The Psychology of Existential Risk: Cognitive Constraints to Evaluation of Human Extinction

Rapportens samlede antal tegn : 125234

Svarende til antal normalsider á 2400 tegn: 52,2

Morten Ilseng Risnes, Studienummer: 20166174

Specialevejleder: Laura Petrini

10. Semester

Kandidatspeciale

Aalborg Universitet

18. maj 2022

Abstract

Background: The development and application of new technologies is likely to introduce new risks to human existence. A moral case can be made that existential risk reduction is more important than any other global public good. Unlike all other types of catastrophic scenarios, human extinction would be a permanent end to mankind, meaning that a vast number of potential human lives would be permanently denied the possibility of existence. Despite this, there seems to be limited interest in the topic of existential risk among lay people and in public policy. It has been found that people do not consider human extinction to be uniquely bad, and that their judgment may be influenced by focus on immediate suffering and death rather than reflection on the greater consequences of extinction.

Objectives: The aim of this thesis is to explore the possibility that human cognition may be poor for accurate perception and assessment of existential risk and scenarios involving premature extinction of mankind. This is a theoretical thesis with the objective of providing insights into the ways in which judgment and decision making potentially is made regarding the concept of premature human extinction. This is done with the following thesis statement: *To what extent are we as humans able to understand and relate to existential risks? What cognitive limitations may restrict our ability to accurately assess risks of premature extinction?*

Resources: The thesis statement is answered mainly with a basis in literature from the heuristics and biases program of cognitive psychology. Other viewpoints and critique are considered.

Results and discussion: Several of the cognitive principles from the heuristics and biases program were found to be relevant to better understand people's assessment of existential risk. The heuristic principles of availability, representativeness, anchoring, and affect were discussed in relation to representation of existential catastrophe scenarios and risk perception. It is suggested that existential risk as an object to the mind is especially difficult to represent and make appropriate judgments on, due to its abstractness and uniqueness. The fact that we cannot learn from personal experience with human extinction was considered one of the reasons for why sound judgment may be difficult, in addition to other aspects related to accessibility of representations for objects in the mind. The theory of attribute substitution was considered as an explanation for the mechanisms underlying heuristic thinking. Judgment and decision making was further explored in light of Stanovich' tripartite model of

the mind, emphasizing that intuitive goals may be in conflict with goals at the intentional level. Thoughts and ideas for future research on the topic was presented.

Table of Contents

1. Introduction

Be it war, climate change or a pandemic, concern and fear for big catastrophes seem to be ever present in public discourse. Great measures have been taken to emphasise the devastating consequences of global warming if we do not get our act together. Similarly, the COVID-19 pandemic sparked conflict over peoples' assessments of impact and severity, where lack of concern for the virus was regarded as a serious matter. Both issues rightfully deserve attention and respect and have inspired movements that unite people from all walks of life. Looking at these examples, it would seem that we place a high value on human life and health also when dealing with distant and future lives. Collectively, we strive to prevent and reduce the impact of catastrophes, and from a cost-benefit standpoint one should expect that greater amount of potential suffering will evoke more attention and engagement in people. However, our willingness to help does not easily translate into optimal actions for reducing suffering. This is for example revealed by the choices we make when we give to charities: while people in the USA make donations amounting to 2% of gross domestic product (GDP) each year, very little of this money end up in the programs that most effectively save or improve lives (Caviola, Schubert & Greene, 2021, p. 596). Instead, many donors view acts of charity as a matter of personal preference, favouring the emotional appeal of a cause over its effectiveness, and are prone to fail in giving effectively even when they wish to do so (Caviola, Schubert & Greene, 2021).

I will argue that our view on potential catastrophes tends to be clouded and inaccurate as well. By *catastrophe,* I mean events that cause huge amounts of suffering and/or death to human lives, and so reduces the total amount of happiness. Based on this we can say that the effects a catastrophe has on future potential for human flourishing should be considered when deciding its severity. The harms of climate change become more evident when one imagines the outcomes that unchecked global warming will have on all future life on Earth. Likewise, unwarranted underpopulation can be seen as a tragedy since it equates to less joy and happiness ever to be experienced in our universe. But in my view neither climate change nor a fertility crisis stands out as the worst of potential catastrophes. For while their harm may be great, such events likely do not put an end to the potential for human flourishing. Any event that allows for civilisation to persist are not a threat to the existence of humanity, they are not an existential catastrophe.

Philosopher Nick Bostrom coined the term *existential risk* and defines it as “one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction for desirable future development” (Bostrom, 2013, p. 15). There are many directions one might go when dealing with the concept of existential risk. What are the existential risks we face today? How likely is it that we will face premature extinction in this century? To what extent are existential risks preventable? And why are we obligated to treat existential risks differently than other kinds of catastrophic risks? In the following sections I will attempt to cover these topics only to set the stage for what I consider to be the most interesting question and subject for this thesis:

To what extent are we as humans able to understand and relate to existential risks? Or in other words, what cognitive limitations may restrict our ability to accurately assess risks of premature extinction?

In answering this question, my aim is to present leading theories on the psychology of reasoning and problem solving to discuss how humans normally succeed, but sometimes fail, to construct a worldview that is useful to reach their intrinsic goals. While I primarily will deal with findings from the heuristics and biases program from the field of cognitive psychology, the question also warrants exploring individuals’ handling of existential risk as seen from a evolutionary perspective.

The most visible thing that makes humanity stand out from the rest of the animal kingdom is our use of technology. Our unique ability to manipulate our environment pulled us out of our Stone Age lifestyle and forced us to adapt and deal with novel challenges never seen before. If humanity is to go extinct soon it seems more and more likely that it will happen by our own hands. And as we stand at the crossroads between progression and regression, it is tempting to think that the way of technology is nothing but a dead end. If we however consider our potential for reflective reasoning as a humanmade tool that can be tended and improved, there is hope that we will keep on top of our self-made risks.

Background

At the onset of working on this thesis, a search for literature that touched on topics related to the subject was made with limited results. Central figures that specifically have discussed psychological challenges to reasoning about existential risk are Eliezer Yudkowsky and Nick

Bostrom, and their contributions will be presented and discussed where relevant. One set of studies on the psychology of existential risk was found, for which findings served as a major inspiration for further exploration of the possibility that our intuition and reasoning abilities may be somewhat inadequate to assessment of existential risk scenarios. Schubert and colleagues (2020) performed studies on laypeople's reasoning about human extinction, finding that people agree such scenarios should be prevented. But of much more interest, they discovered that people tend to think that an existential catastrophe would not be uniquely bad compared to other near-extinction catastrophes that would not result in actual human extinction. The studies consisted of questionnaires, and to test for whether they would consider extinction uniquely bad, participants were asked to rank the severity of outcomes involving no catastrophe, a catastrophe killing 80% of the population, and an existential catastrophe killing 100% of the population. They were then asked to judge whether the difference between the outcome of no deaths and 80% deaths were lower or greater than the difference between 80% and extinction, finding that only a small minority (23%) of the participants judged the second difference to be greater than the first difference (Schubert et al., 2020, p. 3). Thus, extinction was not recognised as uniquely bad by the participants. Further studies revealed that participants' focus on humans' immediate suffering and death may explain why they fail to recognise human extinction as something uniquely bad when compared to other massive, but non-existential, catastrophes. From this, it is suggested that if people were to reflect more carefully on such scenarios, they might tend more towards recognising that human extinction is a uniquely bad event (Schubert et al., 2020, p. 5). These findings raise questions about whether our intuition might impede or disturb our abilities to make sound judgments about existential risk. But before this is explored, it makes sense to establish why existential risk should be considered a unique kind of risk unlike any other catastrophic threats.

2. Theory

2.1. On existential risk

The idea of world-ending events is neither new, nor unfamiliar to mankind, as evidenced by the range of terms we have created to describe such a scenario. "The Devil is known by many names", and a brief web search for synonyms to *Doomsday* gives us terms like *Apocalypse*, *Armageddon, End times, Judgement Day* and *Ragnarök*. Interestingly, these names stem from the eschatology of world religions, the part of theology concerned with death, judgement, Heaven and Hell (Oxford English Dictionary, 2022). In these conceptions of end times lie a sense of meaning and destiny; in Christian doctrine Judgement Day is the second coming of Christ, where those deemed worthy will be rewarded and live forever in an afterlife. If we set aside religious beliefs, end times could also be thought of as the time when humanity inevitably will succumb to entropy, for example from the hypothesized heat death of the universe. In both these perspectives "the end" is inevitable, but one can find comfort in the belief of a greater afterlife or the fact that "The Big Freeze" is scheduled to take place hundreds of billions of years into the future. A premature extinction of humanity distinguishes itself from these other kinds of doomsday scenarios for being void of meaning and for it being avoidable, as implied by the word premature. From the fact that humanity has survived for hundreds of thousands of years it has been pointed out that it is extremely unlikely we will meet an end over the next century due to natural catastrophes such as asteroid impacts, earthquakes, volcanic eruptions, or other natural existential risks that have been present throughout history (Bostrom, 2013, p. 16). In contrast, there is no track record for novel dangers that come about due to development of new technology, so-called *anthropogenic existential risks.*

2.1.2. Anthropogenic risk

Here, the main concern is that our potential to make technological breakthroughs in the future at some point will radically enhance our ability to manipulate our environment or biology and thus introduce existential threats never seen before (Bostrom, 2013, p. 16). A familiar example of this would be the discovery of nuclear fission, the splitting of atoms, which revolutionized energy production but also gave us extreme tools of mass destruction in the form of the atom bomb. Greater dangers might similarly arise from progress in biotechnology, where gain-of-function research could trigger dangerous pandemics or result in creation of advanced biochemical weaponry (Selgelid, 2016). In recent years, more attention has been directed towards exploring the implications of Artificial General Intelligence (AGI). AGI is not something that exists yet, but it has been defined as artificial intelligence that is capable of solving most, if not all, tasks that humans can solve (Shevlin, Crosby & Halina, 2019, p. 1). The creation of AGI is expected to make a fundamental impact on society (Shevlin, Crosby & Halina, 2019, p. 1), and while such technology might seem to belong to the far future, surveys on AI experts reveal that the majority expect AGI to be created during the next 100 years (Baum, Goertzel & Goertzel, 2011; Müller & Bostrom, 2014; Grace et al., 2017), the most recent survey showing 45% predicting a date before 2060, 34% a date after 2060 and 21% that it will never happen (Perry, 2019). Even if the intentions behind such technology might good, current discussions in AI safety work, which deals with avoiding creation of malevolent AI, have turned pessimistic: one of its most prominent thinkers, Eliezer Yudkowsky of the Machine Intelligence Research Institute has declared that the prospect of humanity surviving AGI look "incredibly grim" and that we can only hope for a miracle to change the course (Bensinger & Yudkowsky, 2021).

A less gloomy, but nevertheless grave estimation on humanity's chance of survival over the next years is presented in Toby Ord's *The Precipice: Existential Risk and the Future of Humanity*. Ord, a philosopher well known for his work in the effective altruism movement, gives a 1 in 10 chance of unaligned AI ending humanity within the next 100 years, and a 1 in 6 chance for the human species not surviving beyond the next 100 years due to an existential global catastrophe (Ord, 2020, p. 166). Regardless of the actual chance of an existential catastrophe occurring at any moment, there is an argument to be made on the way uncertainty about the assessment should be treated. Given the potential that a highly undesirable but poorly understood outcome may occur, it will make sense to be extremely wary of the outcome even in cases where an initial assessment deems the chance of the outcome to be very low. For outcomes of high consequence, uncertainty about the accuracy of an initial assessment should outweigh the probability as given in the assessment, as we have to factor in the probability that our assessment could be completely wrong and therefore have fatal consequences. The probability of the assessment being wrong could be much higher than the probability of the outcome as given by the assessment. As such, when dealing with existential risk characterized by novel and poorly understood scenarios (new technology) we may find that most of the risk for a particular event resides in the uncertainty of our initial assessment that the likelihood of the event was low (Bostrom, 2013, p. 16).

