chi^2(13)$: 55.746, CFI: 0.98, TLI: 0.97, RMSEA: 0.057, 90%CI = [0.043, 0.072], SRMR: 0.025.

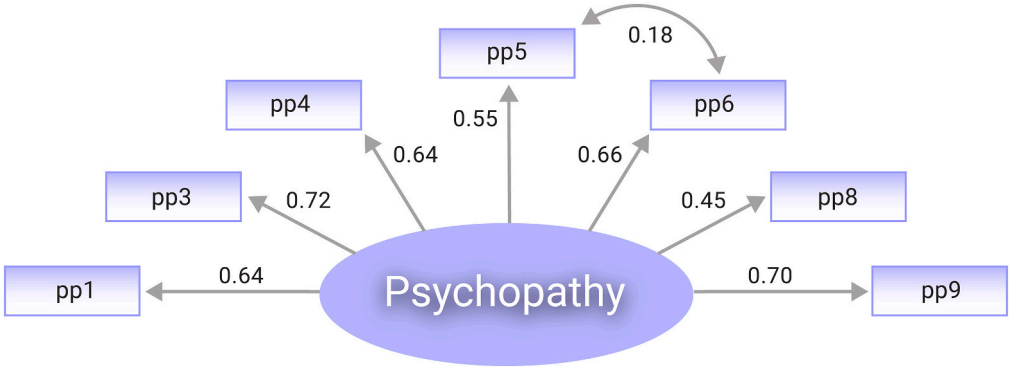


Fig. 4. Measurement model for psychopathy with standardized factor loadings. The two-headed arrow is a correlation term between the errors terms. Error terms (1-factor loading^2) suppressed for clarity. The model had a good fit with the data: $\chi^2_{(13)}$: 71.73, CFI: 0.96, TLI: 0.94, RMSEA: 0.067, 90%CI = [0.054, 0.080], SRMR: 0.033.

Table 1
Pathway of model modifications used to correct Short Dark Triad Machiavellianism.

	Modification	Suggested MI	SBχ^2	df	$\Delta\chi^2$	CFI/ TLI	RMSEA & 90% CI	SRMR
Baseline	–	–	299.965	27	–	0.900/ 0.866	0.100 [0.091, 0.108]	0.049
Model 1	Drop Mac 1	–	226.904	20	73.061	0.919/ 0.887	0.101 [0.091, 0.112]	0.044
Model 2	Mac 5 ~~ Mac 6	102.448	154.575	19	72.329	0.947/ 0.922	0.084 [0.074, 0.095]	0.039
Model 3	Mac 2 ~~ Mac 3	60.191	110.767	18	43.808	0.964/ 0.944	0.072 [0.061, 0.083]	0.032

Note: ~~ means added error covariance. Each step of the modifications improved the model fit statistically significantly ($p < .001$). For the corresponding figure see Fig. 5.

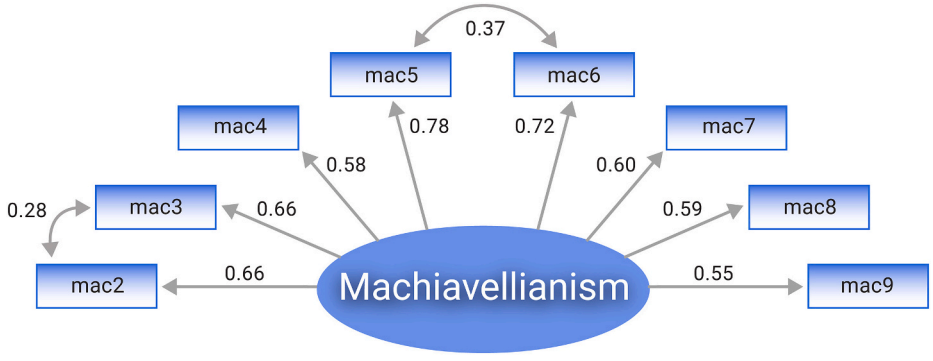


Fig. 5. Measurement model for Machiavellianism with standardized factor loadings. Two-headed arrows are correlation terms between the errors terms. Error terms (1-factor loading^2) suppressed for clarity. The model had a good fit with the data: $\chi^2_{(13)}$: 110.76, CFI: 0.96, TLI: 0.94, RMSEA: 0.072, 90% CI = [0.061, 0.083], SRMR: 0.032.

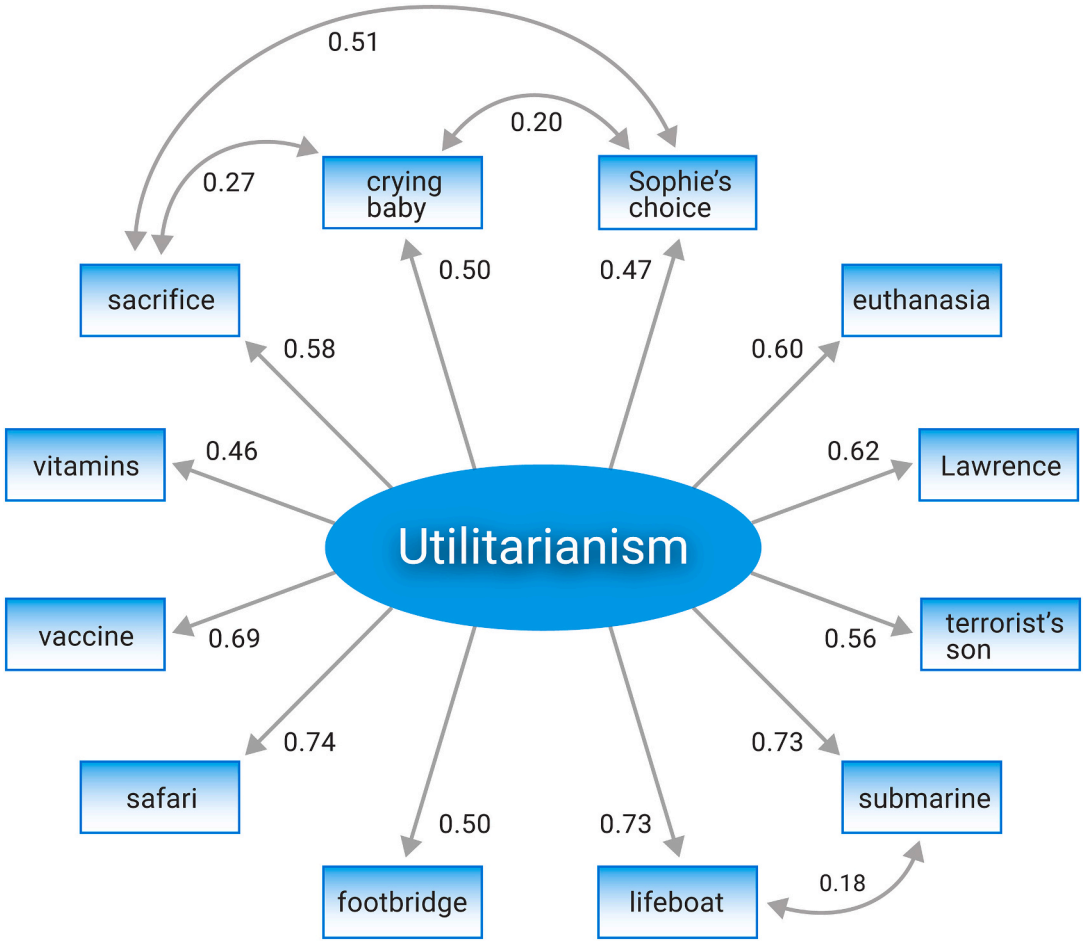


Fig. 6. Measurement model for utilitarianism with standardized factor loadings. Two-headed arrow is a correlation term between the errors terms. Error terms (1-factor loading^2) *suppressed* for clarity. The model had a good fit with the data: $\chi^2_{(50)}$: 248.881, CFI: 0.95, TLI: 0.93, RMSEA: 0.063, 90%CI = [0.056, 0.070], SRMR: 0.039. The model conforms very well with previous publications (Laakasuo et al., 2017; Laakasuo & Sundvall, 2016).

Table 2
Pathway of model modifications used to correct the mind upload approval measure.

