# Reducing long-term risks from malevolent actors

David Althaus, Tobias Baumann


## Contents

## 1 Summary


- Dictators who exhibited highly narcissistic, psychopathic, or sadistic traits were involved in some of the greatest catastrophes in human history. (More)
- Malevolent individuals in positions of power could negatively affect humanity's long-term trajectory by, for example, exacerbating international conflict or other broad risk factors. (More)
- Malevolent humans with access to advanced technology—such as whole brain emulation or other forms of transformative AI—could cause serious existential risks and suffering risks. (More)
- We therefore consider interventions to reduce the expected influence of malevolent humans on the long-term future.
  - The development of manipulation-proof measures of malevolence seems valuable, since they could be used to screen for malevolent humans in high-impact settings, such as heads of government or CEOs. (More)
  - We also explore possible future technologies that may offer unprecedented leverage to mitigate against malevolent traits. (More)
  - Selecting against psychopathic and sadistic tendencies in genetically enhanced, highly intelligent humans might be particularly important. However, risks of unintended negative consequences must be handled with extreme caution. (More)
- We argue that further work on reducing malevolence would be valuable from many moral perspectives and constitutes a promising focus area for longtermist EAs. (More)


## 2 What do we mean by malevolence?

Before we make any claims about the causal effects of malevolence, we first need to explain what we mean by the term. To this end, consider some of the arguably most evil humans in history—Hitler, Mao, and Stalin—and the distinct personality traits they seem to have shared.[1]

Stalin repeatedly turned against former comrades and friends (Hershman & Lieb, 1994, ch. 15, ch. 18), gave detailed instructions on how to torture his victims, ordered their loved ones to watch (Glad, 2002, 2002, p. 13), and deliberately killed millions through various atrocities. Likewise, millions of people were tortured and murdered under Mao's rule, often according to his detailed instructions (Dikötter, 2010, 2016; Chang & Halliday, 2007, ch. 8, ch. 23, 2007). He also took pleasure in watching acts of torture and imitating in what his victims went through (Chang & Halliday, 2007, ch. 48, 2007). Hitler was not only responsible for the death of millions, he also engaged in personal sadism. On his specific instructions, the plotters of the 1944 assassination attempt were hung by piano wires and their agonizing deaths were filmed (Glad, 2002, 2002). According to Albert Speer, "Hitler loved the film and had it shown over and over again" (Toland, 1976, 1976, p. 818). Hitler, Mao, and Stalin—and most other dictators—also poured enormous resources into the creation of personality cults, manifesting their colossal narcissism Dikötter (2020). (The section Malevolent traits of Hitler, Mao, Stalin, and other dictators in Appendix B provides more evidence.)

Many scientific constructs of human malevolence could be used to summarize the relevant psychological traits shared by Hitler, Mao, Stalin, and other malevolent individuals in positions of power. We focus on the **Dark Tetrad** traits Paulhus (2014) because they seem especially relevant and have been studied extensively by psychologists. The Dark Tetrad comprises the following four traits—the more well-known Dark Triad (Paulhus & Williams, 2002, 2002) refers to the first three traits:

- Machiavellianism is characterized by manipulating and deceiving others to further one's own interests, indifference to morality, and obsession with achieving power or wealth.
- Narcissism involves an inflated sense of one's importance and abilities, an excessive need for ad-

[1] Of course, assessing other people's personality is always fraught with uncertainty, especially if they are long dead.

miration, a lack of empathy, and obsession with achieving fame or power.
- Psychopathy is characterized by boldness, callousness, impulsiveness, a lack of empathy and guilt, and antisocial behavior, including violence and crime.
- Sadism involves deriving pleasure from inflicting suffering and pain on others.

There is considerable overlap between the Dark Tetrad traits. In general, almost all plausible operationalizations of malevolence tend to positively correlate with each other and negatively with "benevolent" traits such as altruism, humility or honesty. (See the section Correlations between dark traits and other traits in Appendix B for more details.)

This suggests the existence of a **general factor** of human malevolence[2]: the Dark Factor of Personality (Moshagen et al., 2018, et al., 2018)—analogous to $g$, the general factor of intelligence—characterized by egoism, lack of empathy[3] and guilt, Machiavellianism, moral disengagement, narcissism, psychopathy, sadism, and spitefulness. Like most personality traits (Johnson et al., 2008, et al., 2008), malevolent traits seem relatively stable over the lifespan Obradović et al. (2007) and influenced by genetic factors Vernon et al. (2008), but more on this below.[4]

Throughout this article, we will assume a dimensional—rather than categorical, "black-or-white"—conception of malevolence. That is, we believe that malevolent traits exist on a continuum—just like most other human traits such as extraversion or intelligence cf. Haslam et al. (2012); (Plomin, 2019, ch. 5). Slight Machiavellian or sadistic tendencies, for example, are common. Many humans seem to flatter their superiors and enjoy seeing (non-tragic) mishaps of their political opponents. But only a few individuals will derive pleasure from witnessing human torture or will kill their former friends just to consolidate their power.

It is this latter type of human—showing clear signs of at least some highly elevated Dark Tetrad traits—who we have in mind when we use the term "malevolent".

## 3 Malevolent humans in power pose serious long-term risks

In this section we discuss why and how malevolent individuals in highly influential positions—such as political leaders or CEOs of notable companies—could negatively affect humanity's long-term trajectory, ultimately increasing existential risks (including extinction risks) and risks of astronomical suffering (s-risks).

### 3.1 Malevolent humans often rise to power

Malevolent humans are unlikely to substantially affect the long-term future if they cannot rise to power. But alas, they often do. The most salient examples are dictators who clearly exhibited elevated malevolent traits: not only Hitler, Mao, and Stalin, but also Saddam Hussein, Mussolini, Kim Il-sung, Kim Jong-il, Duvalier, Ceauşescu, and Pol Pot, among many others.

In fact, people with increased malevolent traits might even be overrepresented among business Babiak et al. (2010); Boddy et al. (2010); Lilienfeld et al. (2014), military, and political leaders Post (2003); Lilienfeld et al. (2012), perhaps because malevolent traits—especially Machiavellianism and narcissism—often entail an obsession with gaining power and fame Kajonius et al. (2015); Lee et al. (2013); Southard & Zeigler-Hill (2016) and could even be advantageous in gaining power Deluga (2001); Taylor (2019). Again, Appendix B provides more details.

### 3.2 History suggests that malevolent leaders have caused enormous harm

One reason for expecting malevolent humans in power to pose risks to the future is that they seem to have caused great harm in the past.

Hitler, Mao, and Stalin were directly involved in several of the greatest atrocities in history, such as World War II, the Holocaust, the Great Leap Forward, the Cultural Revolution, and the Great Terror. There thus seems to be a correlation between the malevolence of (autocratic) political leaders and the amount of harm

[2] Other researchers have suggested similar constructs aimed to represent the common core of "evil" (e.g., (Book et al., 2015, 2015); (D. N. Jones & Figueredo, 2013, 2013); (Marcus et al., 2018, 2018)).

[3] Baron-Cohen (2012) argues that the defining feature of human evil is "zero degrees of empathy." However, some psychopaths can read other people extremely well and would thus score highly on certain items of the empathy questionnaires Baron-Cohen describes in his book. Furthermore, as Baron-Cohen acknowledges, people on the autism spectrum tend to have less empathy—at least certain forms of it—but they are not more malevolent than the population average. Therefore, reducing malevolence to "zero degrees of empathy" could be problematic or at least crucially depends on how we define and operationalize empathy.

[4] As of now, there is no established treatment of malevolence. Harris and Rice (2006) review the empirical findings on the treatment of psychopathy but are quite pessimistic about their effectiveness.

[5] A more rigorous analysis would be valuable, though it would also be methodologically challenging—assessing the personality traits of historical figures, for example, is rather difficult.

that occurred under their rule; at least according to our understanding of history.[5] If the past is any guide to the future, individuals with highly elevated dark traits could again manage to rise to positions of extreme power and cause extraordinary harm.