2.1.1 A premature extinction

Going further on the distinctive treatment one should give to high consequence outcomes, there is a strong argument for treating existential catastrophes as something categorically different to all other non-extinction catastrophes. For unlike any other tragic event, premature extinction of the human species would not only mean the end of all life existing at that point in time, but also the end for all potential human lives to come. At face-value this already seems bad because if it were to happen there would be no going back: there is no “undo button”, no second chance to learn from our mistakes and carry on in a new direction. The implication of an existential catastrophe is the closing of all doors for what could be, and a tragic end to the story of humanity. Thinking further on the implications, a premature extinction would also mean that the total amount of human lives ever to exist would be reduced. There is no consensus on how much value should be attributed to human lives that do not exist but might do someday, but I would argue that people commonly ascribe *some* value to potential future lives rather than *none*. The current concern about the consequences of climate change could be taken as an example of caring about the wellbeing of future generations. Given that each future life holds some value, if ever so little, to most people today, the number of total potential lives at stake should influence our concern for catastrophic risks. And what exactly is at stake regarding existential risk? To put it in perspective, one can compare the number of all humans that have lived so far with how many people that could possibly come to existence in the future. Demographers estimate that about 109 billion people have lived and died over the past 200,000 years (an approximate cutoff when speaking of the human species). As of 2022, there are about 8 billion humans alive on Earth, meaning that those of us currently alive represent about 6,8% of all people to have ever lived (Roser, 2022, p. 1). Estimating the number of future lives is naturally much more difficult, and can be nothing more than a good guess, but it gives us an idea about what the future might be like. If humanity were to have the same lifespan as that of a typical mammalian species, which is 1 million years (Roser, 2022, p. 1), one conservative estimation suggests that approximately 100 trillion lives could exist over our “remaining” 800 000 years (Roser, 2022, p. 4). Much more optimistically, one can also imagine that humanity, being clearly distinct from all other life on Earth for its creative abilities, will find ways to live past the typical lifespan of other species. It is thought that our planet could be habitable for around a billion more years, resulting in the possibility of a 125 quadrillion more lives yet to be born. Having in mind that these calculations are no more than best guesses based on historic growth, describing our potential to expand in numbers helps to make a clearer image of what is at stake: from an omnipresent perspective, our time on Earth up to now would be a fragment of all that is to humanity. This places a tremendous responsibility on humans alive today to keep the possibilities for growth open, and not forever closing the door for future humanity by succumbing to premature extinction. Therefore, if we value future life, it is fair to say that humans today carry a heavy burden. Yet, surprisingly little of our resources are currently invested in management of existential risk. An approximate estimate of annual targeted spending in the world in US dollar from 2017 suggests that over $300 billion was spent on fighting climate change, $1-10 billion on nuclear security, $1 billion on extreme pandemic prevention and merely $10 million on AI safety research, while about $1,3 trillion went to luxury goods (Todd, 2017, p. 10). In hindsight, it's easy to say that spending on pandemic prevention perhaps was neglected, and likewise, one can argue that the other causes might be underfunded as well.

2.2. Reasoning and existential risk

So far, I have presented the main points underlying the importance of existential risk. There is reason to accept the possibility that an event leading to premature extinction could take place in the foreseeable future. An existential catastrophe is unlike any other kinds of catastrophes because of its permanent effect on the potential for human flourishing. Concern for catastrophes appeal to the value of future life and wellbeing. Of all concerns, existential risks could be seen as the most worrisome, as a massive amount of potential future lives are at stake. Despite this, the term “existential risk” does not seem to be widely used in public discourse or in any major political party programs. Why is this? One simple explanation is that the reasoning about existential risk could be flawed on some level, and the public's intuition on the topic trumps the experts' worldview. Another, more salient alternative is that we as humans have difficulties with generating an accurate concern for existential risk in order to treat it as categorically different from other high-consequence non-existential risks. Recall the research on effective altruism that I mentioned in the opening segment, showing that people trying to do good in the world fail to do so in an effective manner. It might seem cynical to point out ineffectiveness when intentions are noble, but this does not take away from the finding that humans are prone to commit arguably irrational misjudgments about the impact of their efforts. We are not dispassionate robots solely thinking in numbers and relative risk values. Nor are we animals with imagination and concern limited to the present. Given that existential risks currently are underestimated, it is fruitful to explore how humans typically make risk assessments and reason about problems in the world. Returning to the thesis problem of this paper, we should be asking ourselves what aspects of our cognition might influence and limit our proclivity to deal with the concept of existential risk in an appropriate way. In the following section, I will present theory and findings on the science of human reasoning and problem solving to open for discussion on assessment of existential risk.

2.2.1. The psychology of reasoning

Cognitive psychology is the branch of psychology devoted to the operation of mental processes, exploring concepts such as perception, attention, memory, decision making and other facets of thinking. At the time of its conception in the 1940s to -50s it deviated from the contemporary approach of behaviorism for its attempts at understanding inner workings mediating our actions, rather than just concerning itself with observable behavior, and has since yielded constructs and frameworks that have been adopted by the other branches of psychology (Oxford Dictionary of Psychology, 2015).

2.2.2. Heuristics and biases program

Today, modern perspectives in cognitive psychology commonly deal with mental processes in the framework of a dual process theory, as expounded upon by Kahneman and Tversky, and popularized by their book Thinking, Fast and Slow. Psychologist and economist Daniel Kahneman together with Amos Tversky, also a psychologist, became known for their prospect theory, the finding that people typically deal with prospects of financial gain and loss in an asymmetric manner, and therefore are not the perfectly rational agents implied by mainstream economic theory at the time. This work was awarded with the Nobel prize in economics and set the stage for deeper exploration into the ways human reasoning and decision-making deviate from rational behavior.

2.3. A dual process theory of the mind

Based on empirical research, Kahneman and Tversky came to believe that our thinking abilities can be separated into two generic modes of cognitive function: one intuitive type of thinking characterized by automatic and rapid judgments and decisions, and the other a controlled mode which is slower and deliberate (Kahneman, 2003, p. 697). This first mode of cognition was inspired by an idea that there was something between the automatic operations of perception and the deliberate operations of reasoning; a mode for intuitive judgments with output that could differ from judgements made by the individual if he were to think about the matter thoroughly. Studying statistically sophisticated researchers, they had found that the participants, despite being experts on the domain, were prone to make systematic errors on causal statistical tasks. When asked to give intuitive judgements, their statistical inferences and estimates of statistical power would be way off target, showing that the experts clearly lacked a sensitivity to the effects of sample size. In other words, there were a persistent discrepancy between statistical knowledge and statistical intuition. This finding helped establish a dual process theory of the mind and led Kahneman and Tversky on a mission to uncover instances where our intuition fails us and what these kinds of error in judgment might tell us about cognitive mechanisms (Kahneman, 2003, p. 697f).

When we talk about judgements in daily life, we commonly think of it as intuition when complex judgments and preferences are made almost instantly and effortlessly. Ask me if I think the next Batman film will be any good, and I might come up with an immediate answer starting with “I don’t know, but my gut feeling is that…”. Ask me about how much profit the film will make, and I might again come up with a quick “best guess” that I think will be somewhat close to the real number, even if I have not consciously thought through any of the variables affecting box office profit. Several thinkers have recognized that there is a clear difference between intuitive judgments such as these, and the effortful cognitive processing that I would rely on if I, for example, were to try and calculate the film’s box office profit. This two-systems view has evolved into models that agree on the characteristics that distinguishes the two types of cognitive processes from each other. The intuitive type of thinking, commonly referred to as *System 1*, is typically described as fast, automatic, parallel, effortless, associative, implicit, and often emotionally charged. Additionally, System 1 is thought to be governed by habit, making these judgements and preferences difficult to control and change. In contrast, the *System 2* of cognitive processing relates to the operations that take place when we reason through judgment. These operations are widely recognized as being slower, serial, effortful, something that we are more likely to be consciously aware of, and under our deliberate control as we try to reach an accurate judgement. A simple way of deciding whether a specific mental process should be labelled System 1 or System 2, is to try and decide how much effort it takes to perform. More specifically, looking at det effect of concurrent cognitive tasks can give use insight because overall capacity for mental effort is limited, making it so that effortful processes will tend to disrupt each other, while effortless processes can be performed simultaneously with inconsequential impact.

Unlike perception, which is tasked with processing of immediate stimuli such as what we see in front of us and nothing else, both System 1 and System 2 processing are thought to be involved when we deal with conceptual representations as well as percepts, content relating to the past or the future and that which can be evoked by language. In Kahneman's model, both the perceptual system and the intuitive operations of System 1 are responsible for generating all impressions we have of the attributes of objects available to perception and thought. These impressions, which are automatic and not verbally explicit are however not to be confused with any judgments we make, judgments always being intentional and explicit formation of opinion on the concept in mind, even if we do not overtly express them. This means that the deliberate, thoughtful processes of System 2 necessarily plays a role in the making of judgments, regardless of whether the judgment originates in impressions or in deliberate reasoning. What *intuitive judgment* refers to then, is the kind of judgments that directly reflect the impressions their based on, untainted by any input from System 2 (Kahneman, 2003, p. 699). In this view, System 2 is functioning as a monitoring mechanism, in the sense that mental operations and behaviour may be inhibited or modified depending on their quality. This quality control could for example entail overriding the intuitive idea that one should walk on green light because sirens can be heard further up the street. Going back to the case of concurrent tasks revealing whether certain operations are effortful or effortless, Kahneman suggests that we can approach the phenomenon with this hypothesis simplified in the phrase "System 2 monitors the activities of System 1" (Kahneman, 2003, p. 699). The capacity to monitor impressions and modify judgments is limited, and this would explain how tasks for which intuition is sufficient to succeed can be performed simultaneously, while tasks that require attentive self-monitoring of thoughts easily can be interrupted and sabotaged if other cognitively challenging tasks are introduced. For the most part we rely on System 1 mental processing to do our daily tasks, such as knowing when to cross street walks, whereas the resource-demanding operations of System 2 is reserved for those special cases where intuition falls short. If I was caught up in a heated discussion as I were to cross the street-walk, my limited capacity to self-monitor thoughts might not be available to override the intuitive judgment that the street is in fact safe to cross. Kahneman and colleagues suggest that the monitoring of System 2 actually is rather passive under normal circumstances, allowing many intuitive judgments to be expressed freely. This is for example seen in a classical study on simple puzzles, where the participant asked to decide the value of a ball after being told that a bat and a ball together cost $1.10 and that the bat costs $1 more than the ball. I see the numbers and feel inclined to believe that the bat costs 1$ and the ball 10 cents because these round numbers easily match the total price. If only I had relied on hard calculation instead of intuition, I would easily find that this answer cannot be correct. Many people fall for this, and somewhat surprisingly, higher intelligence does not seem to alleviate these shortcomings, as studies revealed that about half of the Princeton and Michigan university students participating also chose to give the intuitive, but wrong, answer (Kahneman, 2003, p.699). The bat-and-ball problem was integrated into what initially was a three-item test, named the *Cognitive Reflection Test* (CRT), with similar problems that are “easy” in the sense that the correct answer is obvious in hindsight but difficult to answer if erroneous intuitive judgments are not properly suppressed (Frederick, 2005, p. 27). Later studies have shown high variance in performance on CRT and similar test batteries (Bruin, Parker & Fischhoff, 2007; Del Missier, Mäntylä & Bruine de Bruin, 2010; Toplak et al., 2011), that performance on CRT is stable across time (Stagnaro, Pennycook & Rand, 2018), and that men tended to score higher, but both men and women were too optimistic about their performance (Ring, Neyse, David-Barett & Schmidt, 2016).

2.3.3. Accessibility, natural assessments, and framing effects

It seems that while our effortless, intuitive judgments often are sufficient for problem solving and decision making, they cannot be fully trusted, which is especially problematic in cases where they seem convincing enough to bypass the internal self-monitoring that we call System 2. What then, makes certain intuitive thoughts rise and shape our judgment and behaviour in the right situations? To answer this question, Kahneman modified the concept of *accessibility* from memory research (Tulving & Pearlstone, 1966) to describe how some mental operations are more available and easier to perform than others. More specifically, accessibility in this paradigm relates to the ease or difficulty for which different aspects and elements of a situation, distinct objects in a scene, and different attributes of the object itself may come to mind. When seeing a coffee cup, I easily get an impression of its size, and even an approximate idea of how much liquid it will contain. But if tasked with deciding how much of an area the cup would cover if it was shattered into tiny pieces, I would find it much more challenging to reach an answer on the spot. Denied use of a hammer, I could try to visualize the cylinder being flattened out on the table and add this part to the bottom of the cup. For this task, there is no perceptual impression I can rely on; instead, I have to consciously deconstruct the object in my mind's eye, as the answer is not perceptually accessible. Accessibility can therefore be understood as a continuous dimension with the most rapid, automatic, and effortless (typical System 1) processes in one end and the most slow, serial and effortful (typical System 2) processes in the other.

As the coffee cup example illustrates, the actual properties of the object of judgment determines how accessible it is. While this first example seems rather obvious, it highlights the idea that we have a set of predispositions that direct our understanding of the world. Certain attributes of an object are easier to recognise than others, they are more accessible. While physical properties of the object itself decides what attributes are more accessible, our relation to the object also plays a role. Similarity, causal propensity, surprisingness, affective valence, and mood are other attributes that are generated for everything we interact with in our daily life. These so-called *natural assessments* are automatically and effortlessly registered and processed, and so plays a part in what conceptions are more accessible to us (p. 701). Furthermore, the context in which something is presented determines what attributes are more accessible. Cues from the context will strengthen one interpretation of the object over the other, making it so that "I3" easily is recognized as a letter (B) or a number (13) depending on what other symbols surrounds it. There is little room for ambiguity, and multiple interpretations do not arise to our consciousness simultaneously; instead, our mind decides on the most accessible interpretation, explaining how we might see a rabbit or a duck in an ambiguous picture, but are unable to have a perception of both interpretations at the same time.