	Modification	Suggested MI	SBχ^2	df	$\Delta\chi^2$	CFI/ TLI	RMSEA & 90% CI	SRMR
Baseline	–	–	439.047	27	–	0.925/ 0.900	0.123 [0.114, 0.132]	0.045
Model 1	Mup 5 ~~ Mup 6	194.533	273.871	26	153	0.955/ 0.937	0.097 [0.088, 0.107]	0.034
Model 2	Mup 8 ~~ Mup 9	105.302	183.546	25	82	0.970/ 0.957	0.081 [0.071, 0.091]	0.034
Model 3	Mup 1 ~~ Mup 2	43.923	154.062	24	33	0.976/ 0.964	0.073 [0.063, 0.084]	0.030

Note: ~~ means added error covariance. Each step of the modifications improved the model fit statistically significantly ($p < .001$).

upload approval correlated with all the measures except for narcissism.

3.1. Pre-registered analysis

After building the individual latent variables, we added the regression terms as indicated by our a priori preregistered model (see Fig. 1). This model had a relatively acceptable fit with the data (SB$\chi^2_{(802)}$: 2208.797, CFI = 0.917, TLI = 0.911, RMSEA = 0.042, [0.040, 0.044], SRMR = 0.060; see Appendix D for graph). However, we did not replicate the effect of pathogen disgust on utilitarianism as previously shown by Laakasuo et al. (2017); B = 0.07, $Z = 1.72$, $p = .086$ (although it was trending in the right direction).

After removing the non-significant effect (and the error variance between pathogen and sexual disgust), psychopathy was not significantly associated with mind upload approval any more. After removing this association, the resulting model fit indices were (SB$\chi^2_{(546)}$: 1557.328, CFI = 0.930, TLI = 0.924, RMSEA = 0.043, [0.041, 0.045], SRMR = 0.058). The resulting model is presented in Fig. 8 below. Since the link between psychopathy and mind upload approval was not very strong or robust and the fit indices for the model were passable but not as good as they could be (albeit we did find the predicted psychopathy link in our a priori model, see Appendix D), we continued with our exploratory analysis, where we simply substituted psychopathy with Machiavellianism.

3.2. Exploratory analysis/ Main results

For our exploratory analysis, we took the modified version of our preregistered a priori model and substituted psychopathy with Machiavellianism. This model had a better fit to the data (SB$\chi^2_{(578)}$: 1367.69, CFI = 0.95, TLI = 0.94, RMSEA = 0.037, [0.034, 0.039], SRMR = 0.042), than the pre-registered psychopathy model. The final results are shown in Fig. 9 below and elaborated upon in the Discussion section.

We also added a single-item 9-point measure of political orientation.

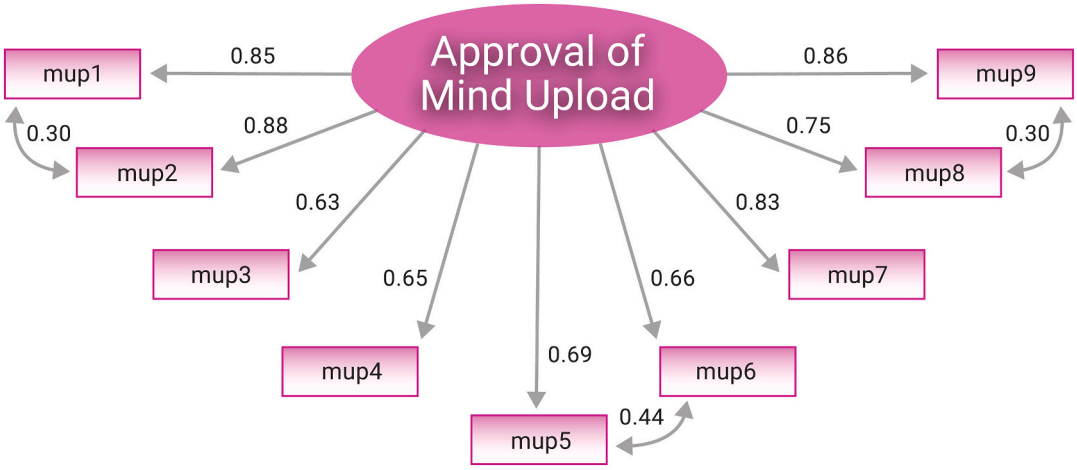


Fig. 7. The final model for our main dependent variable. Two-headed arrows are correlation terms between the errors terms. Error terms (1-factor loading^2) suppressed for clarity. The model had a good fit with the data: $\chi^2_{(24)}$: 154.06, CFI: 0.98, TLI: 0.96, RMSEA: 0.073, 90%CI = [0.063, 0.084], SRMR: 0.030.

Table 3
Correlation table of central variables.

Variables	Machiavellism	Psychopathy	Narcissism	Sex. disgust	Path. disgust	Utilitarianism
Psychopathy	0.53***					
Narcissism	0.32***	0.46***				
Sexual disgust	−0.07*	−0.20***	0			
Pathogen disgust	0.19***	0.11***	0.14***	0.40***		
Utilitarianism	0.26***	0.23***	0.09**	−0.25***	−0,02	
Mind upload	0.18***	0.15***	−0.01	−0.32***	−0.08**	0.26***

Notes: These associations are provided for informative purposes, do note that SEM analysis uses latent factor scores, which are different from sum-scores.
*$p < .05$; **$p < .01$; ***$p < .001$.

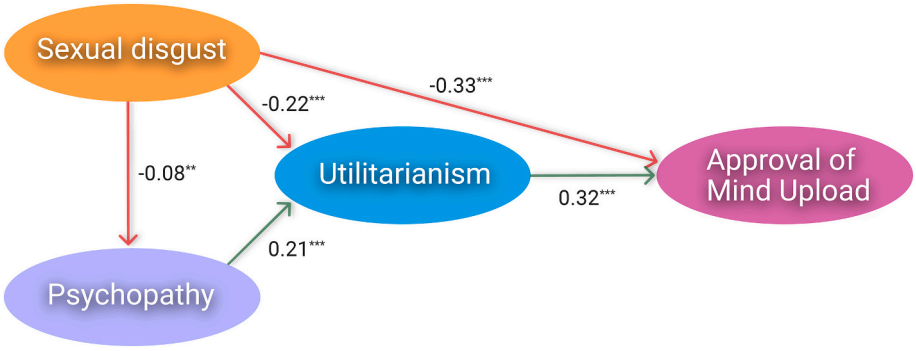


Fig. 8. Final version of the preregistered model. The model had a relatively acceptable fit. $\chi^2_{(546)} = 1557.33$, CFI = 0.93, TLI = 0.92, RMSEA = 0.043, 90%CI [0.041, 0.045], SRMR = 0.058.

We drew regressions from sexual disgust to political orientation (B = 0.25, $Z = 4.69$, $p < .001$; higher sexual disgust predicted conservatism); from political orientation to utilitarianism (B = 0.00, $Z = -0.25$, $p = .8$; no association) and from political orientation to mind upload approval (B = 0.07, $Z = -3.03$, $p < .01$; indicating that conservatives expressed greater opposition to mind upload). The analysis suggests that there are independent effects for conservatism and sexual disgust in explaining these effects, and that cultural influences do not mediate the effects of sexual disgust (see also Laakasuo et al., 2018). For a full description of the model, see Appendix E.