However, correlation does not imply causation. Even if we grant that this correlation indeed exists[6], one could argue that there are better explanations for how these atrocities came about. In particular, it seems plausible that other factors—such as political instability or extremist ideologies—matter most. We discuss these issues in more detail in this section of Appendix A.

It's also worth mentioning that individuals with malevolent personalities are more likely to adopt dangerous ideologies. Dark Tetrad traits are associated with political extremism generally, including supporting the use of violence to achieve political and other ideological goals Duspara & Greitemeyer (2017); Međedović & Knežević (2018); Gøtzsche-Astrup (2019); D. N. Jones (2013).[7]

Thus, while we agree that history is largely shaped by economic, political, cultural, institutional, ideological and other systemic forces, we believe that the personality traits of individual leaders—at the very least in autocratic regimes—can plausibly make a substantial difference as well (see also (Bertoli et al., 2019; Byman & Pollack, 2001, especially p. 115-121), Gallagher & Allen (2014); B. F. Jones & Olken (2005)). After all, there were humans who rose to power within rather autocratic regimes but who nevertheless enacted relatively beneficial policies. Examples include Juan Carlos I of Spain, Mustafa Kemal Atatürk, Mikhail Gorbachev, Lee Kuan Yew, and Marcus Aurelius, who seemed to exhibit more benevolent personality traits than the likes of Hitler, Mao, and Stalin.

### 3.3 Malevolent leaders have the potential to corrupt humanity's long-term future

One could question whether malevolent individuals can substantially influence the long-term trajectory of humanity for the worse, even from positions of extreme power. It is possible that they only cause short-term harm, in which case reducing malevolence may not be a priority from a longtermist perspective.

However, we believe malevolent leaders plausibly have a significant detrimental effect on the long-term future. Hitler, Stalin, and Mao, for instance, seemed to have had a profoundly negative influence on global affairs and international cooperation, some of which can arguably still be felt today, more than half a century after the atrocities they perpetrated.[8] That said, it is difficult—if not impossible—to assess long-term impacts, as we do not know what would have happened counterfactually.

#### 3.3.1 Broad risk factors due to malevolent leaders

Beckstead (2013) asks whether there is "a common set of broad factors which, if we push on them, systematically lead to better futures". It seems plausible that malevolent humans in power would push such factors in the wrong direction.[9]

Specifically, we conjecture that malevolent humans in power would affect the risk factors below in the following ways:

- Increase the spread of political extremism and other dangerous ideologies (see again Appendix A).
- Exacerbate the risk of great power wars and international conflict (Byman & Pollack, 2001, particularly p. 112, 134, 137-138); Gallagher & Allen (2014)[10], including the risk of nuclear war and arms races involving transformative AI.
- Increase the likelihood of the formation of a global

[6] One reason to be hesitant here is that it seems plausible, for instance, that we, as well as journalists and historians, see more signs of malevolent personality traits in leaders who have caused great harm, and will tend to overlook malevolent personality traits in leaders who have done more good.

[7] Again, we refer to Appendix A for more details.

[8] For example, without Mao and Stalin the probability of a communist China is smaller. A non-communist China may have better relations with the U.S., and the probability of great power wars and (AI) arms races may be reduced. However, such claims are necessarily very speculative. For instance, one could also argue that World War II may have increased longer-term stability by leading to the formation of the UN.

[9] Or, as Robert Hare, one of the most well-known researchers of psychopathy, puts it: "Serial killer psychopaths ruin families. Corporate, political and religious psychopaths ruin economies [and] societies." (Ronson, 2012, p. 117).

[10] Gallagher & Allen (2014) found that U.S. presidents scoring higher on the Big Five facet "altruism" were less likely to employ military force.

[11] Also compare MacAskill: "I still endorse the view of pushing resources into the future. The biggest caveat actually I'd have is about the rise of fascism and Stalinism as the thing to push on [...] even though you might not think that a particular ideology will last forever, well, if it lasts as long until you get like some eternal lock-in event, then it lasts forever. [...] I kind of think the rise of fascism and Stalinism was a bigger deal in the 20th century than the invention of nuclear weapons." MacAskill (2020).

totalitarian regime, potentially resulting in a permanent lock-in of harmful values and power structures.[11]

- Increase the likelihood of reckless behaviour, rather than careful reflection, in high-stakes situations (for example, those resembling the Cuban Missile Crisis).
  - Dark Triad traits, psychopathy in particular, are associated with extreme risk-taking Hosker-Field et al. (2016); Visser et al. (2014).
- Increase intranational conflict and undermine public institutions, social coordination, collective decision making and general discourse[12], particularly by:
  - Exacerbating economic and social inequality.[13]
  - Increasing corruption (Bendahan et al., 2015, table 5), rent-seeking, and the risk of financial crises Boddy (2011).
  - Reducing access to information, e.g., through censorship and propaganda.[14]
  - Reducing trust in government and institutions Bowler & Karp (2004).[15]

Such trends would plausibly lead to worse futures in expectation. They also plausibly increase existential risks (including extinction risks) and suffering risks (see the next section). However, the evidence linking these risk factors to malevolent humans in power is fairly weak, for various reasons. We are therefore only somewhat confident in these connections.

### 3.3.2 Existential and suffering risks due to malevolent leaders

In terms of more concrete scenarios, the most extreme risks to the long-term future would arguably result from malevolent humans with access to highly advanced technology, particularly transformative AI.

The following list outlines some (non-exhaustive) examples of how malevolent individuals could increase existential and suffering risks:

- As noted above, malevolent individuals tend to exhibit more risk-taking behaviour. In the context of a project to develop and deploy transformative AI, they are therefore more likely to ignore potential warning signs and omit precautionary measures. This increases the risk of misaligned transformative AI.
- Malevolent humans are likely less opposed to making threats than the average human Jonason et al. (2012); Ullrich et al. (2001)[16] and plausibly less motivated to pursue peaceful bargaining strategies. Conflicts involving malevolent humans are therefore significantly more likely to escalate and result in catastrophic outcomes. Also, it could be dangerous if AI systems inherit some of their values or heuristics, such as an increased willingness to make and carry out threats and/or a reduced willingness to compromise.
- Advanced technology might enable sadistic individuals in power to create suffering on an unprecedented scale.
- A malevolent individual, or a small group of such individuals—e.g., the inner circle of an autocratic state—might manage to obtain control of Earth (cf. MacAskill, 2020)[17], and eventually the observable universe. For example, imagine Hitler or Stalin had access to advanced technology—including aligned AGI and mind uploading, enabling immortality. Such a lock-in of permanent rule by a (global) malevolent dictator would clearly qualify as an existential risk, as it would thwart any prospect of a more valuable future. It also constitutes a significant s-risk as there would be nobody left to keep any sadistic tendencies of the dictator in check.

While specific scenarios are necessarily speculative, it seems clear that malevolent leaders pose a serious threat to humanity's long-term future. Of course, malevolent leaders are not the root of all evil, and

[12] Dark Triad traits in political candidates correlate with more negative campaigns and fear appeals Nai (2019).

[13] Since Dark Triad traits correlate with social dominance orientation (SDO, Jones & Figueredo, 2013), malevolent leaders will, on average, exhibit higher SDO and prefer policies resulting in higher social and economic inequality.

[14] Azizli et al. (2016) find that psychopathy and Machiavellianism are associated with a greater propensity to lie and engage in high-stakes deception.

[15] Bowler & Karp (2004) find that scandals involving politicians tend to lower political trust. It seems plausible that malevolent political leaders are more likely to be involved in scandals.

[16] According to Ulrich et al. (2001), the rate of antisocial personality disorder among criminal offenders, 45% of whom were convicted of "robbery or extortion," is more than 10 times higher than that of the control sample. Jonason et al. (2012) also find that Machiavellianism, narcissism, and psychopathy all correlate with the use of "hard tactics" in the workplace, including "threats of punishment" (see Table 1, p. 451).