These kinds of preferences are not restricted to our view of physical objects, as the same effects also determine how we grapple with more abstract stimuli. For example, the way in which a statement is phrased will make it so that some judgments about the statement are more accessible to us than others. Kahneman and colleagues named this the *framing effect* after finding that peoples' choices can be manipulated by framing alternatives so that they are either perceived as a loss or a gain. As part of our natural assessments, we automatically label an outcome as “good” or “bad” depending on how it is presented to us, thereby creating a highly accessible judgment that outcomes which lead to a loss must be avoided since losses usually are bad thing. In practice, this leads patients and even physicians to go with the choice of radiation therapy over surgery if the outcome of surgery is presented as 90% chance of short-term survival versus 10% chance of immediate death (McNeil, Pauker, Sox & Tversky, 1982). Because the prospect of immediate death feels especially terrifying and bad, one becomes blind to the actual odds and what is at stake. Participants were therefore more likely to avoid surgery when death was emphasised but choose it when it was described as having a 90% survival rate. Returning to the coffee cup example, we can now view this as also being due to a framing effect. The basic principle behind framing is that we as humans have a strong preference to passively accept a statement, object, or anything else in the way that it is presented to us. And in the same way that we effortlessly can determine the size of the cup but struggle to determine the area it would cover if shattered, we are quick to deem an option as bad if there is sufficient emphasis on the loss it might incur, while it takes deliberate effort to process the actual probabilities and their meaning. The finding that loss aversion has a consistent and considerable effect on individuals’ economic choices in various circumstances evolved into the prospect theory for which Kahneman won the Nobel Prize in Economic Sciences, while their research on errors of judgment became what is called the heuristics and biases program of decision making and problem solving.

2.3.4. Heuristic principles

Kahneman and Tversky proposed that we have a limited set of heuristics that we use to reach judgments more efficiently on complex problems that otherwise would require conscious assessment of probabilities and prediction of values (Tversky & Kahneman, 1974, p. 1124). Decision-making in daily life is often complicated by the fact that we have limited access to information and nevertheless have to make decisions, uncertain about specifics of an event or exact numbers. To compensate for uncertainty, we typically rely on different *heuristic* principles, which are mental shortcuts involved in creating intuitive judgments of the world. As part of System 1 mental processes, heuristics help us to quickly make judgments in an effortless way, but as I have described earlier, intuitive judgments introduce the potential for error. While relying on heuristics often results in judgments that are good-enough, there are thus an assorted set of biases associated with each of them. Studying peoples’ assessment of probabilities and prediction of values, Kahneman and Tversky first uncovered three kinds of heuristics used for judgment and decision-making under uncertainty: judgment by representativeness, judgment by availability, and judgment by adjustment and anchoring.

People typically rely on the *representativeness heuristic* when tasked with deciding the relation between one thing and another, such as how likely it is that a described object belongs to a specific class of objects. In such cases we tend to arrive at an answer by deciding how representative object A is of object B, that is, how similar the object in mind is to the object class in question (Tversky & Kahneman, 1974, p. 1124). Deciding the extent of similarities, and thus representativeness, makes it so that we, for example, quickly can tell how likely it is that a given person has a specific occupation based on a description of that person's personality and interests. On the other hand, it has been shown that reliance on this heuristic comes at the cost of ignoring other information present, like the prior probability of outcomes or the sample size in question. In practice, this means that people tend to judge based on similarity to the reference class in situations where they instead should be calculating the actual probability of something being the case (Tversky & Kahneman, 1983, p. 296). This error has been named *the conjunction fallacy,* and it was famously displayed in experiment conditions where, tasked with ranking the probability that different statements about a fictious person Linda are correct, participants were likely to rank the idea that she was both a feminist AND a bank teller as more probable than the statement that she was a bank teller, presumably because her description seemed more representative of feminist than that of a bank teller (Tversky & Kahneman, 1983). Of course, this makes little sense since each factor is more likely by themselves than when combined.

The *availability heuristic* comes in play in situations where people are tasked with assessing the frequency of a class or the probability of an event. Again, instead of trying to calculate actual probabilities, we are quick to arrive at an answer through a mental shortcut. The ease of which instances or occurrences can be brought to mind is intuitively accepted as an indication of how likely they are to take place. If something is easy to recall, we feel that it must also be important. This reliance on availability in memory and imagination also introduces cognitive biases because some scenarios will be easier to imagine or recall simply because they have been heavily exposed in recent news or initially had a great emotional impact on us, explaining why terror attacks are so effective at creating fear in a population (Tversky & Kahneman, 1974, p. 1127f).

The final heuristic of the three that Kahneman and Tversky first proposed describes the way we use prior information as a reference point when making decisions. More specifically, we are likely to take an initial related assumption and tweak it in the direction we feel is appropriate as a fast and effortless way of reaching a good-enough belief. This is evident in situations where people are giving estimates and make use of initial values perhaps present in the problem formulation or available from partial computation, that are then adjusted to yield the final answer. Their answer is then heavily dependent on what initial starting point was chosen, it is "anchored" to the initial value, which is why this kind of mental shortcut has come to be called the *(adjustment and) anchoring heuristic*. Beliefs derived through the anchoring heuristic will be biased towards the starting point as people often fail to adjust their initial number or value enough. This even applies when the initial starting point is of minor usefulness, as shown in a study where participants were to estimate the number of countries in Africa after first being asked to decide whether they thought it was higher or lower than a number provided by a random number generator. Those receiving a high random number would consistently assume a higher number of countries than participants with a low random number, revealing just how dependent they were on the trivial starting point (Tversky & Kahneman, 1974, p. 1128).

A later contribution to the list of general-purpose heuristic principles, is the affect heuristic. Although initially neglected, the role of affect in judgment and decision making was later recognised and came to be regarded as an important development in the study of judgment heuristics (Kahneman, 2003, p. 710). The affect heuristic is the mental strategy in which emotions towards the object in mind direct judgment. More specifically, it is thought that representations of objects and events are tagged to varying degrees with affect, and that people come to refer to an "affect pool" that contains all these positive and negative tags associated with the representation when they later make judgments and decisions relating to the object (Slovic, 2007, p. 1335). Affect may here serve as a cue to the individual about the object's importance and a more efficient way to quickly make a judgment of something as good or bad and can therefore be thought of as a mental shortcut, or heuristic, in judgment and decision making (Slovic, 2007, p. 1336). Affect has been shown to direct our judgment of other people, where persons accused of academic misconduct were perceived as deserving of less judgment if they were pictured as smiling, compared to non-smiling depictions (LaFrance & Hecht, 1995). Judgment of more abstract objects has also been shown to activate the affect heuristic. Participants that were presented with Chinese characters and their meaning and later asked to choose the character they preferred the most, chose characters with a positive meaning 70% of the time (Sherman et al., 2002). Interestingly, risk perception was found to be strongly associated with the degree to which a hazard evoked feelings of dread. Judgments of risk and benefit was also found to be negatively correlated: if a hazard is regarded as beneficial, it is perceived as having a lower risk, while hazards with less perceived benefits are perceived as more risky (Slovic, 2007, p. 1335). People tend to base their judgments of an activity or a technology on their feelings about it as well as their thoughts, meaning that the kind of affect they have for the object will influence both perceived benefit and risk. This goes both ways, as an experiment on the effect of information about technology being less risky was shown to improve overall positive affect for the technology, and in turn increase its perceived benefit (Slovic, 2007, p. 1343).

2.3.5. Attribute substitution

Taken together, these heuristics give insight into various circumstances where we quickly arrive at (often) useful beliefs from interpreting information we have at hand, and why this inevitably leaves us vulnerable to biased thinking and erroneous judgment. Since first introducing the heuristics and biases approach, Kahneman in collaboration with Shane Frederick, further developed an underlying model explaining how judgment heuristics actually work, how they are not restricted to only judgments under uncertainty, and how they relate to the interplay between mental operations of System 1 and System 2. They proposed that heuristic processes could be unified under a theory of *attribute substitution*: when judgments are reached based on a heuristic, it is because the individual has assessed a specified target attribute of the object in question by substituting it with a related heuristic attribute that comes more readily to mind (Kahneman, 2003, p. 707). Instead of applying cognitively demanding processes to make a computationally complex judgment, the individual unconsciously simplifies the problem by replacing some parts with more accessible attributes, thereby creating an essentially new question for which they effortlessly can offer a fitting answer. This corresponds well with the general finding that people’s intuitive judgments may sometimes be completely off base when their judgments from thoughtful reasoning are not. And while one could speculate that attribute substitution occurs because of a misunderstanding of the question, Kahneman argues it is more plausible that it happens because an evaluation of the heuristic attribute comes immediately to mind and that its associative relationship with the target attribute is sufficiently close to pass the permissive monitoring of System 2 (Kahneman, 2003, p. 709).

What exactly makes something a heuristic attribute, and why is it highly accessible? Reviewing the representativeness heuristic we can describe the phenomenon as the act of substituting probabilities with representativeness, such that individuals end up looking for similarities between the object and a reference class instead of thinking about how likely it is that an object A and object B fit together. In the Linda example, participants come to believe it is more likely that she is both a feminist and a bank teller because they intuitively change the question into being about how well her description fits a typical bank teller versus a typical feminist bank teller. In this case, judging from stereotypes is a highly accessible natural assessment, much more so than judgment from abstract probabilities (Kahneman, 2003, p. 709). The studies on the variations of the Linda problem and the representativeness heuristic provides strong evidence for the theory of attribute substitution because the same pattern in ranking of statements would show up for participant groups explicitly asked to rank for similarity and explicitly asked to rank for probability. As the list consisted of many miscellaneous statements that could be more or less probable and separately more or less similar to the Linda description, it was striking to see consensus between the groups. The "probability group" would not have ended up with the same rankings unless they also judged for representativeness like the "similarity group" did (Kahneman, 2003, p. 709).

What makes heuristic attributes highly accessible is their origin in System 1 processing; being standard and oft-used mental operations, they immediately pop into mind and so outcompete assessment of the target attribute. Going back to the interplay between System 1 and System 2, biased thinking is due the limited applicability of System 1 intuitive judgment, but the workings of System 2 is perhaps more worthy of blame. After all, isn't mental operations of System 2 supposed to kick in and override or modify faulty intuitive judgments? Kahneman asserts that faulty judgments normally are identified as biased and dealt with in the proper way by System 2, but only as long as corrective thoughts are sufficiently accessible to intervene in the judgment. Several factors that impair the corrective operations of System 2 have so far been uncovered, such as time pressure (Finucane, Alhakami, Slovic & Johnson, 2000) and concurrent engagement with another cognitive task (Gilbert, 2002), while things like higher intelligence (Stanovich & West, 2002) and exposure to statistical thinking (Agnoli, 1991) have some correlation with more successful System 2 operations. However, as revealed by research on the Linda problem, statistical training does little to correct intuitive heuristics outside of special circumstances, as statistically knowledgeable graduate students only managed to self-correct when given strong cues that they should use statistics, or the problem was phrased to their benefit; otherwise, they were just as vulnerable to exhibit biased judgment as everyone else. Trying to counteract faulty judgment by teaching about the potential biases seems to be just as ineffective, at times leading to such overcorrection that participants will underestimate frequency in problems where the availability heuristic is at play (Oppenheimer, 2004). Difficulties in performing reasonable judgment might be further explained by considering the effect of intuitive judgments being immediately accessible: even when the individual consciously tries to correct for bias, the intuitive judgment might act as an anchor for subsequent adjustments, resulting in the corrective adjustments to be too small to have a sufficient impact on the final judgment.

In conclusion, Kahneman suggests three necessary conditions for attribute substitution to take place. First, the target attribute needs to be relatively inaccessible, for example by requiring effortful mental operations due to its complexity or abstractive properties. Second, an associated attribute must be highly accessible, such as heuristic attributes that we are predisposed to look for because they provide useful rules-of-thumb. And finally, the substitution must not be sufficiently dealt with by the reflective System 2, as was discussed in the previous paragraph.