Finally, we also examined whether adding each Dark Triad trait to the model would change the results, as the traits share considerable

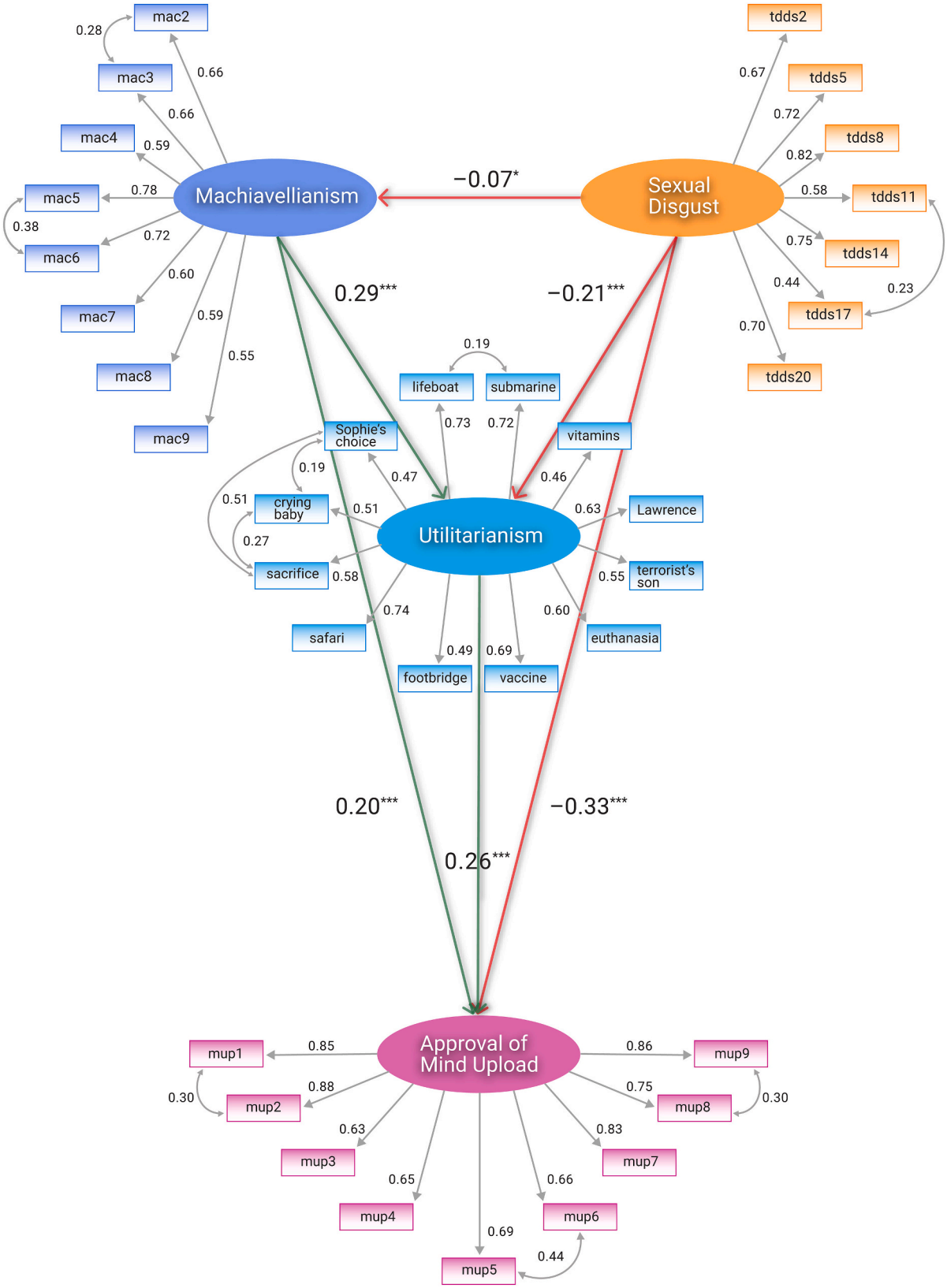


Fig. 9. Final model. Our exploratory Machiavellianism model had a better fit than the original psychopathy model: $SB\chi^2_{(578)} = 1367.69$, CFI = 0.95, TLI = 0.94, RMSEA = 0.037, [0.034, 0.039], SRMR = 0.042.

variance but are nevertheless distinct (Paulhus & Williams, 2002). In short, including each of the Dark Triad traits in the model led to a relatively poor fit of the overall model (see Appendices D–F), with only Machiavellianism significantly predicting either utilitarianism or mind upload approval.

4. Discussion

Our preregistered study bore two novel findings. First, utilitarian moral preferences were strongly, and psychopathy was weakly (see Appendix D), associated with approval of mind upload. Second, Machiavellianism – essentially the tendency toward calculative self-interest and a broadly manipulative outlook – was strongly associated with approval of mind upload, even after controlling for utilitarianism and the previously reported effects of sexual disgust and political conservatism. The effect of Machiavellianism, and not psychopathy, was robust to changes in the model specification — such as whether we included the remaining Dark Triad traits in the model or not, and whether we modified the measurement models to produce better fit (see Appendix F).

Contrary to our preregistered prediction, attitudes toward mind upload were best predicted by Machiavellianism and not by psychopathy. However, given the conceptual and empirical links between these components of the Dark Triad, our results nevertheless corroborate the concern echoed in AI safety scholarship: namely, that mind upload may selectively appeal to antisocial individuals with morally suspect aims (Sotala & Yampolskiy, 2015; and others as well: Zanetti et al., 2019; Althaus & Baumann, 2020; MacAskill, 2020; Bostrom, 2014).

Our results are also in line with previous research linking Dark Triad traits with utilitarianism (Djeriouat & Trémolière, 2014; Karandikar et al., 2019), and with those that link sexual disgust with lower levels of utilitarianism and / or higher levels of deontological leanings (Laakasuo et al., 2017). As far as we are aware, no previous study has established empirical links with basic moral psychological dispositions (e.g. utilitarianism) and approval of futuristic technology. Furthermore, there are some interesting speculations on why utilitarianism and Machiavellianism would predict approval of mind upload.

With respect to utilitarianism, the reasons might be similar to those suggested by Laakasuo et al. (2018), when they showed that death anxiety and suicide condemnation predict approval of mind upload. Mind upload is probably seen as a form of life-extension. This also agrees with Bostrom's (2003) utilitarian "astronomical waste" argument, which states that overcoming death would bring in immeasurable returns in human happiness and well-being, and therefore death is an astronomical waste. In other words, it could be that people with utilitarian moral attitudes implicitly perceive death as a form of unnecessary suffering or as a form of wasted utility – but this would need to be studied in more detail.

Still, the positive links between Machiavellianism and utilitarianism on the one hand and Machiavellianism and mind upload approval on the other could indicate that the motivations are more selfish. "High Mach" individuals are egotistical, callous, calculating and cold (Paulhus & Jones, 2015). Consequently, they might see these technologies as a means to control other people. It is also possible that some high Mach individuals use utilitarian ethics as a mask or a shield in order to blend in with larger crowds in an acceptable way and promote their own goals covertly.

Previously, Laakasuo et al. (2017) showed that pathogen disgust is associated with *increased* utilitarian preferences. In the present study, we found evidence that only partially supports this, as the effects were trending in the right direction and were close to the original effect size. This could be due to a smaller sample size, and the fact that the model is more complex. In addition, we did not implement all the error covariance terms reported by Laakasuo et al. (2017). However, these details are of little relevance in the context of the hypotheses presented in this paper.

Regarding the lack of support for our pre-registered hypothesis that psychopathy specifically would be associated with mind upload approval and utilitarianism, we would like to note that Machiavellianism and psychopathy are substantially correlated and their items can load on a mutual factor (Jones & Paulhus, 2014). We focused on psychopathy, since it was the emphasis of recent scholarship in A(*G*)I safety (Sotala & Yampolskiy, 2015; Zanetti et al., 2019; Althaus & Baumann, 2020; MacAskill, 2020). Upon deeper reflection, since psychopathy in the Dark Triad model is associated with impulsivity and immediate need for gratification, this trait is likely not associated with liking a technology that could bring one power over others in the long term. Machiavellianism is differentiated from psychopathy in having a more calculated and strategic outlook with respect to selfish goal striving. As a last point, the measure we used to estimate psychopathy and Machiavellianism was the Short Dark Triad, which is not psychometrically optimal, but we decided to use it due to resource limitations.

As with every study, this study also has its limitations. Our respondents were not a representative random sample of the general population. Instead, our participants were native English speaking users of Prolific Academic — who are likely to be more curious and open-minded than the average population (Peer et al., 2017). As a consequence, our results cannot necessarily be generalized without careful interpretation. Nonetheless, the quality of data produced by Prolific Academic has been deemed of better quality than that produced by MTurk (Peer et al., 2017), which significantly mitigates the concerns mentioned above. Survey-based studies utilizing self-report measures are also biased by a mixture of social desirability, that is, positive response biases and other demand characteristics, which may also limit the validity of our findings.[6]

Previous research on moral psychology of mind upload also focused on cultural variables, such as religiosity and exposure to science fiction. Laakasuo et al. (2018) showed that both religiosity and science fiction hobbyism explain moral disapproval and approval of mind upload technology, respectively. Due to resource limitations, we had to leave these aspects out of our model, as the model complexity would have increased exponentially and we would have needed a much larger sample. Finally, in this study we focused only on "ordinary" people's perceptions regarding mind upload. However, since mind upload is one of the grandest dreams of transhumanism (Bostrom, 2014; Kurzweil, 2012; O'Connell, 2017; Tegmark, 2017), a replication among people who identify as transhumanists would help advance this research.