[17] MacAskill (2020, emphasis added): "[...] when you look at history of well what are the worst catastrophes ever? They fall into three main camps: pandemics, war and totalitarianism. Also, totalitarianism or, well, autocracy has been the default mode for almost everyone in history. And I get quite worried about that. So *even if you don't think that AI is going to take over the world, well it still could be some individual.* And if it is a new growth mode, I do think that very significantly increases the chance of lock-in technology."

many conflicts, wars and atrocities would happen without them. Nevertheless, we believe that preventing malevolent individuals from rising to power is likely valuable and robustly positive, according to almost all moral perspectives (compare also Beckstead, 2013; Tomasik, 2013a, 2013b).

# 4 Interventions to reduce the influence of malevolent actors

## 4.1 Advancing the science of malevolence

Further research into the construct of malevolence and its consequences would allow us to make more rigorous statements about the links between malevolent leaders and bad outcomes.

A more established science of malevolence would also help raise awareness of malevolent personality traits and how to detect them among the general public, influencers, politicians, researchers, and academics. Generally, the more we know about malevolence, the easier it is to accomplish many of the interventions discussed below.

### 4.1.1 Developing better constructs and measures of malevolence

It seems worthwhile to develop constructs capturing more precisely the constellation of traits most worrisome from a longtermist perspective, as existing constructs will not always do so.

For example, one of the most commonly used scales to measure psychopathy, the Psychopathy Checklist-Revised by Hare et al. (1990), consists of 20 items, grouped into two factors. Factor 1—characterized by cruelty, grandiosity, manipulativeness, and a lack of guilt—arguably represents the core personality traits of psychopathy. However, scoring highly on factor 2—characterized by impulsivity, reactive anger, and lack of realistic goals—is less problematic from our perspective. In fact, humans scoring high on factor 1 but low on factor 2 are probably *more* dangerous than humans scoring high on both factors (more on this below). Generally, most measures of psychopathy include items related to increased impulsivity (e.g., Cooke & Michie (2001); Levenson et al. (1995); Lilienfeld & Andrews (1996)).

#### 4.1.1.1 The most dangerous individuals tend to go undiagnosed

Individuals officially diagnosed as malevolent—e.g. those diagnosed with psychopathy, antisocial or narcissistic personality disorder—are probably unrepresentative of the most dangerous individuals. This is because an official diagnosis is only made when somebody suffers from immediate and severe problems (relating to their malevolence) or was forced to seek therapy, e.g., because they committed a crime.

In contrast, malevolent humans with good impulse-control and otherwise decent mental health have no reason to seek out a therapist and will generally not be convicted of crimes. The most dangerous malevolent humans will realize that not being unmasked as malevolent is of the highest importance, and will have sufficient motivation, cunning, self-awareness, charisma, social skills, intelligence and impulse-control to avoid detection Perina et al. (2020).[18] Such individuals might even deliberately display personality characteristics entirely at odds with their actual personality. In fact, many dictators did precisely that and portrayed themselves—often successfully—as selfless visionaries, tirelessly working for the greater good (e.g., Dikötter (2020)). It may therefore be very valuable to conduct more research on this hard-to-detect type of conscientious, strategic malevolence (cf., e.g., Gao & Raine (2010); Lilienfeld et al. (2015); Mullins-Sweatt et al. (2010)).

### 4.1.2 Manipulation-proof measures of malevolence

To prevent malevolent humans from reaching highly influential positions, we need to be able to reliably detect those traits.

Currently, most measures of dark traits take the form of either interviews or self-completed questionnaires. Smart malevolent humans can easily manipulate these types of instruments and evade detection by lying. It is key, therefore, that we develop *manipulation-proof* measures of malevolence, i.e., measures that cannot (easily) be gamed.

One possibility would be to ask peers and previous associates to evaluate the personality traits of the person in question.[19] Of course, this raises several problems. Malevolent humans could have charmed and fooled

[18] Kaja Perina on the Manifold podcast Perina et al. (2020): "Most of the studies on psychopaths [...] are done on inmates. For that reason, we're forced to conjecture about the really successful ones because I think the more successful, the more they evade detection, perhaps, lifelong. So there is this disconnect wherein a lot of them, the violent ones, the less intelligent ones, really end up in jail, and these are the ones who are studied, but these are not the ones who are highly Machiavellian, necessarily, these are not the ones who are brilliantly manipulative. These are the ones who are committing violent crimes and get caught".

[19] Extensive background checks, for example with the help of private investigators, would be another promising possibility. Intelligence agencies do this already for somewhat related purposes. Generally, the competitive nature of the political process can often uncover past immoral behavior—though swaying partisan views seems to require evidential strength that is difficult to achieve. Thanks to Mojmír Stehlík for raising these points.

many of their (former) friends and colleagues. They could also bribe or manipulate others to lie. So, while other-report measures (e.g., 360 degree assessments) may be harder to manipulate than self-reported ones and are therefore valuable, they are unlikely to completely solve the problem.[20]

Physiological or neurobiological measures based on methods like EEG or fMRI might be particularly difficult to manipulate—though this would probably require substantial technological and scientific progress. Neuroimaging techniques might allow us to identify abnormal brain structures or detect suspicious behavior, such as showing neurological signs of pleasure and/or no distress when seeing other humans or animals in pain. Therefore, more neurobiological research on the neurological signatures of pleasure and displeasure (e.g., Berridge & Kringelbach (2013)), and on the neurobiology of sadism and psychopathy, might be very valuable.[21] (Note that we have not investigated this in detail, so it is probably best to start with a systematic literature review). However, such methods also raise ethical questions about judging people by brain scans rather than their actual behavior.

#### 4.1.2.1 Potential misuse and negative consequences

Manipulation-proof measures of malevolence could also be misused—like all technology. For instance, governments might falsely brand political opponents as psychopaths.

Another concern is that such tests may constitute an unfair form of discrimination against humans with certain traits. This is because they measure innate characteristics that are impossible to change, rather than exclusively considering the actual behaviour of individuals. Also, even if this is deemed acceptable in the case of malevolence, advocating for testing in this context might lead to the widespread adoption of personality testing in general, which some believe could have negative consequences. (On the other hand, existing selection procedures also implicitly or explicitly select based on innate traits such as intelligence, and also include various kinds of tests.)

Lastly, unless tests of malevolence have perfect validity and reliability, there will be measurement errors: Some people will be diagnosed as highly malevolent even though they aren't, and some truly malevolent people will escape detection.

#### 4.1.2.2 How valuable would manipulation-proof measures of malevolence be in practice?

Given the potentially enormous benefits, why has there been so little interest in the development of manipulation-proof tests of malevolence? First, doing so is likely difficult and, especially if it involves neuroscience research, expensive[22] (as an example, MRI machines cost between $0.3M and $3M). Second, malevolent humans might, in some cases, actually benefit individual companies or political parties: high levels of psychopathy and narcissism could be useful for things like negotiating, motivating employees, or winning public approval. Third, most people likely overestimate their ability to discern malevolent traits in others, making them less interested in such tests. Finally, it seems that tests in general are not used much in at least some contexts; for example, most elected or appointed positions in government do not require intelligence, knowledge, or personality tests.[23]

One might argue that it was obvious to most people that dictators such as Hitler, Mao, and Stalin were malevolent even before they gained power, and that manipulation-proof measures of malevolence would therefore have been useless. However, we are doubtful that people can easily detect malevolence, at least in the most dangerous types of individuals, as mentioned above. (See also the section How well can people detect malevolent traits in Appendix A for more details.) Of course, many did suspect that Hitler, Mao, and Stalin were malevolent. However, this was not common knowledge—and without objective evidence, calling an individual malevolent can easily be dismissed as slander. Not to mention that anyone making such accusations risks serious reprisals. So even if a majority had realized early on that Hitler, Mao, and Stalin are malevolent, it might not have helped.

[20] Yet another possibility would be to use "objective" personality tests that don't rely on self- or other-report but use actual performance tests to evaluate personality traits (without the test-taker knowing which trait is supposed to be measured). However, according to our cursory reading of the literature, the few "objective" personality tests that exist seem to have low validity (e.g., Kline & Cooper (1984)).