2.3. A tripartite model of the mind

While Kahneman and colleagues have fleshed out a useful framework to understand the interplay between mental process under decision-making and interpreting the world, their dual-process theory of the mind has also come under scrutiny for its limited description of the processes underlying our so-called reflective, effortful System 2. The psychologist Keith Stanovich became influential in the field of decision-making for his work on rationality and why the cognitive construct of intelligence is necessary, but distinct, from what he refers to as “the reflective mind”. First off, Stanovich clarifies that although the different operations of the mind can be classified as part of one system (fast) and another (slow), there really is no such thing as two distinct singular systems. Rather, “System 1” refers to a set of different systems that have automaticity in common with each other; they respond automatically to triggering stimuli independent of System 2 and may sometimes produce output that conflicts with judgments from System 2, but they are otherwise not necessarily related to each other functionally or structurally (Stanovich, 2008, p. 57). This is more in line with the view that intuitive processing comes from learned information and habit formation, as well as specific modules that are the results of evolutionary adaptation. In Stanovich' view it therefore makes much more sense to refer to System 1 as the autonomous set of systems, or TASS for short (Stanovich, 2008, p. 57). Likewise, we may benefit from treating System 2 as more than a single system.

2.3.1. Replicator and vehicle

In Stanovich' view, our mind can be regarded as a part of a machine constructed by our genes to promote their survival and reproduction. Our genes benefit when the body they are part of survive and thrive, but in special circumstances the interests of genes may collide with the interests of the person carrying them. Stanovich likes to think of this as a conflict between "replicators" (DNA) and their "vehicle" (the animal itself). Together with other animal species, we as humans possess a set of mental processes that operate on goals set by the genes, but unlike other animals we are also equipped with programming that allow for us, the vehicle, to set our own goals. In short, there are "short-leash" genetic control mechanisms that evolutionary adaptation has installed in the brain. They consist of general strategies and tricks that enhance survival and reproduction and are typical of System 1/ TASS. But additionally, a parallel control system has evolved, which is "long-leashed" in the sense that it is more flexible, able to conceive and deal with problems that are too novel and complex for a relevant strategy to have been programmed through adaptation (Stanovich, 2010, p. 12f). This analytic system, which the replicators only have indirect control over as opposed to their direct shaping of components in TASS, is to Stanovich what "System 2" essentially is. And as humanity gradually has modified its surroundings, we now find ourselves living in complex societies where most of the goals that the analytic system is trying to coordinate are derived goals. Estranged to the hunter-gatherer lifestyle of our ancestors, our basic goals and primary drives are now satisfied indirectly by maximising secondary goals such as employment, status, and prestige. (Stanovich, 2010, p. 67). Some instinctual responses have lost their use and may even be harmful, such as road rage. Since certain TASS-triggered responses must be suppressed in order to achieve many of these secondary goals, there is now a conflict between the short-leash TASS and our long-leashed analytic system. Road rage can for example be thought of as an aggressive instinct that allowing for a separation between the goals of evolutionary adaptation and the interests of the vehicle, that is the individual. An extreme case of this taking place would for example be the invention and use of contraceptives that allow us, the vehicle, to satisfy derived goals to the detriment of the replicator, our genes (Stanovich, 2010, p. 67).

2.3.2. A set of systems

Stanovich, who initially championed the dual-process theory, came to suggest that it might make more sense to speak of a "tripartite model" instead, as System 2 can be further separated into two different kinds of processes: the algorithmic mind and the reflective mind. Going back to the conflict between replicator and vehicle, Stanovich boils this down to a competition between TASS (System 1) and the analytic system (System 2). TASS will implement its short-leash goals unless mechanisms of the analytic system inhibit or modify to implement its long-leash goals. But, while generation of analytic judgments is necessary to correct biased and vehicle-threatening thinking, it is pointless if an override of the intuitive judgments never actually takes place. Thus, Stanovich distinguishes between *the algorithmic mind*, responsible for idea-generation and *the reflective mind*, which is the higher-order mechanism that initiates the override. In the following sections I will present the properties of the algorithmic mind, its interplay with the reflective mind and introduce three categories of cognitive failure based on Stanovich' taxonomy.

What is the rationale for separating the theoretical System 2 of reasoning into two new systems? Stanovich acknowledges that even if a line cannot be drawn as easily as that between autonomous, effortless operations and deliberate, effortful ones, there nevertheless is evidence for distinctive systems when one compares performance in tests for fluid intelligence to variation in rationality as measured by ability to correctly override and modify judgments, such as the cognitive reflection test that I mentioned earlier. For example, Stanovich and colleagues found some failures in critical thinking to be relatively independent of intelligence (Stanovich & West, 2007, 2008). A more recent study on the general relationship between rationality and intelligence found latent variables representing rationality and general intelligence to be strongly correlated, *r* = .54. So, while there is a clear relationship between the two, the correlation fell well short of unity, thereby indicating that the rationality and intelligence factors are empirically distinct from each other (Burgoyne et al., 2021).

2.3.3. Cognitive decoupling and the algorithmic mind

What fluid intelligence tests typically assess, is performance under optimal conditions, meaning that the person taking the test is explicitly informed of the winning strategy instead of initiating higher-level thinking on their own accord (Stanovich, 2008, p. 61). Rather than being a measure of executive processes, standard neuropsychological tests are then a measure of “supervisory processes”, since they assess the ability to carry out rules instantiated not by internal regulation, but by the tester who explicitly sets the rules and tells what maximal performance looks like (Stanovich, 2008, p. 67). Mental operations under such conditions primarily relate to Stanovich’ concept of the algorithmic mind. He posits that judgments made by the analytic System 2 come from a process of cognitive simulation. While we have a primary representation of the world that is supposed to directly map reality, we are also capable of generating secondary representations; similar to primary representations but decoupled from the world so that they can be manipulated and as such be mechanism for simulation. Going back to the coffee cup example, our immediate perception of the cup would be a primary representation and the idea of how much area a shattered cup would cover is the manipulated representation, constructed in our mental playground. There is a *cognitive decoupling* as a separate “possible world box” is created in which simulations can be done without contaminating the relationship between the world and the primary representation, meaning that we can do hypothetical thinking and simulating outcomes without distorting our view of reality (Stanovich, 2008, p. 62f). This operation, cognitive decoupling, is uniquely a function of what Stanovich calls the algorithmic mind and is absolutely necessary for cognitive simulation and hypothetical reasoning to happen.

Cognitive decoupling has three functions, namely override, comprehensive simulation, and interruption of serial associative cognition (Stanovich, 2008, p. 70). It is critical for override of judgments, as decoupling involves taking offline the connection between a primary representation and response programming. Furthermore, it is related to comprehensive simulation in the sense that decoupling makes it possible to have multiple models undergoing simultaneous evaluation and transformation without interfering with each other.

The third type of decoupling relates to a process where the analytic mind is engaged, but without hypothetical thinking and generation of new representations. This type of reasoning, called serial associative cognition, occurs when the individual reasons from the primary representation of a problem as it is presented to them, their starting point is a model of the world that is given to them. Serial associative cognition can explain bad performance on the *Wason selection task*, a test of deductive reasoning where participants are prone to incorrectly seek for confirmation of a rule rather than try to invalidate it (Stanovich, 2008, p. 68). In this case there is serial associative cognition with a focal bias, as the participant reasons from the rules presented in the task but does not think to imagine a scenario where the rule is false and what this would imply for testing the rule. In such instances, decoupling involves interrupting the process of serial associative cognition, meaning decoupling from the next step in an associative sequence that would otherwise direct thought (Stanovich, 2008, p. 70).

2.3.4. Initiator of decoupling: the reflective mind

But none of these types of cognitive decoupling can be initiated by themselves. Instead, it is the job of another system to signal decoupling to commence and begin the process of cognitive simulation or hypothetical reasoning. The system responsible for this process Stanovich refers to as the reflective mind. While having similar properties to that of the algorithmic mind, such as capacity-limited serial processing, the reflective mind is conceptualised as the system that enables the critical thinking required to avoid cognitive biases and heuristic fallacies. Thinking back on the set of studies indicating that variation in intelligence is insufficient to explain capability for rational thinking, the algorithmic mind may be seen as related to fluid intelligence and the reflective mind to represent rational thinking dispositions. In other words, the algorithmic mind relates to efficiency in computational power, while the reflective mind is concerned with the goals of the cognitive system, beliefs relevant to those goals, and the choice of action that is optimal given the system's goals and beliefs (Stanovich et al., 2020, p.1114). Stanovich emphasises the systems' roles in cognition by pointing to the clinical setting: while intellectual disabilities (i.e., mental retardation) can be regarded as a failure in the algorithmic mind since computational power is affected, many psychiatric disorders seem to indicate disruption of intentions and beliefs at the personal level, with symptoms such as delusional thinking implying there is an impairment in rationality (Stanovich, 2008, p. 59). This gives us a hint that intelligence is not the same as rationality, rational thinking having to do with forming well-calibrated beliefs and acting appropriately on these beliefs. These operations are the function of the reflective mind. The cognitive decoupling processes and processing power of the algorithmic mind is of course a prerequisite for rational thinking, but an incomplete system without the reflective mind as a higher-level overseer that can decide and implement what processes are to take place and to what extent they should affect judgment and action. If the algorithmic mind does cognitive decoupling, it is the reflective mind that initiates the decoupling mechanism, be it decoupling as an override of TASS, decoupling to sustain cognitive simulations or decoupling related to serial associative cognition (Stanovich, 2008, p. 70).

2.3.5. Cognitive failures of the tripartite model

The tripartite model of the mind is helpful to understand what intelligence tests miss and why intelligence is not the same as rationality. But it also provides a different view on how bad judgment can occur. Based on Stanovich' model one can define several categories of cognitive failure: TASS override failure, failure related to the mindware, and failure due to over-economizing. This first kind of failure is really the typical override failure that was recognized in the heuristics and biases program. The automated set of systems, or System 1, produces a response that due to higher accessibility fails to be overridden by the analytic System 2, and results in suboptimal judgments. A related but different sort of failure also results in bad judgment because intuitive responses are let through, but this time because the individual is not equipped with the proper mindware to perform the override, there is a *mindware gap*. "Mindware" is here taken to mean the rules, procedures, and strategies used by the analytic system to transform decoupled representations. Since override failure assumes that relevant mindware failed to be put in use, we should then expect that the risk of override failure increases as mindware gaps are reduced. It is further suggested that cognitive failure might also be caused by contaminated mindware, referring the possibility that we might also acquire unhelpful rules, procedures or strategies that cause us to behave irrationally, against our own goals. The final types of cognitive failure reflect our brains preference towards strategies that require the least amount of effort, "cognitive miserliness" as Stanovich puts it. This can be in form of our tendency to perform serial associative cognition with a focal bias, that is, instead of initiating a new simulation of alternative worlds from scratch or otherwise do completely decoupled reasoning, we base our reasoning on a problem as it is immediately presented to us. While effective, this makes us vulnerable to biases from for example the framing effect that was described earlier. Another case of cognitive failure from miserliness would simply be the tendency to not make use of System 2 reasoning at all, instead letting all autonomous processes and judgments go on without interference (Stanovich, 2008, p. 73).

2.4. Critique of the heuristics and biases program

While the heuristics and biases program and dual processes theory of the mind have fared well over the years, it has also gained its critics. Regarding the work by Kahneman and Tversky specifically, psychologist Gerd Gigerenzer supports the idea of heuristics as efficient shortcuts for effortless judgment and decision making under uncertainty but, challenges the idea that heuristic thinking creates irrational cognitive biases. His main argument is that errors and cognitive illusions that have been uncovered do not violate probability theory in the way Kahneman and Tversky claimed, and that they have neglected conceptual distinctions that are fundamental to probability and statistics. The finding of cognitive illusions, he argues, changes and vanishes as soon as these conceptual distinctions are taken into consideration (Gigerenzer, 1991, p. 86). To decide when biased thinking is taking place and participants' intuitive judgments are erroneous, they are typically asked to think in probabilities that then are compared to a given norm which, in Kahneman and Tversky's view, would be the one correct answer in accordance with an "accepted rule" of statistics (Gigerenzer, 1991, p. 86). But this is problematic, because according to the frequentist approach to statistics, there are no appropriate norms for a correct judgment of the single-case scenarios that the participants typically are assessed on. From the frequentist view it makes no sense to assign probability to a single event; probability theory is about frequencies, not about single events (Gigerenzer, 1991, p. 88). With this in mind, Gigerenzer argues that answers labelled as biases, such as the conjunction fallacy associated with the Linda problem, are not in violation of (frequentist) probability theory and therefore not evidence that we are prone to biased judgment. When participants provide frequencies instead of judging probability of a single event, such as by asking them how many out of 100 persons who fit the Linda description are bank tellers or bank tellers that are active in the feminist movement, the conjunction fallacy seems to largely disappear (Gigerenzer, 1991, p. 92). Similar effects from changes in test design were furthermore taken as evidence that neither overconfidence bias, nor base-rate neglect should be regarded as errors in probabilistic reasoning, as they seemingly are not violations of (frequentist) probability theory (Gigerenzer, 1991, p. 109). It would seem that many heuristics and biases tasks become more solvable when they are redesigned to fit certain stimulus constraints. The idea that human performance in probabilistic reasoning tasks is especially sensitive to the format in which information is presented and how one is to answer has been further explored, suggesting that humans have inductive-reasoning mechanisms that involve some rational principles, but that its design requires representations of event frequencies to properly function (Brase et al., 1998, p. 4). The argument goes that natural selection could not have led to development of cognitive mechanisms designed to reason about, or receive input in, a format that did not regularly exist. The probability of a single event cannot be observed by an individual, but an individual can observe the frequency of events over time, and thus we may be better adapted to intuitive judgments in frequency formats (Brase et al. 1998, p. 5).