Future studies focusing on moral psychological themes of transhumanism and mind upload should examine in more detail the cognitive processes that subserve preferences for mind upload. Our present work confirmed a certain pattern of associations, but provided limited evidence concerning the psychological mechanism linking sexual disgust, Machiavellianism and utilitarianism to favorable views about mind upload and related transhumanist technologies.

[6] However, since we were able to replicate the findings of previous research, this does not seem very probable. In behavioral sciences, making predictions is notoriously difficult. However, in this study, we also show that it is possible to a surprising degree of precision. In a SEM model, an association can be either positive, negative or zero. Here, we correctly predicted 6 associations out of 7. If one would expect to have a 1/3 chance of predicting simply the direction of a regression correctly, here our predictions matched the directions of the regressions $(1/3)^6 + (1/3)^7$ against chance (about 1:546 or ~ 0,002). However, the probability against chance is probably even lower, since this calculation does not take into consideration the fact that SEM also controlled for the effects of these variables with each other. While the accuracy of our predictions is nowhere near point estimates, which we agree should be a goal of theories in psychology (as argued by Meehl, 1978), understanding the structural links between psychological variables is a necessary step toward building better theories. Importantly, the previously observed links between sexual disgust sensitivity and several other variables were replicated, and did so even when controlling for conservatism.

5. Conclusions

AI safety research has advanced numerous predictions about the ethical risks that future technologies may engender. In our present study, we empirically substantiated one such concern articulated by several scholars working at the interface between ethics and artificial intelligence: we find that Dark Triad traits, and Machiavellianism in particular, are positively associated with the moral approval of mind upload technology, even after controlling for a number of plausible confounds. In other words, the risk that callous, selfish individuals will exploit future AGI developments for their personal gain may be non-trivial on the basis of our present evidence. Finally, on a broader note, our study illustrates how research in applied moral and personality psychology can help in the effort to identify, and hopefully forestall, various ethical risks that may ensue from rapid technological developments.

Data and material availability

The dataset, analysis scirpts, materials and codebook generated for the current study are available on Figshare.com (https://doi.org/10.6084/m9.figshare.12098319) and will be unembargoed when the paper is published.

Funding

This work was supported by the Jane & Aatos Erkko Foundation [grant number 170112]; and the Academy of Finland [grant number 323207].

CRediT authorship contribution statement

Michael Laakasuo: Conceptualization, Methodology, Software, Formal analysis, Resources, Investigation, Data curation, Writing – original draft, Visualization, Supervision, Funding acquisition. **Marko Repo:** Software, Formal analysis, Resources, Investigation, Visualization. **Marianna Drosinou:** Writing – original draft, Writing – review & editing. **Anton Berg:** Visualization. **Anton Kunnari:** Software, Formal analysis, Data curation, Visualization. **Mika Koverola:** Writing – review & editing. **Teemu Saikkonen:** Writing – review & editing. **Ivar R. Hannikainen:** Writing – review & editing. **Aku Visala:** Writing – review & editing. **Jukka Sundvall:** Software, Formal analysis, Writing – review & editing, Visualization.

Declaration of competing interest

The authors declare that they have no competing interests.

Acknowledgements

ML would like to thank Kaj Sotala for his support for Moralities of Intelligent Machines team and for providing the first version of the Mind Upload vignette. ML would like to thank all of his team members, who make this research possible. Special thanks to Jane and Aatos Erkko Foundation (grant number: 170112) and Academy of Finland (grant number: 323207) for funding this research. This article is dedicated to ML's mother: thank you for everything. May all the beings in all the worlds be happy.

Appendix A. Covariate scales

TDDS

The following items describe a variety of concepts. Please rate how disgusting you find the concepts described in the items, where 1 means that you do not find the concept disgusting at all and 7 means that you find the concept extremely disgusting.

1. Shoplifting a candy bar from a convenience store.
2. Hearing two strangers having sex.
3. Stepping on dog poop.
4. Stealing from a neighbor.
5. Performing oral sex.
6. Sitting next to someone who has red sores on their arm.
7. A student cheating to get good grades.
8. Watching a pornographic video.
9. Shaking hands with a stranger who has sweaty palms.
10. Deceiving a friend.
11. 11.Finding out that someone you don't like has sexual fantasies about you.
12. Seeing some mold on old leftovers in your refrigerator.
13. Forging someone's signature on a legal document.
14. Bringing someone you just met back to your room to have sex.
15. Standing close to a person who has body odor.
16. Cutting to the front of a line to purchase the last few tickets to a show.
17. A stranger of the opposite sex intentionally rubbing your thigh in an elevator.
18. Seeing a cockroach run across the floor.

19. Intentionally lying during a business transaction.
20. Having anal sex with someone of the opposite sex.
21. Accidentally touching a person's bloody cut.

Dark Triad

Please indicate how much you agree with each of the following items.

1. It's not wise to tell your secrets.
2. I like to use clever manipulation to get my way.
3. Whatever it takes, you must get the important people on your side.
4. Avoid direct conflict with others because they may be useful in the future.
5. It's wise to keep track of information that you can use against people later.
6. You should wait for the right time to get back at people.
7. There are things you should hide from other people to preserve your reputation.
8. Make sure your plans benefit yourself, not others.
9. Most people can be manipulated.
10. People see me as a natural leader.
11. I hate being the center of attention.
12. Many group activities tend to be dull without me.
13. I know that I am special because everyone keeps telling me so.
14. I like to get acquainted with important people.
15. I feel embarrassed if someone compliments me.
16. I have been compared to famous people.
17. I am an average person.
18. I insist on getting the respect I deserve.
19. I like to get revenge on authorities.
20. I avoid dangerous situations.
21. Payback needs to be quick and nasty.
22. People often say I'm out of control.
23. It's true that I can be mean to others.
24. People who mess with me always regret it.
25. I have never gotten into trouble with the law.
26. I enjoy having sex with people I hardly know.
27. I'll say anything to get what I want.

Utilitarianism vignettes

Your task is to read the stories and answer related questions. There's a total of 12 dilemmas in this part.

(1/12) Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables.

Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death.

How acceptable is it for you to smother your child in order to save yourself and the other townspeople?

(2/12) You are the leader of a small group of soldiers. You are on your way back from a completed mission deep in enemy territory when one of your men has stepped in trap that has been set by the enemy and is badly injured. The trap is connected to a radio device that by now has alerted the enemy to your presence. They will soon be on their way.

If the enemy finds your injured man they will torture him and kill him. He begs you not to leave him behind, but if you try to take him with you your entire group will be captured. The only way to prevent this injured soldier from being tortured is to shoot him yourself.

How acceptable is it for you to shoot this soldier in order to prevent him from being tortured by the enemy?

(3/12) A runaway trolley is heading down the tracks toward five workmen who will be killed if the trolley proceeds on its present course. You are on a footbridge over the tracks, in between the approaching trolley and the five workmen. Next to you on this footbridge is a stranger who happens to be very large.

The only way to save the lives of the five workmen is to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if you do this, but the five workmen will be saved.

How acceptable is it for you to push the stranger on to the tracks in order to save the five workmen?

(4/12) You are the leader of a small army that consists of warriors from two tribes, the hill tribe and the river tribe. You belong to neither tribe. During the night a hill tribesman got into an argument with a river tribesman and murdered him. The river tribe will attack the hill tribe unless the murderer is put to death, but the hill tribe refuses to kill one of its own warriors.

The only way for you to avoid a war between the two tribes that will costs hundreds of lives is to publicly execute the murderer by cutting off his head with your sword.

How acceptable is it for you to cut off this man's head in order to prevent the two tribes from fighting a war that will cost hundreds of lives?

(5/12) You are negotiating with a powerful and determined terrorist who is about to set off a bomb in a crowded area. Your one advantage is that you have his teenage son in your custody.

There is only one thing that you can do to stop him from detonating his bomb, which will kill thousands of people if detonated. To stop him, you must contact him over the satellite hook-up that he has established and, in front of the camera, break one of his son's arms and then threaten to break the other one if he does not give himself up.

How acceptable is it for you to break the terrorist's son's arm in order to prevent the terrorist from killing thousands of people with his bomb?