[21] However, one needs examples to train such predictors in the first place. One could start by looking for differences in the brains of normal people and, say, diagnosed psychopaths, but this metric will be biased towards diagnosed psychopaths who are at least somewhat unrepresentative of non-diagnosed malevolent humans (as explained above). One needs to correct for this ascertainment bias.

[22] Relatedly, neuroscience research is often underpowered, resulting in low reproducibility of the accumulated findings Button et al. (2013).

[23] However, the education system also involves a lot of tests and grading; and is at least somewhat related to career advancement. Such tests are also common for military entry and sometimes civil service. In the context of elections, the key question is why voters do not generally demand such tests (including related objective measures, such as tax returns).

However, if manipulation-proof and valid measures of malevolence had existed—alongside strong norms to use them to screen political leaders and widespread trust in their accuracy—it could have been common knowledge that these individuals were malevolent, which would have significantly reduced their chance of rising to power. Essentially, manipulation-proof and valid measures of malevolence could serve as an objective arbiter of good intentions, analogous to the scientific use of experiments as objective arbiters of truth. Their role in a hiring process could then be compared to security clearances, for instance.

It is not clear whether even perfectly diagnostic measures of malevolence would ever become widespread—for example, because of the abovementioned ethical concerns. However, for highly influential positions, people are most willing to make use of all available evidence (and candidates for such positions have an incentive to provide credible signals of trustworthiness). For instance, receiving a top-secret security clearance involves extensive interviews with one's (former) spouses, colleagues, friends and neighbors, alongside reviews of medical and psychiatric records, and sometimes even polygraph examinations. This elaborate and arguably privacy-violating process would be unacceptable for a routine job, but is considered appropriate given the stakes at hand. Last, it could already be valuable if only a few companies or government departments started using manipulation-proof measures of malevolence; near-universal adoption of such measures is by no means necessary.

Despite these caveats, we believe that work on manipulation-proof measures of malevolence is promising. Subject to personal fit, it may be worthwhile for some effective altruists to consider careers in psychology or neuroscience. This would allow them to advance the science of malevolence, contribute to the development of manipulation-proof measures of malevolence, and improve their chances to convince decision-makers to take such measures seriously.

### 4.2 Political interventions

Many factors determine whether an individual can rise to a position of power, and it is important to include (non-)malevolence as a criterion when selecting leaders. Ideally, we should establish strong norms against allowing highly malevolent leaders to rise to power—even in cases where elevated Dark Tetrad traits may be instrumental in advancing the interests of a company or nation.

While the notion of Dark Tetrad traits is not foremost in most people's minds, one could argue that much political debate is about related concepts like the trustworthiness or honesty of candidates, and voters do value those attributes.[24]

Perhaps the key issue, then, is not a lack of awareness; rather the non-availability of reliable objective measures and the overestimation of people's ability to detect malevolence. In fact, humans seem *too* eager to view their political opponents as inherently malevolent and ill-intentioned. Conversely however, humans also tend to view members of their own tribe as inherently good and overlook their misdeeds. (See again Appendix A for more details.)

The media also tends to depict impulsive psychopaths—say, ruthless serial killers with a long history of violence or crime. These are relatively easy to detect, potentially leading to a false sense of security (compare also (Babiak et al., 2010, p.174-175)). As mentioned above, it may therefore be valuable to raise awareness that at least some types of malevolent humans are difficult to detect.

Alternatively, we could influence political background factors that make malevolent leaders more or less likely. It seems plausible that political instability, especially outright revolutions, enable malevolent humans to rise to power (Colgan, 2013, p. 662-665). Generally, democracies plausibly select for more trustworthy, predictable and benevolent leaders (Byman & Pollack, 2001, p.139-140). Thus, interventions to promote democracy and reduce political instability seem valuable—though this area seems rather crowded.

Even within established democracies, we could try to identify measures that avoid excessive polarization and instead reward cross-party cooperation and compromise. Mitigating the often highly combative nature of politics would plausibly make it harder for malevolent humans to rise to power.[25] (For example, effective altruists have discussed electoral reform as a possible

[24] However, there exists the frightening possibility that some voters want their political leaders to be at least moderately malevolent. Most Russians and Chinese, for example, seem to think highly of Stalin and Mao, respectively—though this is likely at least partly due to propaganda. Generally, many voters seem to like "strong men" like Putin and overlook or even appreciate elevated Dark Tetrad traits in their political leaders. Also, according to the (potentially biased) assessment of "experts", politicians with autocratic tendencies—many of whom nonetheless received the majority of votes—score significantly higher on Dark Triad traits than the average politician Nai & Toros (2020).

[25] It is also worth noting that in some forms of government, such as allocating political positions to randomly selected individuals or hereditary monarchy, those in positions in power are exactly as likely to be malevolent as the population at large. This may be better than fierce competition for positions of power if the latter advantages the most ruthless and malevolent individuals. On the other hand, good selection procedures could also reduce malevolence in positions of power below the baseline; and of course this is only one consideration among many when evaluating different forms of government.

lever that could help achieve this.)

Since elevated Dark Tetrad traits are significantly more common among men Paulhus & Williams (2002); Plouffe et al. (2017), it also seems beneficial to advance gender equality and increase the proportion of female leaders.

Other potential factors that might facilitate the rise of malevolent individuals include social and economic inequality, poverty, ethnic, military or religious conflicts, and a "widespread sense of grievance or resentment" (Glad, 2002, p. 4). Thus, identifying cost-effective interventions to improve these factors (as well as identifying factors we haven't thought of) could be promising. A more thorough study of the history of malevolent humans rising to power would also be valuable to better understand which factors are most predictive.

Overall, it seems plausible that many promising political interventions to prevent malevolent humans from rising to power have already been identified and implemented—such as, e.g., checks and balances, the separation of powers, and democracy itself. After all, much of political science and political philosophy is about preventing the concentration of power in the wrong hands.[26] We nevertheless encourage interested readers to further explore these topics.

## 4.3 Future technologies and malevolence

In this section, we explore how possible future technologies could be used to reduce the influence of malevolent actors.

### 4.3.1 Whole brain emulation

Whole brain emulation is the hypothetical process of scanning the structure of a brain and replicating it on a computer. Hanson (2016) explores the possible implications of this technology. In his scenario, brain emulations ("ems") will shape future economic, technological and political processes due to their competitive advantage over biological minds.

One key question is: which human brains will be uploaded? We believe that it would be crucial to screen potential ems for malevolence—particularly the first individuals to be uploaded. Considering the power that the first ems would likely have to shape this new "Age of Em", it could be disastrous for humanity's long-term future if a malevolent individual forms the basis for (some of) the first ems (cf. Bostrom, 2002, p. 12). Conversely, by screening for malevolence, using manipulation-proof measures, we could effectively reduce malevolence among ems. This offers an unprecedented opportunity to ensure that malevolent forces have significantly less influence over the long-term future.[27]

### 4.3.2 Transformative AI

Many longtermist effective altruists think that shaping transformative artificial intelligence, and in particular solving the alignment problem, is a particularly good lever to improve the long-term future. Some concrete proposals for alignment—such as Iterated distillation and amplification—involve a "human-in-the-loop" whose feedback is used to align increasingly capable AI.

In these scenarios, the "human-in-the-loop" plausibly has enormous responsibility and leverage over the long-term future. It is therefore extremely valuable to ensure that the relevant individual or individuals—if e.g. a jury or parliament fulfills the role of "human-in-the-loop"—do not exhibit malevolent traits. (Again, this requires or is at least facilitated by the availability of manipulation-proof measures.)

Even without human involvement, artificial agents may exhibit behaviour that resembles malevolence (to the extent that this notion makes sense in non-human contexts) if such heuristics prove useful in its training process. After all, the fact that malevolent traits such as psychopathy or sadism evolved in some humans suggests that those traits provided fitness advantages, at least in certain contexts Book et al. (2015); McDonald et al. (2012); Nell (2006); Jonason et al. (2015).