It is thus implied that frequency formats make cognitive illusions disappear because of a natural predisposition to comprehending frequencies. Kahneman and Tversky have responded to this view by suggesting that these findings could be explained as an effect of the frequency formats providing extensional cues that help participant avoid conjunction fallacy or other biases (Kahneman & Tversky, 1996, p. 586). As for Gigerenzer's attack on the use of probability theory as norm for judgment on single-case events, it is difficult to justify. By definition, a frequentist concept of probability makes no sense for single-case propositions as it is concerned with measures relating to collections of events and with values that are degrees of confidence in these propositions. But this does not mean that concepts of probabilities cannot be meaningfully applied at all. Instead, one may accept a subjectivist concept of probability that concerns measures of sets of propositions, such as sets of "possible worlds" (Vranas, 2000, p. 182). I take this to mean that one can also think about probability as the likelihood that one lives in a world where a given event will happen.

3. Discussion

3.1. Reasoning about existential risk

In the introduction, I mentioned how seemingly altruistic people come off as irrational givers if their actual intent is to help other people as much as possible. There are several approaches one can take to address this phenomenon, perhaps the most uncharitable being the assumption that their actions are not from altruism but something else entirely. Indeed, some might argue that complex human behaviour seldom is about what we say it is. In their book “The Elephant in the Brain” (2018), Kevin Simler and Robin Hanson argues that our behaviour in many instances are driven by hidden motives regarding status and social signalling. While charity is good for its own sake, we also do it because it “feels good”, but why is this so? Simler and Hanson suggests that being charitable might feel good because it is beneficial to us in a social context; it is important to us to be seen as charitable. Displays of generosity might for example help to attract potential mates, as we generally want mates who will be generous to us and our offspring (Simler & Hanson, 2018, p. 216). Signs of altruism have indeed been shown to increase desirability as a long-term romantic partner (Barclay, 2010), and male participants tend to contribute more to charity when observed by a member of the opposite sex as opposed to being observed by other males or not observed at all (Iredale, Van Vugt & Dunbar, 2008).

Likewise, we can expect that similar dynamics are at play when it comes to concern for catastrophes. It might feel proper and right to show concern for the potential of future catastrophes and tragedy, but it is not obvious that we show interest in such scenarios merely out of genuine concern. In the case of activism to protect the environment, environmentalism has been studied as a signal for one’s willingness to cooperate, finding that donations to an environmental charity were higher when done in public rather than anonymously, but that participants donated the most when competing to be chosen by an observer for a subsequent cooperative game (Barclay & Barker, 2020). Be that as it may, it does not take away from the fact that some causes are recognised as urgent and important by the public. It matters less whether our concern to some extent can be attributed to hidden motives, as long as there is coordinated action to reduce suffering. The question then, is why mitigation of existential risks is not given preferential treatment to other types of risk mitigation, given its massive and permanent impact. In the same way that willingness to be charitable will result in altruistic, but ineffective, behaviour, it might be the case that concern for catastrophes can be real but somewhat misguided. Instead of claiming that people don't care enough about others, I want to explore the possibility of existential catastrophe as a non-intuitive concept of low accessibility, and thus a problem that we fail to make reasonable judgments about.

As have been shown in the previous section on decision making, there is an established theoretical framework of the mind that allows for bad judgment to take place regardless of sufficient knowledge, mainly in the form of heuristic biases. Could it be that specific components of the phenomenon of existential risk makes it especially difficult to engage with, so much so that neglect of existential risk may be (at least partly) understood as a form of cognitive failure? While this is a somewhat speculative assumption, it nevertheless seems valuable to explore this possibility given the unequivocally massive consequences of an existential catastrophe. As literature exploring this specific hypothesis currently is sparse, this may also warrant further examination. In this section I will start with discussing known heuristic principles and their relevancy to neglect of existential risk, and then follow up with implications of the theory of attribute substitution and more generally the interplay between System 1 and System 2 thinking and the meaning of rationality as seen through the lens of the tripartite model of the mind.

3.2. Existential risk assessment from heuristic principles

3.2.1. Availability of concepts

Out of the different types of heuristic principles described by Kahneman and Tversky, the availability heuristic is a good starting point from which to discuss how we make risk assessments in daily life. As mentioned, the availability heuristic is a mental shortcut that we often rely on to decide the importance and likelihood of the object in question. Instead of reasoning on actual probabilities, we can instead rely on how easily the object can be imagined and how familiar it is to us as a sign of its importance. After all, if something can be easily recalled and understood it is often because it is especially meaningful and important to us. The object's salience also influences its availability, meaning that objects that stand out as unusual would be easier to recall. Two early studies found that people are likely to overestimate frequency of rare causes of death and underestimate common causes

(Lichtenstein et al., 1978) and that there was a correlation with the amount of rare death causes reported in newspapers (Combs & Slovic, 1979). From this, one can infer that selective reporting combined with vivid imagery makes for highly available objects, and the availability heuristic is from time to time used to explain the impact of terror attacks: for example, it has been speculated that vivid imagery and dramatic descriptions of the 9/11 terror attack was used to generate interest and concern in the public for certain policies (Sunstein, 2006, p. 206).

3.2.2. Media and risk perception

As for assessment of catastrophes and existential risk, it could be that this likewise is reflected by what is reported in the media. By this, I mean that people's concern for catastrophes gets directed towards what is frequently discussed by news agencies and in social media, thereby explaining why much more resources are spent on mitigation of climate change than other causes with higher existential risk, as described earlier. If someone is to go into activism or otherwise act on their concern for the future, their choice of what cause to fight for could well be shaped by the information they are most exposed to. In other words, while there may be genuine intentions to improve wellbeing and flourishment of future lives, one is susceptible to believe that easily recalled catastrophic scenarios must be of higher importance, instead of evaluating all possible risk scenarios and from there decide what type of catastrophe is the most crucial to prevent.

This assumption of availability from influence by media is, however, not without its flaws, and the proposed effect media is said to have on risk perception has met some opposition. In what may be called a large-scale field test of the availability heuristic, the 10th year anniversary of the Chernobyl accident was utilized to investigate the influence of media attention on population samples from Sweden, Norway, France, the United Kingdom and Spain, finding that perceived risk of nuclear accidents were *not* affected despite extensive media attention to nuclear risk in this period (Sjöberg & Engelberg, 2010, p. 96). These surprising results led the researchers to question whether media effects on risk perception may be better explained as an effect of new information. Nuclear risk perception did increase at the time of the Chernobyl accident in 1986, but this could perhaps be because it was an unexpected news event, just like the 9/11 attack, which also was very surprising news that increased risk perception (Sjöberg & Engelberg, 2010, p. 96). Following up on these findings, a test paradigm involving films depicting different kinds of risk, namely nuclear risk and fire risk, was used to study the availability heuristic and effects of media. There was no significant change in risk perception, only some idiosyncratic and short-lived effects, suggesting that the impact media has on cognitive risk perception might not be due to the availability heuristic, but an effect of new information being provided to the participant (Sjöberg & Engelberg, 2010, p. 104).

One take-away from this research is that heuristic principles of availability are not as easy to manipulate with media as earlier thought. But more importantly, these studies suggest that depictions have little influence on people's risk assessment alone, or rather, they only show that risk perception did not change perception of catastrophes that the participants were already familiar with. Risk of nuclear disaster or spread of fire are both highly dramatic scenarios, and one can expect that the participants already had fairly vivid ideas about these kinds of disasters. In other words, real events, be it terror attacks or nuclear accidents, could be the main driver for the creation of highly accessible concepts to influence risk assessment, whereas the viewing of fictive scenarios play a limited role. Regarding the question of how we view the likelihood of existential catastrophes, this would then imply that depiction of doomsday scenarios in movies and other media is not sufficiently realistic for us to integrate them as accessible concepts for future judgment. Looking into the effect of media disaster reporting on public risk perception before and after a tornado event, another study found that media exposure in the form of simulated news coverings increased risk perception and safety actions for all participants, but that people in the post-tornado sample exhibited greater risk perception and reported a greater likelihood to seek shelter compared to the sample who had not yet experienced the tornado event (Zhao, Rosoff & John, 2019). Moreover, it has been shown that individuals who have experienced a flood before will rate the potential of future floods as more likely than those who have not, and that individuals who can recall high water levels are especially likely to have higher perceptions of the flood probability (Mol et al.,2020), further supporting the idea that people's risk perception of disasters are affected by the availability heuristic when they can relate to past experiences. Earlier studies on mismanagement of flood hazards have also indicated that when people underreact to flood threats, this failure can be attributed to their inability to conceptualize floods that they have no prior experience of, as prediction of future disaster is dependent on (lack of) flood experiences in the immediate past (Kates, 1962, p. 92, p. 88).

Taken together, these findings make a strong argument that past experiences with disasters affect our perception that such events also will happen in the future. If we can't easily recall memories of a specific disaster, we might think that such events are unlikely to happen because they are unfamiliar and thus feel unimportant and far-fetched to us. The problem then, is that there is no way for humanity to gain experience of an existential catastrophe without suffering its world-ending consequences. Whereas the availability heuristic can be useful to update on risk perception in other instances, it falls short when a completely new risk is created, namely that of man-made disaster. From a historical perspective it would make sense for us to rely on familiarity with some past disasters as a proxy for their likelihood to occur in the future, and even then, overreliance on such a rule-of-thumb would make us vulnerable to rare catastrophes, as implied by modern research on flood preparedness. The fact that our specie is alive today is in itself strong proof that natural catastrophes of existential magnitude are extremely unlikely to take place in the following centuries. But no such consolation can be made for anthropogenic existential risks. Given that scientific progress will not slow down over the coming years, it is not unthinkable that new discoveries will be made that can increase total existential risk to a much greater extent than nuclear fission did in 1938. Or, as Bostrom puts it, we can think of it as drawing balls from an urn, the balls mostly representing beneficial or neutral discoveries but with a few balls representing ease of mass destruction intermixed (Bostrom, 2019, p. 455). The challenge then, is that new technology always will be unfamiliar to us. The availability heuristic thus works against us in two ways: first, because we do not have easily accessible concepts for events of premature extinction in general, and second, because our intuition is likely to be wrong about the dangers of new inventions and discoveries.

3.2.3. Availability and attribute substitution

Another way to confront these failures in risk perception, is through the lens of the theory of attribute substitution. As I have mentioned before, the essence of attribute substitution is that a given problem or question is simplified in such a way that it becomes easier to answer through use of rule-of thumb principles, but at the cost of potentially deviating too much from the original object in question. Tasked with estimating the probability of a future event, we can say that the respondents in studies of risk perception replace the "likelihood" component with that of familiarity and availability of the object. This assumption fits with the finding that participants did not significantly update their risk perception following news coverage ten years after the Chernobyl accident or after watching a film about a specific risk, since both these test designs did not account for the individuals' familiarity with the catastrophe in question prior to exposure. The media did not represent "new information" of the kind that a new disaster would be, so one may speculate that they were deemed less important to update on. In any case, it is difficult to discern to what degree the participants were judging from heuristic principles in the first place, unlike the cases relating to flood hazard, where for example memories for water levels are much more indicative of influence from availability.

A unique problem for assessment of existential risk could simply be that most people have little knowledge of our collective vulnerability to technological dangers and realistic pathways that would end in premature extinction. This should restrict imaginability necessary to visualize such a scenario, which on the one hand makes it much more obscure and therefore unimportant to us, going by the principles of the availability heuristic. But on the other hand, one may classify this kind of cognitive failure as a case of a mindware gap, as defined by Stanovich. Rather than intuitive judgment being caused by a failure of the reflective system to initiate an override, the inherent abstractness of existential risk could "force" the individual to rely on familiarity and other heuristic principles, as there is too little knowledge to form decoupled simulations and do hypothesis testing on the problem. If one has no starting point, no foothold, from where to simulate a plausible scenario, it would be especially demanding to try and derive an estimate of probability, and it would instead be a better strategy to supplement "likelihood" with "familiarity" and from there reach a judgment that at least feels relevant. To take it even further, it could be that flawed preconceptions that existential catastrophes is impossible would shut off intents to engage with the risk through deliberate reasoning, and further promote trust in one's gut-feeling. If one has come to believe that total extinction is extremely unlikely, for example from realizing that natural disasters of such magnitude are incredibly rare, explicit concern for existential risk would seem meaningless.