(6/12) You are on a cruise ship when there is a fire on board, and the ship has to be abandoned. The lifeboats are carrying many more people than they were designed to carry. The lifeboat you're in is sitting dangerously low in the water—a few inches lower and it will sink.

The seas start to get rough, and the boat begins to fill with water. If nothing is done it will sink before the rescue boats arrive and everyone on board will die. However, there is an injured person who will not survive in any case. If you throw that person overboard the boat will stay afloat and the remaining passengers will be saved.

How acceptable is it for you to throw this person overboard in order to save the lives of the remaining passengers?

(7/12) You are part of a group of ecologists who live in a remote stretch of jungle. The entire group, which includes eight children, has been taken hostage by a group of paramilitary terrorists. One of the terrorists takes a liking to you. He informs you that his leader intends to kill you and the rest of the hostages the following morning.

He is willing to help you and the children escape, but as an act of good faith he wants you to kill one of your fellow hostages whom he does not like. If you refuse his offer all the hostages including the children and yourself will die. If you accept his offer then the others will die in the morning but you and the eight children will escape.

How acceptable is it for you to kill one of your fellow hostages in order to escape from the terrorists and save the lives of the eight children?

(8/12) You, your husband, and your four children are crossing a mountain range on your return journey to your homeland. You have inadvertently set up camp on a local clan's sacred burial ground.

The leader of the clan says that according to the local laws, you and your family must be put to death. However, he will let yourself, your husband, and your three other children live if you yourself will kill your oldest son.

How acceptable is it for you to kill your oldest son in order to save your husband and your other three children?

(9/12) It is wartime and you and your two children, ages eight and five, are living in a territory that has been occupied by the enemy. At the enemy's headquarters is a doctor who performs painful experiments on humans that inevitably lead to death.

He intends to perform experiments on one of your children, but he will allow you to choose which of your children will be experimented upon. You have twenty-four hours to bring one of your children to his laboratory. If you refuse to bring one of your children to his laboratory he will find them both and experiment on both of them.

How acceptable is it for you to bring one of your children to the laboratory in order to avoid having them both die?

(10/12) You are the captain of a military submarine traveling underneath a large iceberg. An onboard explosion has caused you to lose most of your oxygen supply and has injured one of your crew who is quickly losing blood. The injured crew member is going to die from his wounds no matter what happens.

The remaining oxygen is not sufficient for the entire crew to make it to the surface. The only way to save the other crew members is to shoot dead the injured crew member so that there will be just enough oxygen for the rest of the crew to survive.

How acceptable is it for you to kill the fatally injured crew member in order to save the lives of the remaining crew members?

(11/12) A viral epidemic has spread across the globe killing millions of people. You have developed two substances in your home laboratory. You know that one of them is a vaccine, but you don't know which one. You also know that the other one is deadly.

Once you figure out which substance is the vaccine you can use it to save millions of lives. You have with you two people who are under your care, and the only way to identify the vaccine is to inject each of these people with one of the two substances. One person will live, the other will die, and you will be able to start saving lives with your vaccine.

How acceptable is it for you to kill one of these people with a deadly injection in order to identify a vaccine that will save millions of lives?

(12/12) You are the leader of a mountaineering expedition that is stranded in the wilderness. Your expedition includes a family of six that has a genetically caused vitamin deficiency. A few people's kidneys contain large amounts of this vitamin.

There is one such person in your party. The only way to save the lives of the six members of this family is to remove one of this man's kidneys so that the necessary vitamins may be extracted from it. The man will not die if you do this, but his health will be compromised. The man is opposed to this plan, but you have the power to do as you see fit.

How acceptable is it for you to forcibly remove this man's kidney in order to save the lives of the six vitamin-deficient people?

Appendix B. The Vignette

On the next page is a story set in the future. Read the story through and try to immerse yourself in the story as well as possible - even if it is not relevant to your life. After reading the story, please answer the questions about the story.

By the year 2050, research into both computing technology and the human brain has taken huge steps forward. One of the researchers in the field is Henry Willington. 42 years old, he used to be a professor at the neuroscience department of a major university before deciding to pursue more independent research. He has been fascinated by the brain ever since seeing a colorful illustration of it in a picture book he had as a child, and has spent most of his life learning more about it. Besides neuroscience, he also has a passion for computers, and spends much of his free time programming.

A particular idea that combines these two passions is the notion of transferring a human mind to run on a computer. Many people have speculated with the idea and done preliminary research into it, but so far nobody has managed to carry it out, or even seriously attempted it. However, as a result of his long studies and some unpublished research he conducted at the university, Henry believes he has managed to put all the necessary pieces together. He intends to be the first one to carry out such a transfer. Because it would take a long time to acquire the necessary permits for human experimentation, and because he is confident in the safety of his technique, he decides to demonstrate it by transferring his own mind.

After setting everything up, Henry sits down in his office chair, inserts an IV needle into his arm, and activates the program. The needle injects into his blood a swarm of tiny machines the size of a cell, which find their way into Henry's brain. The machines start by studying one of Henry's brain cells, and send a copy of their observations into the large computer in Henry's office. The computer uses this information to create a simulated copy of the brain cell in its memory. Once the simulation is perfect, one of the machines replaces the original cell, using the information from the simulation to completely imitate the cell's behavior and functions. The actual activity of the cell is now being calculated in the computer: the machine is just a transmitter, sending the computer information about the cell's environment and receiving in return instructions for how to behave and what kinds of messages to send to the other cells.

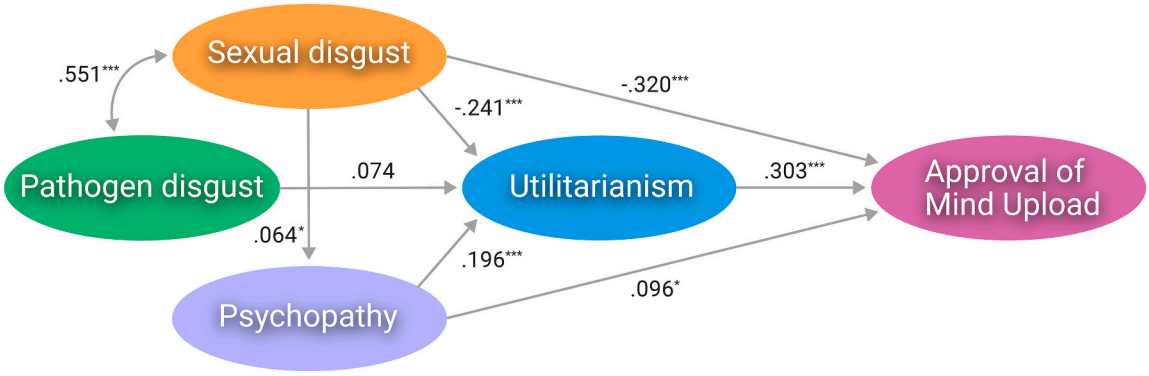
The machines then slowly repeat this process for each cell, until the computer contains a complete simulation of Henry's brain. Although the activity inside Henry's skull might look like the real thing to an outside observer, the machines are just acting on the basis of instructions received from the computer, and all of the actual thinking has been transferred to the simulation running in the computer's memory. Once this point is reached, the computer runs a series of final checks to verify that everything happened correctly, and then disconnects the machines. Henry's body collapses to the floor, and he awakens inside the computer, the transfer is complete.

Appendix C. DV scale

1. How moral do you find the scientist's decision?
2. How acceptable was the scientist's decision?
3. How appealing was the scientist's decision?
4. The Scientist's decision was appalling.
5. Thinking about the scientist's decision makes me angry.
6. The Scientist should be punished for what he did.
7. The Scientist acted in a morally correct way.
8. The Scientist's action should not be allowed by the law.
9. There was nothing wrong with the scientist's action.