In particular, it is possible that domain-general capabilities will emerge via increasingly complex multi-agent interactions (Babiak et al., 2019). In this case, it is crucial that the training environment is set up in a way that prevents the evolution of undesirable traits like malevolence, and instead rewards cooperative and trustworthy behaviour.

To the extent that artificial intelligence designs are inspired by the human brain ("neuromorphic AI"), it seems important to understand the neuroscientific basis of malevolence in humans to reduce the risk of neuromorphic AIs also exhibiting malevolent traits.

### 4.3.3 Genetic enhancement

A third class of relevant new technologies are those that make it possible to change the genetic makeup of future humans. This would offer unprecedented leverage to change personality traits and "human nature", for better or for worse (cf. Genetic Enhancement as a Cause Area). In particular, selection against malevolent traits could significantly reduce the influence of

[26] Thanks to Richard Ngo for making this point.

[27] However, initial distributions may change due to competitive pressures or other factors. Even if none of the first ems are malevolent, there is no guarantee that malevolence will remain absent in the long run.

malevolent individuals.

This is because most variance in adult personality is due to genetic influences ( 30–50%) and nonshared environment effects ( 35–55%), leaving comparatively little room for the shared environment ( 5–25%)[28] (e.g., (Knopik et al., 2018, ch. 16); Johnson et al. (2008); Plomin (2019); Vukasović & Bratko (2015)). (See the sections "Broad-sense heritability estimates of dark traits" and "Is selecting for personality traits possible?" in Appendix B for more details.) By contrast, nonshared environmental influences—which include measurement error, chance life events, and de novo mutations—seem to be mostly unsystematic, idiosyncratic, and unstable, and therefore difficult to influence (Plomin, 2019, ch. 7).

Genetic enhancement technologies might also result in the creation of humans with extraordinary intelligence (see, e.g., Shulman & Bostrom, 2014, p.2-3). Such humans, if created, will likely be overrepresented in positions of enormous influence and would thus have an outsized impact on the long-term future. Reducing malevolence among those individuals is therefore especially important.

#### 4.3.3.1 Overview of genetic enhancement technologies

There are various technologies that would make it possible to modify the genetic makeup of future humans. We think the following four are most relevant:

- In vitro fertilization (IVF) is a process of fertilisation where an egg is combined with sperm outside the body. In several Western countries, 2-8% of newborns are already conceived in this way. While this is primarily used to address infertility, it is possible to create several fertilized eggs and select among those.
- Gene editing (e.g., via CRISPR) is the insertion, deletion, modification or replacement of DNA in the genome of an organism.
- Iterated embryo selection (IES, Shulman & Bostrom, 2014; Sparrow (2014)) takes a sample of embryos and repeats two steps: a) select embryos that are higher in desired genetic characteristics; b) extract stem cells from those embryos, convert them to sperm and ova, and cross those to produce new embryos.
- Genome synthesis is the artificial manufacturing of DNA, base pair by base pair.

Note that these methods can interact with each other and should thus not be viewed as being completely separate. For more details, we highly recommend Gwern's Embryo selection for intelligence.

How far away are these technologies? Gwern writes that "IES is still distant and depends on a large number of wet lab breakthroughs and finetuned human-cell protocols." Nonetheless, he states that: "[...] it seems clear, at least, that it will certainly not happen in the next decade, but after that...?". He concludes that "IES has been badly under-discussed to date."

Regarding genome synthesis, Gwern writes that the "cost curve suggests that around 2035, whole human genomes reach well-resourced research project ranges of $10-30m" and that it "is entirely possible that IES will develop too slowly and will be obsoleted by genome synthesis in 10-20 years."

Gwern gives the following summary:

"CRISPR & cloning are already available but will remain unimportant indefinitely for various fundamental reasons; [...] massive multiple embryo selection is some ways off but increasingly inevitable and the gains are large enough on both individual & societal levels to result in a shock; IES will come sometime after massive multiple embryo selection but it's impossible to say when, although the consequences are potentially global; genome synthesis is a similar level of seriousness, but is much more predictable and can be looked for, very loosely, 2030-2040 (and possibly sooner)."

#### 4.3.3.2 Dangers

Genetic enhancement is widely criticized. Numerous atrocities have been committed in the quest to forge a new, "better" kind of human. We would like to emphasize that we do not advocate for genetic enhancement per se. We only argue that *if* genetic enhancement happens, it seems prima facie important to select against malevolent traits—comparable to the rationale behind differential intellectual progress and differential technological progress.

Still, we are treading dangerous waters. Even just bringing up the possibility of selection for or against personality traits might inspire misuse of such methods. One particularly worrisome scenario is selection against all forms of rebellion and independence, branded as "antisocial tendencies", which could enable extreme totalitarianism. Generally, the currently dominant individuals and classes could abuse these powerful technologies to cement their power.

It is also worth noting that very high levels of usually beneficial traits can be negative: too much trust, for example, might result in naïvety and an increased likelihood of being exploited. Similarly, completely eliminating usually harmful traits could backfire as well: for

[28] Extreme events like severe abuse or violence can make a huge difference for the victims, but such events are relatively rare and therefore do not explain much variance in the general population Plomin (2019).

example, in certain situations, some degree of narcissism and Machiavellianism may benefit entrepreneurs and politicians. Generally, different personality traits are useful for different roles in society, so some diversity is beneficial.

For these and other reasons, it could be net negative to shift personality traits by more than one or two standard deviations.[29] However, we are primarily interested in shrinking the *right tail* of the distribution—e.g., by selecting against embryos with polygenic scores for Dark Tetrad traits above, say, the 99th percentile. This could be done while only marginally decreasing the mean. Generally, if we apply the (double) reversal test, current rates of dark traits—particularly highly elevated ones—appear very far from optimal.

Moreover, many dark traits appear to be genetically correlated with each other and negatively genetically correlated with benevolent personality traits Vernon et al. (2008). Thus, selecting against one dark trait will tend to decrease other dark traits and increase benevolent traits. This plausibly makes selection efforts more robust, though this could also have some downsides.[30]

Lastly, we should arguably be especially cautious in scenarios that involve genetically enhanced humans of extraordinary intelligence. Extremely intelligent sadists and psychopaths would pose risks that outweigh any plausible benefits.

#### 4.3.3.3 Interventions

Funding or otherwise encouraging more research on the genetic basis of malevolent traits would allow us to better select against these traits. Ideally, we would have a good understanding of the genetic basis of malevolent traits *before* technologies such as genome synthesis arrive. Thus, it is plausibly time-sensitive to do this research now, even if powerful genetic enhancement technologies will not be developed for the next several decades.

A particularly cost-effective intervention might be to convince personal genomics companies, such as 23andMe, to offer tests of Dark Tetrad traits. 23andMe has over 10 million customers, so even if only a small fraction of customers took these tests (e.g., out of curiosity), we would already achieve sample sizes surpassing those of large GWA studies.[31] Improved psychological measures of malevolence with higher reliability and validity, as discussed in previous sections, would also enable GWA studies to better identify genetic variants associated with such traits.

In general, increasing the social acceptability of screening for Dark Tetrad traits plausibly increases the probability that future projects involving more powerful technology will also do so. The more established and well-known Dark Tetrad traits are, and the less controversial their heritability, the easier it would be to accomplish many of the interventions mentioned above. It might be valuable, for instance, to persuade sperm banks or other institutions responsible for screening sperm (or egg) donors to add measures of Dark Tetrad traits to their screening process and display the results prominently to women choosing sperm donors.

As with non-genetic interventions, we could attempt to raise awareness of malevolent traits, their heritability, and their dangers. Rather than trying to make changes to the supply side, it might be easier to increase demand by popularizing Dark Tetrad traits.[32] Most parents want their children to be responsible, empathic, and kind. If they are willing to pay for screening for malevolent traits, then sperm banks or others will offer such services.[33]

However, considering the significant dangers outlined above, we believe that public advocacy of the idea of genetic selection against malevolence would likely be premature. Indeed, more research on how to best avoid negative consequences—such as increased inequality or dehumanization of (un)enhanced humans—of possible interventions in this area would be important.