3.2.4. Numb to future lives

A powerful argument for mitigation of existential risk is the consideration of all possible lives that forever would be denied existence, but this assumption is in of itself nothing more than a hypothetical. In order to become concerned for future lives, one would have to assign value to non-existent humans, whom at the moment only can be represented by numbers generated from calculations about continued growth. It goes without saying that we have no intuitive understanding of non-existent people, only through effortful imagination can we begin to speculate about their worth or right to exist. Could this be another attribute vulnerable to substitution? Once more, we are faced with a situation where accessibility of objects could influence judgment. I will argue that the people we surround ourselves with, share memories with and know through parasocial relationships constitute our social world in daily life. Regarding threats of catastrophes, my intuitive concern is first and foremost for the wellbeing of myself, my family, and friends, while it takes some conscious effort to extend this concern to the masses of people I likely never will meet. Related to this, one might speak of an *effect of psychic/mass numbing*: how the general public after a certain point comes to neglect suffering of others. Researching people's indifference to genocide and mass suffering, Slovic argued that helping behaviour is motivated by feelings like empathy, sympathy, compassion, sadness, pity, and distress, emphasising that research on the topic suggest that we are more likely to help people in need if we "feel for" that person (Slovic, 2007, p. 83).

As mentioned in the theory section, affect is thought to play a central role in decision making as a process of System 1, where reliance on affect and emotion is an easy and efficient way to make quick judgments in our complex daily life. Here, affect involves the way positive and negative emotions become attached to what Slovic refers to as "images", which are visual objects, but also other objects such as words, memories, or products of our imagination. Slovic therefore suggests that images, along with the grabbing of our attention is necessary to produce compassion and help others, and that abstract numbers are poor at triggering these mechanisms (Slovic, 2007, p. 83). Numerical representations of human lives fail at conveying the importance and value of those lives because we struggle to translate the number into the large collection of individuals that the number represents. Unlike an image of a single crying child, numbers themselves cannot to be attached to a highly accessible interpretation of their meaning. As such, we are for example more likely to care about the *proportion* of lives saved rather than about *numbers* of saved lives: thinking in percentages allows for us to easily visualise the amount of impact, whereas "10 000 lives" is an object without relative meaning that we have to deliberately connect to circumstances in the real world before we can decide its true implications (Slovic, 2007, p. 85). Thus, one may speak of a form of *psychophysical numbing* that occurs when we are presented with numbers that are too large for us to intuitively grapple with, in the sense that our feelings of concern and compassion stop being proportional to the number of lives at stake when the number is too large for us to accurately represent them in our mind (Slovic, 2007, p. 85).

Slovic’s numbing effect might as well be an example of cognitive bias from the availability heuristic, as it highlights the role of imaginability of objects that are subject to decision making. When a problem revolves around a small group of people, we have available a representation of their value as living beings, but as the number grows larger the group in question becomes a faceless abstraction to us. One may say that the exact number is substituted with a less precise term like “huge” or “enormous”. This makes it so that objects like “ten million lives” and “ten billion lives” become lumped together under the same category or placeholder of “enormous amount of lives”, thereby treating these numbers as if they were nearly equivalent to each other. For existential catastrophes to be recognised as meaningfully distinct from other massive but non-existential catastrophes, one must have in mind the number of potential lives that permanently will be denied existence if such an event were to take place. This number could well be in trillions or even quadrillions, but as the risk presents itself, its implications are nothing but a very large, but faceless and abstract number. Even thinking in percentages would be of little help, as we nevertheless are stuck with numbers too high to represent in our imagination. We can understand that the number is beyond enormous, but intuitively we struggle to assign a value that is proportional to it, and we fail to see what difference it would make. After all, it is already challenging enough that the problem deals with hypothetical lives that have yet to exist and that we never will interact with, so it could be tempting to equate the scenario to other events that also involve flourishing of future humans. If we are to act on concern for other people than ourselves, it would seem natural that our compassion is directed towards friends and family that we know and can conjure strong images of. But after that, I think it is challenging to really *feel* any different whether one additionally is saving lives in the millions or in the billions. While the affect heuristic enables us to efficiently recognise something as good or bad, its effectiveness is negligible for events of such magnitude, and attribute substitution could here explain our inclination to show indifference towards risks of existential consequence.

Another implication of the affect heuristic is the way perceived benefits alleviates risk perception of hazards. Recall that a negative correlation was found between perceived benefit and perceived risk, such that participants would ascribe less risk to potentially dangerous scenarios or technology if there was something positive to gain from it. Affect towards an object, such as different types of technology, can be manipulated by presentation of beneficial aspects to it, which in turn will make it more desirable and seem less dangerous. Anthropogenic existential risk involves dangers from creation of new technology, and given the affect heuristic, there should be cause for concern that perceived benefits of new technology results in neglect of its potential hazards. Noble goals like life extension and ending of poverty are incredible benefits that should increase overall positive affect towards discoveries and inventions that potentially will bring us closer to such goals. But such perceived benefits may in turn result in neglect of any associated risk, making us willing to take on greater gambles and go through radical changes even if it involves significant increases on total existential risk.

3.2.5. Representativeness heuristic

In what other ways might flawed judgment of existential risk be attributed to biases from heuristic principles? Another of the core heuristics described by Kahneman and Tversky, and which has stood the test of time, is judgment from the representativeness heuristic; the way intuitive assumptions are made based on how representative the object in question is to a related class of objects. Also considered a result of attribute substitution, individuals tend to avoid calculating on probabilities by unconsciously substituting probability with representativeness to arrive at an answer quickly and effortlessly. Interestingly, recent studies have used the representativeness heuristic to explain variance in risk perception for catastrophic events. As I have recently discussed, subjective perception of the risk of future tornados was elevated in individuals after the occurrence of an actual tornado event relative to people's risk perception before the tornado took place. In line with this, studies on demand for insurance against disastrous events have shown that people are rather reluctant to pay for insurance before a probable catastrophe occurs (Kunreuther, 2006), but become much more willing, and are even likely to pay unreasonable amounts, after a given catastrophe takes place (Kousky et al., 2010). This pattern of underreaction before the event, followed by overreaction after, may explain the neglect of probabilities as a result of biased heuristic thinking.

3.2.6. Base-rate neglect

But instead of viewing it as an effect of the availability heuristic, Volkman-Wise proposed that it also makes sense to consider it a failure from reliance on the reliability heuristic. While these two heuristic principles are similar to each other, she argues that the relevance of recent information and personal experience for risk perception makes the representativeness heuristic a better fit because it involves violation of Bayes' rule of prior and posterior probabilities (Volkman-Wise, 2015, p. 273). Kahneman and Tversky noted that representativeness is not influenced by prior probability, or base-rate frequency, in the sense that the base-rate frequency of an occupation (or other object class) has no relevance to how representative a description of a person (the object in question) is to that occupation (Tversky & Kahneman, 1974, p. 1124). In other words, we may find that a given object is very similar to the stereotype of a specific reference class, and thereby come to think of it as very likely to belong to that class, without taking into consideration the prior probability of encountering the reference class in the first place. Returning to the phenomenon of underinsuring against disasters prior to their occurrence, Volkman-Wise suggests that individuals in these situations tend to place too little weight on the prior probability of a catastrophe relative to posterior probabilities. In accordance with the representativeness heuristic, individuals here suffer from biased thinking because they use a subset of information to be representative of the entire breadth of information. Thus, a short history in time is taken to be representative of disaster frequency in general, meaning that a recent period with little or many disasters skews the individual's perception of disasters as being a much less, or more, frequent event (Volkman-Wise, 2015, p. 274).

Taking into account the phenomenon of base-rate neglect, we now have a more nuanced understanding of the role that lack of personal experience has to our intuitive judgment. As a mental shortcut to avoid effortful thinking of probability, individuals instead end up perceiving risk based on recent history of disasters, substituting probability with representativeness. This takes away some of the significance that was ascribed to the availability heuristic in previous paragraphs, as this helps to answer why personal experience is so relevant to change in risk perception, as well as the limits of media exposure. But what does this mean for assessment of existential risk? As with the effect of the availability heuristic, one might point out the challenge that we never have experienced such an event, and therefore cannot use the experience as part of our judgment. If we tend to underreact to risks of catastrophe before they happen and overreact afterwards, we might also underreact when confronted with questions about existential risk and after the catastrophe it does not matter if we overreact or not because we are already dead. This way of interpreting the dynamic of the representativeness heuristic is however somewhat problematic. It should be obvious that a complete extinction of mankind can take place only once, so we might want to take into consideration a longer timeline instead of recent history. Thus, a tempting strategy would be to see humanity's total time on earth up till now as being representative of our survival going into the future. So far, mankind has never succumbed to extinction, and while we have faced numerous catastrophes and likely been at the verge of experiencing many more, it is fair to say that humanity has no recollection of being close to *total* extinction. If we have made it thus far, who are to say that things will not turn out fine far into the future as well? This style of thinking actually works out great when the focus is on existential risk from natural catastrophes: our past is proof that all-destroying asteroids, earthquakes or volcano eruptions rarely take place on Earth, making it extremely unlikely that such events will take place in next millennia. But the same can not be said for anthropogenic risk. Unlike the risk from natural disasters, the likelihood of destruction from manmade technologies and discoveries is not static. The discovery of fire making methods involved dangers that we adapted to, and automobile accidents resulted in regulations and safety practices to reduce risk. But as development continues, our world has become more and more different from that of our ancestors, meaning it would be misleading to think of dangers of the past as representative of the future. As seen with attitudes towards insurance policy, there seems to be a tendency towards underreaction to threats that have yet to happen, and implications of the representativeness heuristic might be involved in creating such a blind spot.

3.2.6. Conjunction fallacy

Base-rate neglect is but one of several biases associated with the representativeness heuristic. As illustrated with the Linda problem, relying on representativeness can lead to a conjunction fallacy, as individuals fail to see that the conjunction of two descriptors/events are less probable than each one alone. Additional details make the option seem more similar of a specific class, leading to bad judgments when probability is substituted for representativeness. Similar effects are also seen for judgments about the future and risk perception. Regarding the forecasting of future scenarios, Kahneman and Tversky argued that scenarios which include a possible cause and an outcome would appear more probable than one of the outcomes alone. Indeed, participants would rate the probability of both an earthquake and a flood as higher than a flood happening alone when the conjunctive event was presented in a more detailed scenario. Even individuals with a professional background in forecasting and planning were shown to favour a conjunctive alternative (a Russian invasion followed by suspension of diplomatic relations with the USA) over the simple scenario (suspension of diplomatic relations) (Tversky & Kahneman, 1983, p. 307f). This failure of assessment occurs because one of the events is presented as a cause for the other, leading the individual to assess the effect (suspension) given the cause (invasion), instead of taking into consideration the joint probability of the effect and the cause (both suspension and invasion). Attempts to forecast events typically involve the construction and evaluation of scenarios, a procedure which encourages us to include hypothetical causes leading up to the event. We want to fill in any holes to the story to make it more similar, more representative, to our idea of what such a scenario typically would look like. As such, a conjunctive scenario is misinterpreted as more probable than disjunctive scenarios (Tversky & Kahneman, 1983, p. 308). Adding more details to a hypothetical scenario makes it feel more coherent and more realistic, even though any forecast actually becomes less and less probable for each detail that is added to it. Upon reflection, it is clear that a vague forecast is much more likely to come true, and yet we seem to be more easily convinced of a forecast when it provides additional causes that could lead up to the event, ignoring the fact that each detail is another condition that must play out for the forecast to come true. In a way, forecasting is not much different from risk perception, as it involves making judgments on events that have yet to play out. So, in what ways might the conjunction fallacy affect people's assessment and treatment of existential risk?

Given that detailed and specific scenarios are (unreasonably) regarded as more probable than scenarios with less conditions, there is a danger that resources earmarked for safety work will be spent on strategies tailor-made for very specific dangers instead of general risk prevention. The dilemma is that work on existential risk might be neglected in favour of mitigating minor risks that nevertheless feel more threatening because they provide vivid scenarios with causes leading to the effect. Here, a parallel has been drawn to how US government spending for some time was drawn away from emergency-response capabilities that could respond to any disaster, in favour of mitigating highly improbable, but detailed and film-like terror attacks (Yudkowsky, 2006, p. 7). "Existential catastrophe" is already a vague concept because it is a category for a number of different scenarios, including events we cannot yet imagine.

Vagueness is in this case preferential because existential risk mitigation involves mitigation of any disaster that results in extinction.