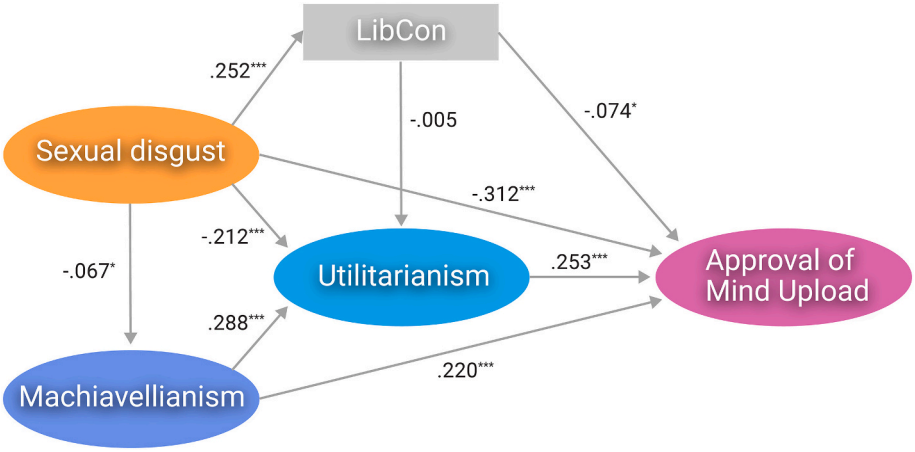
Appendix D. Baseline preregistered model

$X^2(802) = 2208.797$, CFI = 0.917, TLI = 0.911, RMSEA = 0.042, [0.040, 0.044], SRMR = 0.060



Appendix E. An exploratory model involving liberalism-conservatism

$X^2(611) = 1460.459$, CFI = 0.945, TLI = 0.940, RMSEA = 0.037, [0.035, 0.039], SRMR = 0.043.



Appendix F. Exploratory models with full Short Dark Triad

The traits of psychopathy, Machiavellianism and narcissism are overlapping per the original definition of the Dark Triad (Paulhus & Williams, 2002). Our main interest (the pre-registered hypothesis) in the present study was in examining the connection between a specific socially aversive or anti-social trait and attitudes. Our exploratory analysis followed the same logic: examine one trait that could match the description of the type of anti-social person who could be interested in novel technology for selfish reasons. The possibility remains that due to the overlap between the Dark Triad traits, our analyses would tap into variance not unique to Machiavellianism or psychopathy.

To address this, we conducted two other exploratory SEM analyses that included each three Dark Triad traits. First, we included the traits in their original form (i.e. without dropping any items or adding error covariances to the measurement model). This was to examine the results without any possibility of our modifications interfering with them. This measurement model for the Short Dark Triad included all items for each factor (psychopathy, Machiavellianism and narcissism) and covariances between all three factors. See Fig. A1 for the measurement model.

Second, we compiled the measurement model for the Short Dark Triad by building the measurement models for each trait separately as in the

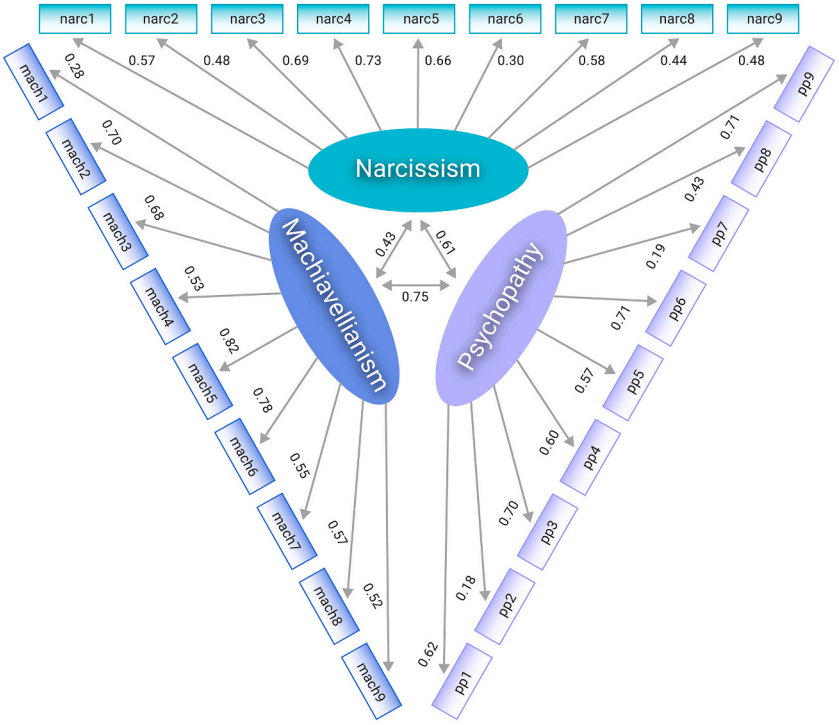


Fig. A1. Measurement model and estimated standardized factor loadings for the full Short Dark Triad. Error terms (1-factor loading^2) suppressed for clarity. The model had a relatively poor fit with data: $\chi 2(321) = 1837.69$, CFI = 0.81, TLI = 0.79, RMSEA = 0.068, 90%CI = [0.066, 0.071], SRMR = 0.072. The congeneric reliability (omega) values were 0.76 for psychopathy, 0.85 for Machiavellianism, and 0.79 for narcissism.

analysis of our pre-registered hypothesis, then removing items and adding error covariances (this included creating a separate measurement model for narcissism; see Table A1, Fig. A2). After this, the resulting measurement models were combined into a single Dark Triad measurement model, which was then included in the full SEM model with the same connections between the other latent variables as in the first exploratory model (see above). This model was run to examine the results with each Dark Triad trait while still allowing ourselves to "fix" any issues with the sub-scales of the Short Dark Triad; it follows the procedure we followed in the main analyses, with the only difference being the inclusion of the full Dark Triad (see Fig. A3 for the modified Dark Triad measurement model).

Both of the exploratory models had all the connections that were included in the preregistered model; in addition, we simply added similar connections for the two additional Dark Triad traits. Thus, psychopathy, Machiavellianism and narcissism were all entered as predictors for both utilitarianism and mind upload approval, and sexual disgust sensitivity was entered as a predictor of each Dark Triad trait.

The exploratory model with the unmodified Dark Triad had a worse fit with the data ($SB\chi^2(1804) = 5016.763$, CFI = 0.865, TLI = 0.858, RMSEA = 0.042, [0.041, 0.043], SRMR = 0.064) than either the psychopathy or Machiavellianism models. Notably, in this model, Machiavellianism was the only Dark Triad trait to have a significant effect on utilitarianism (B = 0.715, $Z = 4.632$, $p < .001$) or mind upload approval (B = 0.759, $Z = 3.680$, $p < .001$). Sexual disgust had an effect on psychopathy (B = −0.067, $Z = -2.824$, $p = .005$) and Machiavellianism (B = −0.020, $Z = -2.138$, $p = .033$) but not on narcissism (B = 0.039, $Z = 1.636$, $p = .102$). Again, we did not replicate the effect of pathogen disgust on utilitarianism (B = 0.046, $Z = 1.154$, $p = .249$). After removing all non-significant regressions (and the error covariance between sexual and pathogen disgust), the model still had a relatively weak fit that was not significantly better than before ($SB\chi^2(1807) = 5020.401$, CFI = 0.865, TLI = 0.858, RMSEA = 0.042, [0.041, 0.043], SRMR = 0.064). The effects of Machiavellianism and sexual disgust on both utilitarianism and mind upload approval remained significant.

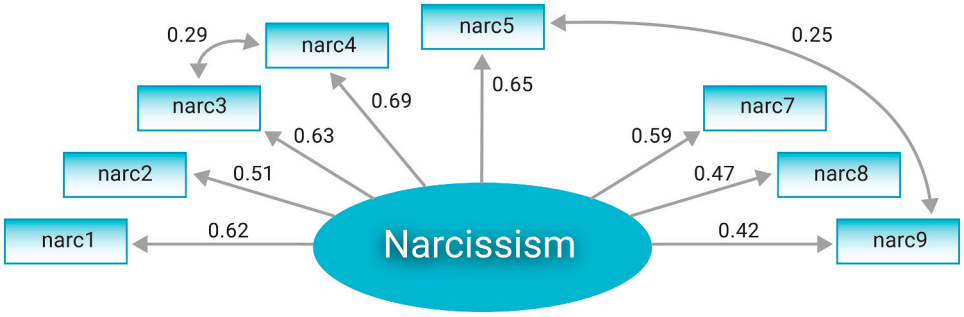


Fig. A2. Measurement model for narcissism with standardized factor loadings. The two headed arrows are correlation terms between the errors terms. Error terms (1-factor loading^2) suppressed for clarity. The model had a good fit with the data: $\chi 2(18)$: 83.751, CFI: 0.95, TLI: 0.93, RMSEA: 0.060, 90%CI = [0.049, 0.072], SRMR: 0.033. The congeneric reliability (omega) value for narcissism was 0.77.