Subject to personal fit, it may also be worthwhile for some effective altruists to consider careers in bioinformatics, social sciences relating to GWA studies, bioethics, or related fields, to be in a good position

[29] Also note that shifting personality traits by more than this would likely be very difficult even if one wanted to do this.

[30] Some dark traits, such as Machiavellianism, can be beneficial under certain circumstances. It might be better if one could single out a dark trait, such as sadism, and only select against it while leaving other dark traits unchanged.

[31] However, GWA studies of personality are still fairly weak even at such scales. Even higher sample sizes might be achieved by identifying proxy variables of malevolence, such as public records on crime. However, this could easily backfire and cause great harm in numerous ways, so one would have to be very careful.

[32] Currently, Dark Tetrad traits seem to be neglected in many relevant areas. Most sperm banks measure traits such as height, attractiveness, physical and mental health but not Dark Tetrad traits. Services that offer pre-implementation diagnostics screen for all sorts of genetic diseases and some even for IQ but not for Dark Tetrad traits. Test batteries of enormous government projects like the UK Biobank measure thousands of variables—including physical health, height, and preferred coffee and cereal type—but they don't measure Dark Tetrad traits or even most personality traits in any sort of rigorous manner.

[33] Generally, reducing the influence of malevolent actors can be done in myriad (often mutually reinforcing) ways. For instance, the more widespread belief systems are that put great value on non-malevolent traits such as compassion and altruism, the more parents might demand screening for malevolent traits, and the more future (government) projects will include measures of malevolence in their test batteries.

to later influence key players.

## 5 Concluding remarks

Many of the above interventions face serious technical challenges. It may be hard to develop manipulation-proof measures of malevolence, and selection on personality traits is probably difficult due to low additive heritability. In addition, many interventions—especially those related to genetic enhancement technologies—entail severe risks of misuse and unintended negative consequences.

However, some of the suggested interventions involve neither speculative future technology nor controversial ideas about genetic enhancement. Overall, we recommend a mix of different interventions, as well as further work aiming to find new types of interventions and checking the assumptions that underlie existing interventions.

Most parents, cultures, and religions feature some notion of "not being evil", so one could argue that reducing malevolence, broadly construed, is already quite crowded. However, we believe the interventions we have explored are more targeted, and are potentially more far-reaching and more neglected than, say, cultural norms or parenting.

Reducing the influence of malevolent actors is not a panacea, of course. Many of the world's biggest problems are not (primarily) due to malevolent intent per se, and instead are mostly caused by incompetence, irrationality, indifference, and our inability to coordinate the escape from undesirable equilibria.

That being said, we believe that reducing the chances of malevolent individuals rising to power would have substantially positive effects under a broad range of scenarios and value systems—whether they place primary importance on avoiding existential risks, reducing suffering, or improving the quality of the long-term future.

## Appendix A

### A.1 How important are situational factors and ideologies compared to personality traits?

In this section, we discuss the extent to which historical atrocities can be attributed to the personality traits of individuals versus structural factors.

First, it seems plausible that background conditions that enable dictatorships in the first place—such as political instability and an absent rule of law—also make it more likely that malevolent humans will rise to power. Individuals who are reluctant to engage in murder and betrayal, for example, will be at a considerable disadvantage under such conditions (also see (Colgan, 2013, especially p. 662-665)).

Similarly, power tends to corrupt (e.g., Bendahan et al. (2015); Cislak et al. (2018)) so it could be argued that most individuals who rise to the top within autocratic regimes, will become more malevolent. Generally, a wealth of social psychology research attests to the importance of situational factors in explaining human behavior Milgram (1963); Burger (2009), though the understanding of modern psychology is that behavior depends on both situational factors and individual personality traits Bowers (1973); Endler & Magnusson (1976).

One particularly relevant factor is the spread of extremist and fanatical ideologies such as fascism, violent communism, and fundamentalist religion, which have undoubtedly contributed to historical atrocities. In fact, such ideologies have plausibly had a much bigger impact on history than the personality traits of individuals and could pose even greater risks to the long-term future. So why focus on personality rather than ideology or structural factors?

For one, tens of millions of people are already combating the dangerous ideologies mentioned above, or work on ensuring political stability and rule of law. These efforts are laudable, but also seem very crowded, which suggests that many of the most cost-effective interventions have already been identified and carried out.

As mentioned above, there is also ample evidence that individuals with malevolent personalities are drawn to dangerous ideologies:[34] Dark Triad traits predict increased intention to engage in political violence Gøtzsche-Astrup (2019). Narcissism and psychopathy are associated with political extremism Duspara & Greitemeyer (2017). Sadistic and psychopathic traits predict endorsing a militant extremist mind-set, in particular the use of violence to achieve political and other ideological goals Međedović & Knežević (2018). Machiavellianism and psychopathy predict racist attitudes, including support for Neo-Nazis and the KKK D. N. Jones (2013). Dark Triad traits correlate with social-dominance orientation D. N. Jones & Figueredo (2013); D. N. Jones (2013), a measure of an individual's preference for economic and social inequality

[34] Adopting certain ideologies could also make one more malevolent. However, we think it's plausible that most of the correlation is explained by causation from malevolent traits to dangerous ideologies, partly because personality traits seem less amenable to change than beliefs.

[35] The Dark Triad also predicts sexism Gluck et al. (2020); O'Connell & Marcus (2016), nationalism Matthews et al. (2018) as well as cognitive and affective prejudice Koehn et al. (2019). Psychopathic traits predict opposition towards free speech and animal

within and between groups Pratto et al. (1994); Dallago et al. (2008).[35]

Most ideologies also seem open for interpretation, leaving sufficient room for the idiosyncratic beliefs and personality traits of leaders to make a difference. Khrushchev and Gorbachev, for example, while broadly sharing Stalin's Marxist-Leninist ideology, have caused much less harm than Stalin. Conversely, as the examples of Stalinism, Maoism, and Juche show, malevolent individuals can develop an existing ideology further, making it even more harmful.

In the end, ideologies, belief systems, and personality traits appear inevitably intertwined. Narcissism, for example, entails inflated beliefs about one's abilities and place in history, by definition. Generally, malevolent individuals tend to hold beliefs that serve as (un)conscious justifications for their behavior, such as a sense of entitlement and grandiosity, and seem more likely to endorse dangerous worldviews and "ideologies that favor dominance (of individuals or groups)" (Moshagen et al., 2018, p. 659).

Finally, it is instructive to compare large-scale atrocities to small-scale atrocities like murder or contract killing. While rates of violent crime surely depend on social background factors and culturally transmitted norms, psychopathy is also considered a strong predictor for homicide, including instrumental, calculated murder Fox & DeLisi (2019). If we accept that malevolent personality traits like psychopathy play a causal role in violent crime, it stands to reason that such traits also play at least some causal role in many large-scale atrocities.

### A.2 How well can people detect malevolent traits?

Historical evidence suggests that even many of their political adversaries—at least for some time—did not realize that Hitler, Mao, and Stalin were malevolent, even *after* they were in power.

Chamberlain famously trusted Hitler's sincerity for far too long. Churchill once remarked that "Poor Chamberlain believed he could trust Hitler. He was wrong. But I don't think that I am wrong about Stalin" (Yergin, 1977, p. 65). Similarly, Truman believed that Stalin "could be depended upon. . . I got the impression Stalin would stand by his agreements" (Larson, 1988, p. 246).[36] At least until the 1940s, many Westerners and Chinese seemed to have been enamored with Mao, potentially partly due to the influential book 'Red Star over China' (Snow, 1937) which painted him in an extremely favorable light (Chang & Halliday, 2007, ch. 18).

Countless famous intellectuals—including G.B. Shaw, H.G. Wells, Sartre, Simone de Beauvoir, Beatrice Webb, Sidney Webb, Susan Sontag, Oswald Spengler, Carl Jung, Konrad Lorenz, and Martin Heidegger—praised authoritarian leaders like Mussolini, Hitler, Stalin or Fidel Castro Hollander (2016, 2017). Even today, most Russians and Chinese think highly of Stalin and Mao, respectively.