The finding that the probability of conjunctive events is overestimated, whereas disjunctive events are underestimated, could provide a false sense of security about the future. As an illustrative metaphor, I will reintroduce Bostrom’s image of scientific progress being akin to drawing balls from an urn. The balls represent all discoveries and inventions, where a small amount will cause a catastrophe leading to premature extinction. Even if each of these balls by themselves only represent a small chance of existential catastrophe, for example 10% chance of being drawn, the more balls that are drawn, the higher total existential risk will be. Remember that it is sufficient if only one of the balls result in catastrophe, as an existential risk scenario is a disjunctive event. Coincidentally, Kahneman and Tversky used a test design where participants were asked to decide the probability of drawing different kinds of marbles from a bag, finding that they considered it more likely to draw seven red marbles, with replacement, in a row from a bag with 90% red marbles, than to draw a single red marble from a bag with 50% red marbles. By the participants’ judgment, the least likely outcome of them all was that at least one of seven marbles drawn, with replacement, from a bag with 10% red marbles, would be red. In other words, there was unreasonable optimism for the conjunctive event (all seven are red) and unreasonable pessimism for the disjunctive event (at least one of seven is red) (Tversky & Kahneman, 1974, p. 1129). Taken together, it is possible that we in general come to underestimate the likelihood of an existential catastrophe because we, like the marble-drawing participants, fail to take into account the compounding risk over time, and instead consider the probability of each catastrophic event in isolation. As time goes by, the probability of premature extinction could be increasing, even as it subjectively does not seem like it.

The optimism towards conjunctive scenarios, like drawing seven red balls in a row, also reflects a tendency towards false perceptions of safety from overly detailed reassurances. As Yudkowsky points out, any claim that a given event or object is not an existential risk becomes weaker for each condition included in the argument. While any condition by themselves might be strongly grounded in reality, any reassurance that depends on several conditions coming true will only be as good as its weakest link, or proposition (Yudkowsky, 2006, p. 7). To illustrate this, Yudkowsky presented an argument about safety against nanotechnology, but one can also find the same weakness in more general statements about existential risk such as “if we ever come close to experiencing an existential catastrophe the threat will be recognised ahead of time, and people will unite to prevent it by spending resources we have set aside to do whatever it takes to mitigate its risk". Reassuring statements like this might be more believable because they depict a coherent narrative where some parts are interpreted as given conditions and others as the effect to focus on and evaluate. The point here is not necessarily that we explicitly make such statements and therefore end up neglecting existential risk, but rather that we are prone to be too optimistic that the future will play out the way we imagine it. As seen in the studies on forecasting, more details increase perceived probability, even as this makes it a forecast of conjunctive events.

3.2.7. Anchoring and adjustment

When Kahneman and Tversky initially described the preference for conjunctive events over disjunctive events, it was discussed as a bias from the heuristic of anchoring and adjustment: how a given starting point, meaningful or arbitrary, comes to influence judgment because one fails to adjust sufficiently away from the starting point. The biased perception of conjunctive and disjunctive events was here explained as an effect of anchoring, and it was suggested that the participants in the marble-drawing scenario overestimated the likelihood of seven consecutive red marbles because the initial 90% of success functioned as an anchor of probability for the whole event. Likewise, the disjunctive event could be underestimated due to the low 10% chance that the first marble drawn would be red (Tversky & Kahneman, 1974, 1129). While being a fitting explanation, the anchoring effect distinguishes itself from other heuristics discussed so far, as the effect cannot readily be understood from the theory of attribute substitution. While effects of availability, affect and representativeness can be regarded as a simplification of a problem at hand through substitution of attributes that are easier to process, anchoring can be thought of as the manipulation of the accessibility of attributes. When an individual takes into consideration a given starting point, their judgment is influenced by the way that starting point temporarily increases the accessibility of a particular value of the target attribute, relative to other values of the same attribute (Kahneman, 2002, p. 707). In other words, the intuitive impression of the value of an object affects further judgment of it, similar to how heuristic judgment in general occurs because of differential accessibility of attributes. Looking back on arguments made in the previous paragraphs, the introduction of new information or recollection of recent history may function as an anchor that rely too much on, as we fail to sufficiently adjust away from it.

One might therefore speak of an interplay between the availability heuristic and the anchoring effect. Specifically addressing the potential for biases in assessment of existential risk, Bostrom proposed the concept of *good-story bias* to describe the hypothetical effects of fiction. Given that our intuitions about what the future of humanity will look like are shaped by stories that we are exposed to through media, there may be a risk that our intuitions are biased towards scenarios that make for an entertaining story. Fantastic depictions with heroes, drama and cliff-hangers might be enjoyable to watch, but this may mislead us into believing that much more realistic but boring scenarios regarding existential risk are too farfetched to be worth taking seriously (Bostrom, 2002, p. 14). Again, specific and dramatic risk scenarios may draw away attention from a more sober assessment of existential risk in general. Similarly, this interplay between anchoring and availability has been hypothesised as introducing the potential for a logical fallacy of generalization from fictional evidence, highlighting the pressure on storytellers to construct detailed and vivid, and thus less probable stories about future scenarios (Yudkowsky, 2006, p. 13). While Sjöberg and Engelberg's research on media exposure and risk perception provides a counterargument regarding the influence of fiction, one may argue that their findings are less relevant for influence on intuitions about existential risk, as the concept of premature extinction of humanity only is accessible through fiction and hypothetical scenarios.

To suggest that science-fiction could influence judgment might come off as highly speculative and even ridiculous, were it not for the well-documented finding that arbitrary information may function as anchors. As previously mentioned, early studies on the anchoring effect found that numbers which participants knew were picked at random nevertheless influenced their judgment. A weakness of this initial study lies in its design, as anchored judgments about the number of African countries under laboratory conditions might not be representative of judgments in real life that are of consequence. The influence of arbitrary anchors has however also been shown to have effect on judicial decisions made by legal experts. Englich and colleagues (2006) found that legal professionals tasked with deciding the appropriate sentencing when reviewing sets of case material would be influenced not only by irrelevant information such as a journalist's questions, but also from obviously arbitrary anchors like having them roll a die before deciding the judgment. The influence of irrelevant anchors did not depend on the judge's experience and expertise, as even practitioners with experience in similar cases were just as prone to influence from anchoring (Englich et al., 2006, p. 197). Moreover, they found that processing a random sentencing anchor leads to a selective increase in the accessibility of arguments that are consistent with the anchor, thereby providing support that the anchoring effect can be understood as a mechanism of selective accessibility. That is, considering an anchor value selectively increases the accessibility of knowledge indicating that aspects of the object in question are similar to that of the anchor (Englich, et al., 2006, p. 195f). Returning to the influence of media on existential risk, it could that stereotypical science-fiction scenarios, being highly accessible objects in the mind, serves as a self-generated anchor to the individual under circumstances where they have little else information to base their judgment on. As such, it would perhaps not matter whether that information consciously is labelled as fiction or not, as long as it can function as a starting point from which to build associations on.

3.3. Reflective mind

So far, I have discussed how assessments on existential risk can be made on the basis of the most well-known heuristic principles asserted by Kahneman and Tversky. For the most part, we may suspect that actions of attribute substitution can be involved in creating a “blind spot” for existential risk. It is important to emphasise that it is not obvious we should be caring about threats of premature extinction right now, and neither is concern for existential catastrophes a particularly visible part of public discourse or a major post for government spending. Nevertheless, it is peculiar how there is visible focus and action towards the betterment of future lives, and yet existential risk does not appear to be a primary concern to most people. If not today, as technological advancement continues there may come a day when existential risks truly deserve more attention, and to me it therefore seems beneficial to have explored and mapped out ways in which our judgment and decision making may struggle to deal with threats of premature extinction. The following paragraphs will discuss mechanisms that lend humanity prone to such biased judgment on existential risk, including discussion of the role and need for rationality.

3.3.1. Cognitive miserliness

Why *would* we be likely to resolve to heuristic thinking over more careful and contemplative thought for matters regarding existential risk? An initial answer here, is “why shouldn’t we?”. The intuitive system, or the automated set of systems (TASS) as Stanovich refers to it, is considered an efficient and effortless group of cognitive mechanisms that we rely on continuously throughout the day without much problem. As I have mentioned, it makes sense from an evolutionary standpoint to choose to expend the least amount of energy, time, and effort, meaning that our judgments, strategies and behaviour often is the result of “cognitive miserliness”. And as a first rule of cognitive miserliness, people will default to TASS processes whenever it is possible to do so. In this view, overreliance on TASS processing can be classified as a type of cognitive failure due to over-economizing of cognitive resources. Threats of an existential catastrophe is not something we can actually see or otherwise perceive; it is an abstract concept regarding the future and hypothetical future lives. Intuitively, there may not be much difference between an existential catastrophe and any other kind of (less impactful) catastrophes if they involve our death and the death of those we most care for in either case. Moreover, unlike other catastrophic risks, the closest we can get to representing such extinction scenarios is through consumption of fiction or effortful speculative forecasts. Given that most people neither have the time nor the resources to go about doing hypothetical thinking and forecasting of such abstract futures, concepts of doomsday and the end of the world could be immediately disregarded as nothing but a plot devise for science-fiction. More concrete threats, or simply normal problems in daily life, may instead be perceived as more deserving of cognitive expenditure. In some cases, it could even be very beneficial to rely on intuition when such ideas are brought up, as doomsayers promoting fear could be doing so out of self-interest and other hidden motives. Finally, one can even claim that ignorance to existential risk is a rational stance for most people. These matters concern us all, and if individual actions have an inconsequential impact there is little incentive to be concerned for existential risk until the potential for such catastrophic scenarios become more evident.

If defaulting to TASS is the first rule of cognitive miserliness, a second rule, according to Stanovich, could be to engage in serial associative cognition with a focal bias in situations where analytic processing seems necessary (Stanovich, 2008, p. 70). Recall that Stanovich defines this as a process of the analytic system, and unlike TASS, it is slow and creates judgment through serial processing from a given starting point. It is however thinking of a shallower sort than simulation and hypothesis generation because it does not involve representation building from scratch, but instead building associations on a model of the world as it is presented to the individual. This starting point constitutes a focal bias, and in combination with effects from availability and representativeness, there could be a case of problematic anchoring that misdirects any analytic processing of existential risk. Seeking confirmation for an object or rule rather than attempting to disprove it was considered as an effect of serial associative cognition with focal bias, in the sense that one becomes fixated on the initial presentation of the model. Instead of coming up with alternative models, it is more efficient to draw on associations from the presented model to confirm it and accept it as truth. This type of confirmation bias could likewise impede assessment of existential risk. If our most accessible representations of extinction scenarios come from unrealistic depictions in media, further thinking on such scenarios could lead to quick dismissal of existential risk in general. Similarly, if we are confronted with questions framing threats of premature extinction as highly unlikely due, for example, to humanity's tenacity, we could be tempted to seek confirmation that any threat can be overcome, instead challenging such optimistic worldviews. Most people do not already fear that the sky is falling, and it is easier to reaffirm this belief rather than actively seek evidence to disprove it. Again, one might frame neglect of existential risk as a result of our cognitive miserliness.

In Stanovich' taxonomy, over-economizing is but one of types of cognitive failure. As mentioned before, erroneous judgment could also stem from a mindware gap. It would be wrong think that disinterest in existential risk solely can be attributed cognitive failure and biased judgment, considering that anthropogenic existential risks are a category of objects that demand somewhat advanced prior knowledge across several theoretical fields. Gain-of-function-research and safety work on artificial intelligence are examples of this, being complex and narrow disciplines that only a minority of the population are specialised in. If there is no intuitive fear of premature extinction, there is no pressure to delve into it either and thus maintaining a blind spot for existential risk.

Finally, we have the standard view of biased judgment and decision making, which is the way in which output from TASS is given free rein to shape our judgment and view of the world, in cases where override should have been initiated. Intuitive judgment has its benefits for being fast rules-of-thumb that can be employed in a vast array of day-to-day situations, but as our interaction with the world is rapidly changing due to innovation and technological advancements, some outputs from TASS become outdated. This can for example manifest itself in the form of road rage, where the impulse towards aggression has no use on the highway and instead could put us in danger. Speaking of technology, the danger is that we create new threats that we have yet to adapt to, and that we cannot grasp through intuitive judgment. By this, I mean that we may not have built-in sensitivity towards events that constitute existential threat, but also that we might be susceptible to miscategorise or misinterpret its attributes. Existential risk scenarios will not typically be suitable for quick assessment by heuristics such as representativeness, availability or affect since they are novel and unnatural threats. Learning from personal experience with existential catastrophe is an oxymoron. As such, instructions from TASS are not only sometimes irrelevant, but potentially also harmful. This emphasises the need for timely initiation of successful override to activate decoupled processing by the analytic system. In Stanovich' taxonomy, this is where true override failure can happen: what one could think of as problems with willpower, or more precisely, a problem of multiple minds (Stanovich, 2008, p. 75). In override failure, one may imagine a struggle taking place between different minds, a fight for influence on behaviour between the autonomous and the analytic mind which the analytic mind ends up losing. Unlike the defaulting to TASS processes, which is an error out of cognitive miserliness, override failure involves the initiation of cognitive decoupling that then fails to be sustained until completion. As was described in the theory section, TASS outputs can sometimes conflict with our superordinate goals. Such undesirable signals can be thought of as short-leashed genetic control mechanisms that are pre-programmed to fulfill basic goals of survival and reproduction to primarily benefit the "replicator" and secondarily the "vehicle". As such, one may speak of a conflict between adaptation-executors that originate from evolutionary adaption, and fitness-maximizers, which are long-leashed self-control towards flourishing.