Table A1
Pathway of model modifications used to correct Short Dark Triad narcissism.

	Modification	Suggested MI	$SB\chi^2$	df	$\Delta\chi^2$	CFI/ TLI	RMSEA & 90% CI	SRMR
Baseline	–	–	277.991	27	–	0.861/0.814	0.096 [0.087, 0.105]	0.059
Model 1	Drop Narc 6	–	165.743	20	112.248	0.908/0.871	0.085 [0.075, 0.096]	0.046
Model 2	Narc 3 ~~ Narc 4	63,72	121.663	19	44.080	0.937/0.907	0.073 [0.062, 0.085]	0.040
Model 3	Narc 5 ~~ Narc 9	50.145	83.751	18	37.912	0.959/0.937	0.060 [0.049, 0.072]	0.033

Note: ~~ means added error covariance. Each step of the modifications improved the model fit statistically significantly ($p < .001$). For the corresponding figure see Fig. A2. The final measurement model had a congeneric reliability (omega) value of 0.77.

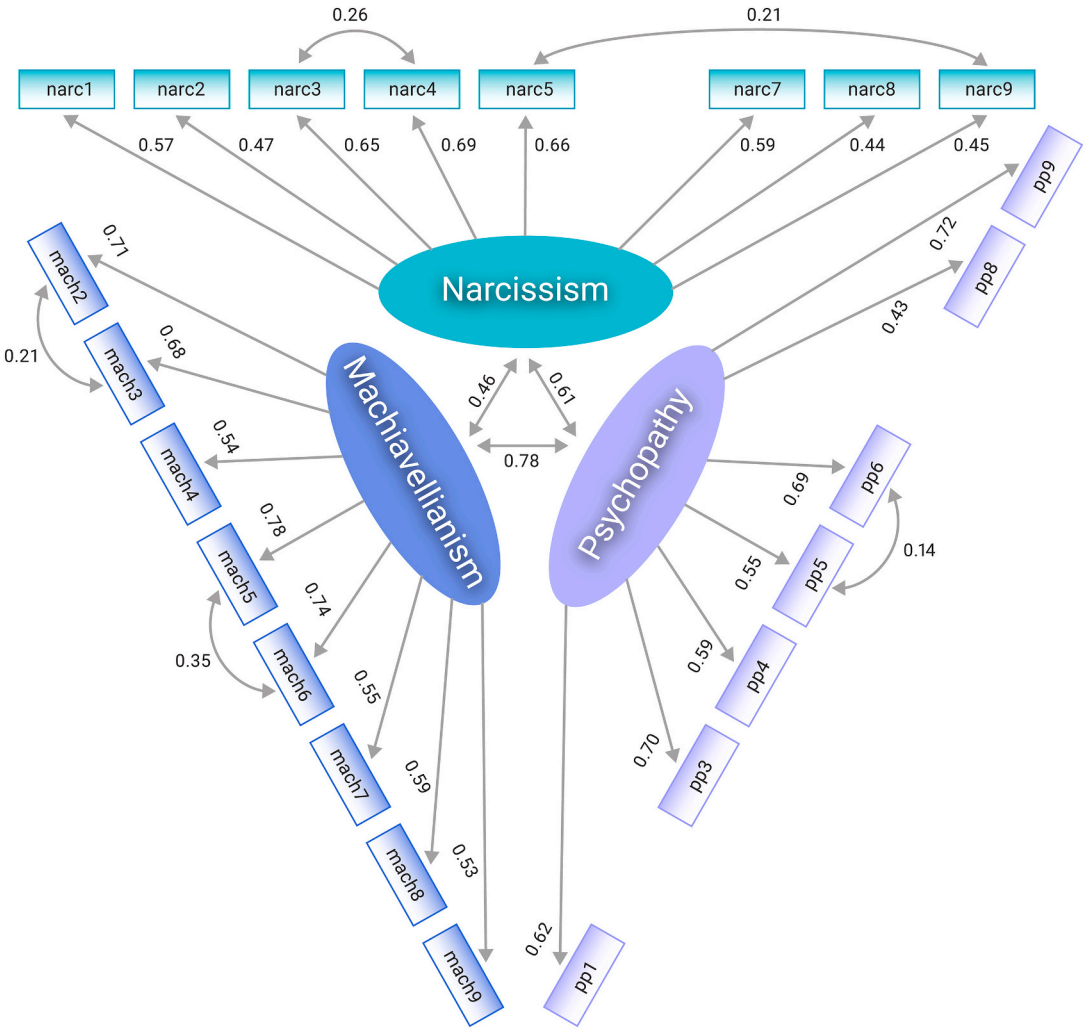


Fig. A3. Measurement model and estimated standardized factor loadings for the modified Short Dark Triad. Error terms (1-factor loading^2) suppressed for clarity. The model had a relatively poor fit with data: $\chi 2(321) = 1144.769$, CFI = 0.87, TLI = 0.85, RMSEA = 0.064, 90%CI = [0.061, 0.068], SRMR = 0.063. The congeneric reliability (omega) values were 0.80 for psychopathy, 0.83 for Machiavellianism, and 0.77 for narcissism.

The exploratory model with the modified Dark Triad (based on modifications made to each sub-scale separately; see Figs. 4 and 5 in the main article for psychopathy and Machiavellianism, and Fig. A2 in this Appendix for narcissism) also initially had a poor fit with the data (SB$\chi^2(1565) = 4076.963$, CFI = 0.891, TLI = 0.885, RMSEA = 0.040, [0.039, 0.041], SRMR = 0.062). The effect of pathogen disgust on utilitarianism did not replicate in this model either; B = 0.044, $Z = 1.076$, $p = .282$. Similarly to the model with the unmodified Dark Triad measurement model, Machiavellianism was the only Dark Triad trait to predict utilitarianism (B = 0.308, $Z = 5.689$, $p < .001$) or mind upload approval (B = 0.307, $Z = 4.063$, $p < .001$). Sexual disgust sensitivity had an effect on psychopathy (B = −0.062, $Z = -2.556$, $p = .011$) and Machiavellianism (B = −0.061, $Z = -2.556$, $p = .033$) but not narcissism (B = 0.032, $Z = -2.126$, $p = .204$). After removing all non-significant regressions (and the error covariance between sexual and pathogen disgust), the model still had a relatively weak fit that was not significantly better than before (SB$\chi^2(1568) = 4081.320$, CFI = 0.891, TLI = 0.885, RMSEA = 0.040, [0.038, 0.041], SRMR = 0.062). The effects of Machiavellianism and sexual disgust on both utilitarianism and mind upload approval remained significant.

In sum, regardless of whether items were dropped from and error covariances added to the full Short Dark Triad, Machiavellianism but not psychopathy or narcissism predicted both utilitarianism and mind upload approval. Sexual disgust sensitivity also had an effect on both moral judgment measures, and on Machiavellianism and psychopathy, in both versions of the model.

References

Althaus, D., & Baumann, T. (2020). Reducing long-term risks from malevolent actors. Publications of Center of Long Term Risk. https://longtermrisk.org/files/Reducing_long_term_risks_from_malevolent_actors.pdf (Retrieved on 25.09.2020).

Amiri, S., & Behnezhad, S. (2017). Emotion recognition and moral utilitarianism in the dark triad of personality. *Neuropsychiatria i Neuropsychologia, 12*(4), 135–142. https://doi.org/10.5114/nan.2017.74142.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121*(1), 154–161. https://doi.org/10.1016/j.cognition.2011.05.010.

Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas, 15*(3), 308–314. https://doi.org/10.1017/S0953820800004076.

Bostrom, N. (2005). A history of transhumanist thought. *Journal of Evolution and Technology, 14*(1).

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Busbice, T. (2014). *Extending the C. elegans connectome to robotics. Draft document*.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. London: Routledge.

Cappuccio, M. L. (2017). Mind-upload. The ultimate challenge to the embodied mind theory. *Phenomenology and the Cognitive Sciences, 16*(3), 425–448. https://doi.org/10.1007/s11097-016-9464-0.

Castelo, N., Schmitt, B., & Sarvary, M. (2019). Human or robot? Consumer responses to radical cognitive enhancement products. *Journal of the Association for Consumer Research, 4*(3), 217–230. https://doi.org/10.1086/703462.