In summary, it seems that many humans fail to detect malevolent individuals, particularly when ideological, patriotic or other biases affect their judgment. Generally, Hitler, Mao, and Stalin—like many narcissists—seem to have been quite polarizing; some thought they were obviously malevolent, others viewed them as benevolent, nearly messianic figures.

## Appendix B

See Appendix B: Reducing long-term risks from malevolent actors for additional details.

## Acknowledgements

Thanks to Jesse Clifton, Jonas Vollmer, Lukas Gloor, Stefan Torges, Chi Nguyen, Mojmír Stehlík, Richard Ngo, Pablo Stafforini, Caspar Oesterheld, Lucius Caviola, Johannes Treutlein, and Ewelina Tur for their valuable comments and feedback. Thanks to Sofia-Davis Fogel for copy editing. All errors and views expressed in this document are our own, not those of the commenters.

David's work on this post was funded by the Center on Long-Term Risk.

## References

Azizli, N., Atkinson, B. E., Baughman, H. M., Chin, K., Vernon, P. A., Harris, E., & Veselka, L. (2016). Lies and crimes: Dark triad, misconduct, and high-stakes deception. *Personality and Individual Differences*, *89*, 34–39.

Babiak, P., Neumann, C. S., & Hare, R. D. (2010). Corporate psychopathy: Talking the walk. *Behavioral sciences & the law*, *28*(2), 174–193.

rights as well as support for using war as a tool for diplomacy (Preston & Anestis, 2018, Table 3). Machiavellianism and narcissism also seem to correlate with overconfidence Campbell et al. (2002, 2004); Macenczak et al. (2016); Jain & Bearden (2011), and thus plausibly negatively correlate with epistemic humility which should serve a protective function against all kinds of extremism and fanaticism. (Note that we don't intend to convey that all these associations are equally dangerous).

[36] Many of these and similar quotes could have been made solely for political reasons, e.g., to strengthen alliances.

Bendahan, S., Zehnder, C., Pralong, F. P., & Antonakis, J. (2015). Leader corruption depends on power and testosterone. *The Leadership Quarterly*, *26*(2), 101–122.

Berridge, K. C., & Kringelbach, M. L. (2013). Neuroscience of affect: brain mechanisms of pleasure and displeasure. *Current opinion in neurobiology*, *23*(3), 294–303.

Bertoli, A., Dafoe, A., & Trager, R. (2019). Leader age and international conflict.

Boddy, C. R. (2011). The corporate psychopaths theory of the global financial crisis. *Journal of Business Ethics*, *102*(2), 255–259.

Boddy, C. R., Ladyshewsky, R., & Galvin, P. (2010). Leaders without ethics in global business: Corporate psychopaths. *Journal of Public Affairs*, *10*(3), 121–138.

Book, A., Visser, B. A., & Volk, A. A. (2015). Unpacking "evil": Claiming the core of the dark triad. *Personality and Individual Differences*, *73*, 29–38.

Bowers, K. S. (1973). Situationism in psychology: An analysis and a critique. *Psychological review*, *80*(5), 307.

Bowler, S., & Karp, J. A. (2004). Politicians, scandals, and trust in government. *Political Behavior*, *26*(3), 271–287.

Burger, J. M. (2009). Replicating milgram: Would people still obey today? *American Psychologist*, *64*(1), 1.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, *14*(5), 365–376.

Byman, D. L., & Pollack, K. M. (2001). Let us now praise great men: Bringing the statesman back in. *International Security*, *25*(4), 107–146.

Campbell, W. K., Goodie, A. S., & Foster, J. D. (2004). Narcissism, confidence, and risk attitude. *Journal of behavioral decision making*, *17*(4), 297–311.

Campbell, W. K., Rudich, E. A., & Sedikides, C. (2002). Narcissism, self-esteem, and the positivity of self-views: Two portraits of self-love. *Personality and Social Psychology Bulletin*, *28*(3), 358–368.

Chang, J., & Halliday, J. (2007). *Mao: The unknown story*. Macmillan, Seize the Hour.

Cislak, A., Cichocka, A., Wojcik, A. D., & Frankowska, N. (2018). Power corrupts, but control does not: What stands behind the effects of holding high positions. *Personality and Social Psychology Bulletin*, *44*(6), 944–957.

Colgan, J. D. (2013). Domestic revolutionary leaders and international conflict. *World Politics*, *65*(4), 656–690.

Cooke, D. J., & Michie, C. (2001). Refining the construct of psychopathy: Towards a hierarchical model. *Psychological assessment*, *13*(2), 171.

Dallago, F., Cima, R., Roccato, M., Ricolfi, L., & Mirisola, A. (2008). The correlation between right-wing authoritarianism and social dominance orientation: The moderating effects of political and religious identity. *Basic and applied social psychology*, *30*(4), 362–368.

Deluga, R. J. (2001). American presidential machiavellianism: Implications for charismatic leadership and rated performance. *The Leadership Quarterly*, *12*(3), 339–363.

Dikötter, F. (2010). *Mao's great famine: The history of china's most devastating catastrophe, 1958-1962*. Bloomsbury Publishing USA.

Dikötter, F. (2016). *The cultural revolution: A people's history, 1962—1976*. Bloomsbury Publishing USA.

Dikötter, F. (2020). *Dictators: The cult of personality in the twentieth century*. Bloomsbury Publishing Plc.

Duspara, B., & Greitemeyer, T. (2017). The impact of dark tetrad traits on political orientation and extremism: an analysis in the course of a presidential election. *Heliyon*, *3*(10), e00425.

Endler, N. S., & Magnusson, D. (1976). Toward an interactional psychology of personality. *Psychological bulletin*, *83*(5), 956.

Fox, B., & DeLisi, M. (2019). Psychopathic killers: a meta-analytic review of the psychopathy-homicide nexus. *Aggression and violent behavior*, *44*, 67–79.

Gallagher, M. E., & Allen, S. H. (2014). Presidential personality: Not just a nuisance. *Foreign Policy Analysis*, *10*(1), 1–21.

Gao, Y., & Raine, A. (2010). Successful and unsuccessful psychopaths: A neurobiological model. *Behavioral sciences & the law*, *28*(2), 194–210.

Glad, B. (2002). Why tyrants go too far: Malignant narcissism and absolute power. *Political Psychology*, *23*(1), 1–2.

Gluck, M., Heesacker, M., & Choi, H. D. (2020). How much of the dark triad is accounted for by sexism?[U+2730]. *Personality and Individual Differences*, *154*, 109728.

Gøtzsche-Astrup, O. (2019). Partisanship and violent intentions in the united states.

Hare, R. D., Harpur, T. J., Hakstian, A. R., Forth, A. E., Hart, S. D., & Newman, J. P. (1990). The revised psychopathy checklist: reliability and factor structure. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*(3), 338.

Haslam, N., Holland, E., & Kuppens, P. (2012). Categories versus dimensions in personality and psychopathology: a quantitative review of taxometric research. *Psychological medicine*, *42*(5), 903–920.

Hershman, D. J., & Lieb, J. (1994). *A brotherhood of tyrants: Manic depression and absolute power*. Amherst, N.Y.: Prometheus Books.

Hollander, P. (2016). *From benito mussolini to hugo chavez*. Cambridge University Press.

Hollander, P. (2017). *Political pilgrims: Western intellectuals in search of the good society*. Routledge.

Hosker-Field, A. M., Molnar, D. S., & Book, A. S. (2016). Psychopathy and risk taking: Examining the role of risk perception. *Personality and Individual Differences*, *91*, 123–132.

Jain, K., & Bearden, J. N. (2011). Machiavellianism and overconfidence.

Johnson, A. M., Vernon, P. A., & Feiler, A. R. (2008). Behavioral genetic studies of personality: An introduction and review of the results of 50+ years of research. *The SAGE handbook of personality theory and assessment*, *1*, 145–173.

Jonason, P. K., Duineveld, J. J., & Middleton, J. P. (2015). Pathology, pseudopathology, and the dark triad of personality. *Personality and Individual Differences*, *78*, 43–47.