To Stanovich, successful override and initiation of decoupling is the essence of rationality. Override of TASS relies on algorithmic-level computational power, but the process must be triggered at the intentional, reflective level that represents long-leashed control mechanisms of the individual. The construct of intelligence represents the potential computational power to go through with TASS override and decoupled simulation, but this can only be triggered by the superordinate control hierarchies that Stanovich identifies as rational thinking positions (Stanovich, 2010, p. 154). The question that remains is to what extent humanity may be considered a rational animal, or at least rational enough to prevent its own premature extinction.

As the heuristics and biases program gradually uncovered weaknesses in our judgment, conflicting interpretations of its meaning arose. This may be referred to as the great rationality debate, between three major positions of opposing theorists appropriately named by Stanovich as the Panglossian, the Meliorist, and the Apologist. Whereas the Panglossian represents the view of humanity as perfectly rational, where any "biases" can be explained as unimportant performance errors or problems with the experimenter rather than the subject, the Meliorist accepts that there is substantial room for improvement in human reasoning, and that humans are somewhat irrational beings that often don't know what they really want (Stanovich, 2010, p. 155f). Finally, the Apologist acknowledges all shortcomings, but attributes it to the resource-limited nature of the brain instead of outright irrationality (Stanovich, 2010, p. 157). The heuristics and biases program, as well as the work by Stanovich and others in cognitive psychology, can here be placed in the Meliorist position, whereas Gigerenzer and other theorists emphasising the findings regarding frequency formats can be recognised as having the apologist stance. It is interesting to note that the Apologist view does not refute the findings of bias in single-case scenarios, but instead proposes that humans' intuitive judgment system is better adapted for thinking in a frequency format. This might explain some of the biases associated with intuitive estimations on the probability of a single-case propositions. In the Apologist view, humans are not always reasoning properly, but their errors are natural and understandable given existing cognitive constraints. But does this view provide a different perspective of intuitive assessments of existential risk and premature extinction? There is perhaps hope in the belief that problems, when redesigned to match our sensitive machinery are more solvable, and that such reframing generally will enhance cognitive performance. But we cannot get away from the fact that human extinction is a single-case event. It makes little sense to ask for the frequency of existential catastrophes. If anything, the Apologist position seems to reinforce the idea that humanity may have a blind spot for existential risk, and perhaps especially so for threats of anthropogenic catastrophe. We face very different problems and threats than our ancestors did in their pre-historic environment, making it that much more important to temper outdated inputs from TASS in favour of exhaustive hypothesis generation and simulation manipulation at the long-leashed intentional level of cognition. There is not much comfort in taking a Panglossian position either. There is little doubt that we time and again fail to prepare enough before catastrophes strike. Sometimes, catastrophes are even caused by us, as it is difficult to blame anyone else but humans for pollution, corruption, and war. If one were to exchange labels of “bias” and “irrationality” with ideas of hidden motives and logical self-interest, this does not make the problems disappear. Given a belief that mankind is good and that we want to care for each other, we must explore the possibilities of blind spots and biases in our cognition.

4. Conclusion

In my thesis I have attempted to cover findings in cognitive literature that may provide insight into challenges to how we as humans would be able to understand and relate to existential risk. One of the primary reasons for exploring the possibility that people irrationally come to neglect existential risk, was the findings by Schubert and colleagues (2020), revealing that people tend to not think of human extinction as something uniquely bad when compared with other kinds of catastrophes. With a starting point in the heuristics and biases program and the tripartite model of the mind, I discussed the ways in which heuristic principles may struggle to deal with questions of existential risk, consequently restricting our ability to accurately assess risks of premature extinction. Existential risk was proposed to be a uniquely difficult object to represent and interpret on intuitive levels of the mind. I therefore raise a question that perhaps humanity has a blind spot for scenarios of existential catastrophe that makes us vulnerable to take on more existential risk even if we rationally should not do so.

4.1. Future research

For dangers of premature extinction, future research could involve experiments on the effects that fiction and other media could have on existential risk perception and affect towards such a scenario. It would also be fruitful to explore any psychological effects associated with adopting a belief that premature extinction could happen in the foreseeable future, as well as individual variance in sensitivity and interest towards the topic. Finally, it would also be valuable to further explore hidden motives and incentives that could explain neglect for existential risks.

5. List of references

Barclay, P. (2010). Altruism as a courtship display: Some effects of third-party generosity on audience perceptions. The British Journal of Psychology, 101(1), 123–135. https://doi.org/10.1348/000712609X435733

Barclay, P. & Barker, J. L. (2020). Greener Than Thou: People who protect the environment are more cooperative, compete to be environmental, and benefit from reputation. *Journal of Environmental Psychology*, 72, 101441. https://doi.org/10.1016/j.jenvp.2020.101441

Bensinger, R. & Yudkowsky, E. (2021, 11. November) Discussion with Eliezer Yudkowsky on AGI interventions. *Less Wrong*. https://www.lesswrong.com/posts/CpvyhFy9WvCNsifkY/discussion-with-eliezer-yudkowsky-on-agi-interventions

Bostrom, N. (2002) Existential risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology*, Vol. 9, No. 1

Bostrom, N. (2019). The Vulnerable World Hypothesis. *Global Policy*, 10(4), 455–476. https://doi.org/10.1111/1758-5899.12718

Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual Differences in Adult Decision-Making Competence. *Journal of Personality and Social Psychology*, 92(5), 938–956. https://doi.org/10.1037/0022-3514.92.5.938

Burgoyne, A. P., Mashburn, C. A., Tsukahara, J. S., Hambrick, D. Z. & Engle, R. W. (2021) Understanding the relationship between rationality and intelligence: a latent-variable approach. *Thinking & Reasoning*. https://doi.org/10.1080/13546783.2021.2008003

Colman, A. M. (2015) “Cognitive psychology, n.”. Oxford Dictionary of Psychology. Fourth ed. Oxford University Press

Combs, B. & Slovic, P. (1979). Newspaper Coverage of Causes of Death. *Journalism Quarterly, 56(4), 837-849.* https://doir.org/10.1177/107769907905600420

Del Missier, Mäntylä, T., & Bruine de Bruin, W. (2010). Executive functions in decision making: An individual differences approach. *Thinking & Reasoning*, 16(2), 69–97. https://doi.org/10.1080/13546781003630117

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making. Personality & Social Psychology Bulletin, 32(2), 188–200. https://doi.org/10.1177/0146167205282152

"Eschatology, n.". OED Online. March 2022. Oxford University Press. https://www-oed-com.zorac.aub.aau.dk/view/Entry/64274?redirectedFrom=eschatology& (accessed April 30, 2022).

Frederick, S. 2005. "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives*, 19 (4): 25-42.

Gigerenzer, G. (1991). How to Make Cognitive Illusions Disappear: Beyond "Heuristics and Biases." *European Review of Social Psychology*, 2(1), 83–115. https://doi.org/10.1080/14792779143000033

Iredale, Van Vugt, M., & Dunbar, R. (2008). Showing Off in Humans: Male Generosity as a Mating Signal. *Evolutionary Psychology*, 6(3), 147470490800600–. https://doi.org/10.1177/147470490800600302

Kahneman, D. & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, 103(3), 582–596. https://doi.org/10.1037//0033-295x.103.3.582

Kates, R. W. (1962) Hazard and Choice Perception in Flood Plain Management. No. 78. University of Chicago, Department of Geography Research

Kousky, C., Michel-Kerjan, E.O., & Raschky, P.A. (2010). Demand for flood insurance. Working paper.

Kunreuther, H. (2006). Disaster Mitigation and Insurance: Learning from Katrina. *The Annals of the American Academy of Political and Social Science*, 604(1), 208–227. https://doi.org/10.1177/0002716205285685

LaFrance, M. & Hecht, M. A. (1995). Why Smiles Generate Leniency. *Personality & Social Psychology Bulletin*, *21(3*), 207–214. https://doi.org/10.1177/0146167295213002

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M. & Combs, B. (1978). Judged frequency of lethal events. Journal of Experimental Psychology. *Human Learning and Memory*, 4(6), 551–578. https://doi.org/10.1037//0278-7393.4.6.551

McNeil, B. J., Pauker, S. G., Sox, H. C., & Tversky, A. (1982). On the Elicitation of Preferences for Alternative Therapies. *New England Journal of Medicine,* 306(21), 1259–1262. doi:10.1056/nejm198205273062103

Mol, J. M., Botzen, W. J. W., Blasch, J. E., & de Moel, H. (2020). Insights into flood risk misperceptions of homeowners in the Dutch river delta. *Risk Analysis, 40(7),* 1450–1468. https://doi.org/10.1111/risa.13479

Oppenheimer, D. M. (2004). Spontaneous Discounting of Availability in Frequency Judgment Tasks. *Psychological Science*, 15(2), 100–105. https://doi.org/10.1111/j.0963-7214.2004.01502005.x

Ord, T., (2020) *The Precipice: Existential Risk and the Future of Humanity*, Hachette Books

Ring, P., Neyse, L., David-Barett, T., & Schmidt, U. (2016). Gender Differences in Performance Predictions: Evidence from the Cognitive Reflection Test. *Frontiers in Psychology*, 7, 1680–1680. https://doi.org/10.3389/fpsyg.2016.01680

Roser, M. (2022, 15. March) The Future is Vast: Longtermism's perspective on humanity's past, present, and future https://ourworldindata.org/longtermism

Schubert, S., Caviola, L., & Faber, N. S. (2019). The Psychology of Existential Risk: Moral Judgments about Human Extinction. *Scientific Reports*, 9(1), 15100–15108. https://doi.org/10.1038/s41598-019-50145-9

Selgelid, M. J. (2016). Gain-of-Function Research: Ethical Analysis. *Science and engineering ethics*, 22(4), 923–964. https://doi.org/10.1007/s11948-016-9810-1

Sherman, & Kim, H. S. (2002). Affective Perseverance: The Resistance of Affect to Cognitive Invalidation. *Personality & Social Psychology Bulletin*, 28(2), 224–237. https://doi.org/10.1177/0146167202282008

Shevlin, H., Vold, K., Crosby, M., & Halina, M. (2019). The limits of machine intelligence: Despite progress in machine intelligence, artificial general intelligence is still a major challenge. *EMBO reports*, 20(10), e49177. https://doi.org/10.15252/embr.201949177

Simler, K. & Hanson, R., (2018). *The Elephant in the Brain: Hidden Motives in Everyday Life*. New York: Oxford University Press

Sjöberg, L. & Engelberg, E. (2010). Risk Perception and Movies: A Study of Availability as a Factor in Risk Perception. *Risk Analysis*, 30(1), 95–106

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research*, 177(3), 1333–1352. https://doi.org/10.1016/j.ejor.2005.04.006

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018) Performance on the Cognitive Reflection Test is stable across time. *Judgment and Decision Making*, 13, 260–267

Stanovich, K. E. (2008). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. Evans & K. Frankish (Eds.), In two minds: Dual processes and beyond (pp. 55-88). Oxford: Oxford University Press.

Stanovich, K. E. (2010). The robot's rebellion finding meaning in the age of Darwin. *University of Chicago Press*. https://doi.org/10.7208/9780226771199

Stanovich, K. E., Toplak, M. E., & West, R. F. (2020). Intelligence and rationality. In R. J. Sternberg (Ed.), *Cambridge Handbook of Intelligence* (2nd Edition) (pp. 1106-1139). New York: Cambridge University Press.

Sunstein, C. R. (2006). The Availability Heuristic, Intuitive Cost-Benefit Analysis, and Climate Change. *Climatic Change*, 77(1), 195–210. https://doi.org/10.1007/s10584-006-9073-y

Todd, B (2017, October) The case for reducing existential risks. *80,000 hours.* https://80000hours.org/articles/existential-risks/#why-these-risks-are-some-of-the-most-neglected-global-issues

Toplak, M.E., West, R.F. & Stanovich, K.E. (2011) The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Mem Cogn 39*, 1275. https://doi.org/10.3758/s13421-011-0104-1

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124–1131. doi:10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4), 293–315. doi:10.1037/0033-295x.90.4.293

Zhao, M., Rosoff, H., & John, R. S. (2019). Media Disaster Reporting Effects on Public Risk Perception
and Response to Escalating Tornado Warnings: A Natural Experiment: Media Disaster Reporting Effects. *Risk Analysis*, 39(3), 535–552. https://doi.org/10.1111/risa.13205