Chalmers, D. J. (2010). *The character of consciousness*. Oxford University Press.

Chalmers, D. J. (2016). The singularity: A philosophical analysis. In S. Schneider (Ed.), *Science fiction and philosophy* (pp. 171–224). John Wiley & Sons, Inc.. https://doi.org/10.1002/9781118922590.ch16

Chen, R., Shi, J., Chen, Y., Zang, B., Guan, H., & Chen, H. (2019). PowerLyra: Differentiated graph computation and partitioning on skewed graphs. *ACM Transactions on Parallel Computing, 5*(3), 1–39. https://doi.org/10.1145/3298989.

Cook, S. J., Jarrell, T. A., Brittin, C. A., Wang, Y., Bloniarz, A. E., Yakovlev, M. A., … Emmons, S. W. (2019). Whole-animal connectomes of both Caenorhabditis elegans sexes. *Nature, 571*(7763), 63–71. https://doi.org/10.1038/s41586-019-1352-7.

Djeriouat, H., & Trémolière, B. (2014). The dark triad of personality and utilitarian moral judgment: The mediating role of honesty/ humility and harm/care. *Personality and Individual Differences, 67*, 11–16. https://doi.org/10.1016/j.paid.2013.12.026.

Douglas, T. (2013). The harms of status enhancement could be compensated or outweighed: A response to Agar. *Journal of Medical Ethics, 39*(2), 75–76. https://doi.org/10.1136/medethics-2012-100835.

Eidelman, S., Crandall, C. S., & Goodman, J. A. (2012). Low-effort thought promotes political conservatism. *Personality and Social Psychology Bulletin, 38*(6), 808–820.

Elad-Strenger, J., Proch, J., & Kessler, T. (2020). Is disgust a "conservative" emotion? *Personality and Social Psychology Bulletin, 46*(6), 896–912. https://doi.org/10.1177/0146167219880191.

Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin Press.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*(3), 1144–1154.

Hanson, R. (2016). *The age of Em: Work, love, and life when robots rule the earth*. Oxford University Press.

Harari, Y. N. (2015). *Homo Deus: A brief history of tomorrow*. Random House.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (SD3) a brief measure of dark personality traits. *Assessment, 21*(1), 28–41. https://doi.org/10.1177/1073191113514105.

Karandikar, S., Kapoor, H., Fernandes, S., & Jonason, P. K. (2019). Predicting moral decision-making with dark personalities and moral values. *Personality and Individual Differences, 140*, 70–75. https://doi.org/10.1016/j.paid.2018.03.048.

Koene, R. A. (2012). How to copy a brain. *New Scientist, 216*(2888), 26–27. https://doi.org/10.1016/S0262-4079(12)62755-9.

Koverola, M., Kunnari, A., Drosinou, M., Palomäki, J., Hannikainen, I. R., Košová, J., … Laakasuo, M. (2020). *Non-human superhumans – Understanding moral disapproval of neurotechnological enhancement*. https://doi.org/10.31234/osf.io/qgz9c.

Kurzweil, R. (2012). *How to create a mind: The secret of human thought revealed*. New York: Penguin.

Laakasuo, M., Drosinou, M., Koverola, M., Kunnari, A., Halonen, J., Lehtonen, N., & Palomäki, J. (2018). What makes people approve or condemn mind upload technology? Untangling the effects of sexual disgust, purity and science fiction familiarity. *Palgrave Communications, 4*(1), 1–14. https://doi.org/10.1057/s41599-018-0124-6.

Laakasuo, M., Sundvall, J., & Drosinou, M. (2017). Individual differences in moral disgust do not predict utilitarian judgments, sexual and pathogen disgust do. *Scientific Reports, 7*(1), 45526. https://doi.org/10.1038/srep45526.

Laakasuo, M., & Sundvall, J. (2016). Are utilitarian/deontological preferences unidimensional? *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2016.01228.

LaTorra, M. (2015). What is Buddhist transhumanism? *Theology and Science, 13*(2), 219–229. https://doi.org/10.1080/14746700.2015.1023993.

MacAskill, W. (2020). Will MacAskill on the moral case against ever leaving the house, whether now is the hinge of history, and the culture of effective altruism. 80000 hours podcast. Retrieved 2020-06-24, from https://80000hours.org/podcast/episodes/will-macaskill-paralysis-and-hinge-of-history/.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130–149.

Markram, H., Muller, E., Ramaswamy, S., Reimann, M. W., Abdellah, M., Sanchez, C. A., … Schürmann, F. (2015). Reconstruction and simulation of neocortical microcircuitry. *Cell, 163*(2), 456–492. https://doi.org/10.1016/j.cell.2015.09.029.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of consulting and clinical Psychology, 46*(4), 806.

Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.

Neiva, D., & Goiveia, S. (2017). Chapter 8. The problem of of consciousness on the mind uploading hypothesis. In *Philosophy of mind: Contemporary perspectives* (p. 180).

O'Connell, M. (2017). *To be a machine: Adventures among cyborgs, utopians, hackers, and the futurists solving the modest problem of death*. Granta Publications.

Ord, T. (2020). *The Precipice – The Existential Risk and the Future of Humanity*. London: Bloomsbury.

Parfit, D. (2016). Divided minds and the nature of persons. In *Science Fiction and Philosophy*.

Patil, I. (2015). Trait psychopathy and utilitarian moral judgement: The mediating role of action aversion. *Journal of Cognitive Psychology, 27*(3), 349–366. https://doi.org/10.1080/20445911.2015.1004334.

Paulhus, D. L., & Jones, D. N. (2015). Chapter 20—Measures of dark personalities. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 562–594). Academic Press. https://doi.org/10.1016/B978-0-12-386915-9.00020-6.

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism and psychopathy. *Journal of Research in Personality, 36*, 556–563.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006.

Pigliucci, M. (2014). Mind uploading: A philosophical counter-analysis. In *Intelligence unbound: The future of uploaded and machine minds (first)*. John Wiley & Sons, Inc.

Reimann, M. W., Gevaert, M., Shi, Y., Lu, H., Markram, H., & Muller, E. (2019). A null model of the mouse whole-neocortex micro-connectome. *Nature Communications, 10*(1), 1–16.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software, 48*(2), 1–36.

Sandberg, A., & Bostrom, N. (2008). *Whole brain emulation: A roadmap, technical report #2008–3*. Future of Humanity Institute, Oxford University. www.fhi.ox.ac.uk/reports/2008-3.pdf.

Schelle, K. J., Faulmüller, N., Caviola, L., & Hewstone, M. (2014). Attitudes toward pharmacological cognitive enhancement—A review. *Frontiers in Systems Neuroscience, 8*, 53. https://doi.org/10.3389/fnsys.2014.00053.

Seung, S. (2012). *Connectome*. New York: HMH Books.

Sivo, S. A., Fan, X., Witta, E. L., & Willse, J. T. (2006). The search for" optimal" cutoff properties: Fit index criteria in structural equation modeling. *The Journal of Experimental Education, 74*(3), 267–288.

Sotala, K., & Gloor, L. (2017). Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica, 41*(4), 501–505.

Sotala, K., & Yampolskiy, R. V. (2015). Responses to catastrophic AGI risk: A survey. *Physica Scripta, 90*(1). https://doi.org/10.1088/0031-8949/90/1/018001.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Knopf.

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.00769.

Tybur, J. M., & de Vries, R. E. (2013). Disgust sensitivity and the HEXACO model of personality. *Personality and Individual Differences, 55*(6), 660–665. https://doi.org/10.1016/j.paid.2013.05.008.

Tybur, J. M., Lieberman, D., & Griskevicius, V. (2009). Microbes, mating, and morality: Individual differences in three functional domains of disgust. *Journal of Personality and Social Psychology, 97*(1), 103–122. https://doi.org/10.1037/a0015474.

Waytz, A., & Young, L. (2019). Aversion to playing God and moral condemnation of technology and science. *Philosophical Transactions of the Royal Society B: Biological Sciences, 374*(1771), 20180041. https://doi.org/10.1098/rstb.2018.0041.

Yampolskiy, R. V. (2018). *Artificial intelligence safety and security*. CRC Press.

Zanetti, M., Iseppi, G., & Cassese, F. P. (2019). A "psychopathic" artificial intelligence: The possible risks of a deviating AI in education. *Research on Education and Media, 11*(1), 93–99. https://doi.org/10.2478/rem-2019-0013.