Jonason, P. K., Slomski, S., & Partyka, J. (2012). The dark triad at work: How toxic employees get their way. *Personality and individual differences*, *52*(3), 449–453.

Jones, B. F., & Olken, B. A. (2005). Do leaders matter? national leadership and growth since world war ii. *The Quarterly Journal of Economics*, *120*(3), 835–864.

Jones, D. N. (2013). Psychopathy and machiavellianism predict differences in racially motivated attitudes and their affiliations. *Journal of Applied Social Psychology*, *43*, E367–E378, doi:10.1111/jasp.12035.

Jones, D. N., & Figueredo, A. J. (2013). The core of darkness: Uncovering the heart of the dark triad. *European Journal of Personality*, *27*(6), 521–531.

Kajonius, P. J., Persson, B. N., & Jonason, P. K. (2015). Hedonism, achievement, and power: Universal values that characterize the dark triad. *Personality and Individual Differences*, *77*, 173–178.

Kline, P., & Cooper, C. (1984). A construct validation of the objective-analytic test battery (oatb). *Personality and Individual Differences*, *5*(3), 323–337.

Knopik, V. S., Neiderhiser, J. M., DeFries, J. C., & Plomin, R. (2018). *Behavioral genetics* (7th ed.). Macmillan Learning. New York.

Koehn, M. A., Jonason, P. K., & Davis, M. D. (2019). A person-centered view of prejudice: The big five, dark triad, and prejudice. *Personality and Individual Differences*, *139*, 313–316.

Larson, D. W. (1988). Problems of content analysis in foreign-policy research: notes from the study of the origins of cold war belief systems. *International Studies Quarterly*, *32*(2), 241–255.

Lee, K., Ashton, M. C., Wiltshire, J., Bourdage, J. S., Visser, B. A., & Gallucci, A. (2013). Sex, power, and money: Prediction from the dark triad and honesty–humility. *European Journal of Personality*, *27*(2), 169–184.

Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of personality and social psychology*, *68*(1), 151.

Lilienfeld, S. O., & Andrews, B. P. (1996). Development and preliminary validation of a self-report measure of psychopathic personality traits in noncriminal population. *Journal of personality assessment*, *66*(3), 488–524.

Lilienfeld, S. O., Latzman, R. D., Watts, A. L., Smith, S. F., & Dutton, K. (2014). Correlates of psychopathic personality traits in everyday life: Results

from a large community survey. *Frontiers in psychology*, *5*, 740.

Lilienfeld, S. O., Waldman, I. D., Landfield, K., Watts, A. L., Rubenzer, S., & Faschingbauer, T. R. (2012). Fearless dominance and the us presidency: Implications of psychopathic personality traits for successful and unsuccessful political leadership. *Journal of personality and social psychology*, *103*(3), 489.

Lilienfeld, S. O., Watts, A. L., & Smith, S. F. (2015). Successful psychopathy: A scientific status report. *Current Directions in Psychological Science*, *24*(4), 298–303.

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/

Macenczak, L. A., Campbell, S., Henley, A. B., & Campbell, W. K. (2016). Direct and interactive effects of narcissism and power on overconfidence. *Personality and Individual Differences*, *91*, 113–122.

Marcus, D. K., Preszler, J., & Zeigler-Hill, V. (2018). A network of dark personality traits: What lies at the heart of darkness? *Journal of Research in Personality*, *73*, 56–62.

Matthews, G., Reinerman-Jones, L. E., Burke, C. S., Teo, G. W., & Scribner, D. R. (2018). Nationalism, personality, and decision-making. *Personality and Individual Differences*, *127*, 89–100.

McDonald, M. M., Donnellan, M. B., & Navarrete, C. D. (2012). A life history approach to understanding the dark triad. *Personality and Individual Differences*, *52*(5), 601–605.

Međedović, J., & Knežević, G. (2018). Dark and peculiar. *Journal of Individual Differences*.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of abnormal and social psychology*, *67*(4), 371.

Moshagen, M., Hilbig, B. E., & Zettler, I. (2018). The dark core of personality. *Psychological review*, *125*(5), 656.

Mullins-Sweatt, S. N., Glover, N. G., Derefinko, K. J., Miller, J. D., & Widiger, T. A. (2010). The search for the successful psychopath. *Journal of Research in Personality*, *44*(4), 554–558.

Nai, A. (2019). Disagreeable narcissists, extroverted psychopaths, and elections: A new dataset to measure the personality of candidates worldwide. *European Political Science*, *18*(2), 309–334.

Nai, A., & Toros, E. (2020). The peculiar personality of strongmen: comparing the big five and dark triad traits of autocrats and non-autocrats. *Political Research Exchange*, *2*(1), 1–24.

Nell, V. (2006). Cruelty's rewards: The gratifications of perpetrators and spectators. *Behavioral and Brain Sciences*, *29*(3), 211–224.

Obradović, J., Pardini, D. A., Long, J. D., & Loeber, R. (2007). Measuring interpersonal callousness in boys from childhood to adolescence: An examination of longitudinal invariance and temporal stability. *Journal of Clinical Child and Adolescent Psychology*, *36*(3), 276–292.

O'Connell, D., & Marcus, D. K. (2016). Psychopathic personality traits predict positive attitudes toward sexually predatory behaviors in college men and women. *Personality and Individual Differences*, *94*, 372–376.

Paulhus, D. L. (2014). Toward a taxonomy of dark personalities. *Current Directions in Psychological Science*, *23*(6), 421–426.

Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of research in personality*, *36*(6), 556–563.

Perina, K., Hsu, S., & Washington, C. (2020). *Manifold podcast.* Retrieved from https://manifoldlearning.com/episode-036/

Plomin, R. (2019). *Blueprint: How dna makes us who we are.* Mit Press.

Plouffe, R. A., Saklofske, D. H., & Smith, M. M. (2017). The assessment of sadistic personality: Preliminary psychometric evidence for a new measure. *Personality and individual differences*, *104*, 166–171.

Post, J. M. (2003). Assessing leaders at a distance: The political personality profile. *The psychological assessment of political leaders: With profiles of Saddam Hussein and Bill Clinton*, 69–104.

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of personality and social psychology*, *67*(4), 741.

Preston, O. C., & Anestis, J. C. (2018). Psychopathic traits and politics: Examining affiliation, support of political issues, and the role of empathy. *Personality and individual differences*, *131*, 142–148.

Ronson, J. (2012). *The psychopath test: A journey through the madness industry*. Riverhead Books.

Southard, A. C., & Zeigler-Hill, V. (2016). The dark triad traits and fame interest: Do dark personalities desire stardom? *Current Psychology*, *35*(2), 255–267.

Sparrow, R. (2014). In vitro eugenics. *Journal of Medical Ethics*, *40*(11), 725-–731. Retrieved from https://doi.org/10.1136/medethics-2012-101200

Taylor, S. (2019). *Pathocracy. when people with personality disorders gain power.* Psychology Today. Retrieved 2019-07-31, from https://www.psychologytoday.com/us/blog/out-the-darkness/201907/pathocracy

Toland, J. (1976). *Adolf hitler*. New York: Doubleday.

Ullrich, S., Borkenau, P., & Marneros, A. (2001). Personality disorders in offenders: Categorical versus dimensional approaches. *Journal of Personality Disorders*, *15*(5), 442–449.

Vernon, P. A., Villani, V. C., Vickers, L. C., & Harris, J. A. (2008). A behavioral genetic investigation of the dark triad and the big 5. *Personality and individual Differences*, *44*(2), 445–452.

Visser, B. A., Pozzebon, J. A., & Reina-Tamayo, A. M. (2014). Status-driven risk taking: Another "dark" personality? *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *46*(4), 485.

Vukasović, T., & Bratko, D. (2015). Heritability of personality: a meta-analysis of behavior genetic studies. *Psychological bulletin*, *141*(4), 769.

Yergin, D. (1977). Shattered peace: The origins of the cold war and the national security state.