·5·

Cognitive biases potentially affecting judgement of global risks

Eliezer Yudkowsky

5.1 Introduction

All else being equal, not many people would prefer to destroy the world. Even faceless corporations, meddling governments, reckless scientists, and other agents of doom, require a world in which to achieve their goals of profit, order, tenure, or other villainies. If our extinction proceeds slowly enough to allow a moment of horrified realization, the doers of the deed will likely be quite taken aback on realizing that they have actually destroyed the world. Therefore I suggest that if the Earth is destroyed, it will probably be by mistake.

The systematic experimental study of reproducible errors of human reasoning, and what these errors reveal about underlying mental processes, is known as the heuristics and biases programme in cognitive psychology. This programme has made discoveries *highly* relevant to assessors of global catastrophic risks. Suppose you are worried about the risk of Substance P, an explosive of planet-wrecking potency which will detonate if exposed to a strong radio signal. Luckily there is a famous expert who discovered Substance P, spent the last thirty years working with it, and knows it better than anyone else in the world. You call up the expert and ask how strong the radio signal has to be. The expert replies that the critical threshold is probably around 4000 terawatts. 'Probably?' you query. 'Can you give me a 98% confidence interval?' 'Sure', replies the expert. 'I'm 99% confident that the critical threshold is above 500 terawatts, and 99% confident that the threshold is below 80,000 terawatts.' 'What about 10 terawatts?' you ask. 'Impossible', replies the expert.

The above methodology for expert elicitation looks perfectly reasonable, the sort of thing any competent practitioner might do when faced with such a problem. Indeed, this methodology was used in the Reactor Safety Study (Rasmussen, 1975), now widely regarded as the first major attempt at probabilistic risk assessment. But the student of heuristics and biases will recognize at least two major mistakes in the method – not logical flaws, but conditions extremely susceptible to human error. I shall return to this example in the discussion of anchoring and adjustments biases (Section 5.7).

The heuristics and biases programme has uncovered results that may startle and dismay the unaccustomed scholar. Some readers, first encountering the experimental results cited here, may sit up and say: 'Is that really an experimental result? Are people really such poor guessers? Maybe the experiment was poorly designed, and the result would go away with such-and-such manipulation.' Lacking the space for exposition, I can only plead with the reader to consult the primary literature. The obvious manipulations have already been tried, and the results found to be robust.

5.2 Availability

Suppose you randomly sample a word of three or more letters from an English text. Is it more likely that the word starts with an 'R' ('rope'), or that 'R' is its third letter ('park')?

A general principle underlying the heuristics and biases programme is that human beings use methods of thought – *heuristics* – which quickly return good approximate answers in many cases; but which also give rise to systematic errors called *biases*. An example of a heuristic is to judge the frequency or probability of an event by its *availability*, the ease with which examples of the event come to mind. 'R' appears in the third-letter position of more English words than in the first-letter position, yet it is much easier to recall words that begin with 'R' than words whose third letter is 'R'. Thus, a majority of respondents guess that words beginning with 'R' are more frequent, when the reverse is the case (Tversky and Kahneman, 1973).

Biases implicit in the availability heuristic affect estimates of risk. A pioneering study by Lichtenstein et al. (1978) examined absolute and relative probability judgements of risk. People know in general terms which risks cause large numbers of deaths and which cause few deaths. However, asked to quantify risks more precisely, people severely overestimate the frequency of rare causes of death and severely underestimate the frequency of common causes of death. Other repeated errors were also apparent: accidents were judged to cause as many deaths as disease. (Diseases cause about 16 times as many deaths as accidents.) Homicide was incorrectly judged a more frequent cause of death than diabetes or stomach cancer. A follow-up study by Combs and Slovic (1979) tallied *reporting* of deaths in two newspapers, and found that errors in probability judgements correlated strongly (.85 and .89) with selective reporting in newspapers.

People refuse to buy flood insurance even when it is heavily subsidized and priced far below an actuarially fair value. Kates 1962 suggest that underreaction to threats of flooding may arise from 'the inability of individuals to conceptualize floods that have never occurred ... Men on flood plains appear to be very much prisoners of their experience ... Recently experienced floods appear to set an upward bound to the size of loss with which managers believe they ought to be concerned'. Burton et al. (1978) report that when dams and levees are built, they reduce the frequency of floods, and thus apparently create a false sense of security, leading to reduced precautions. While building dams decreases the *frequency* of floods, damage *per flood* is so much greater afterwards that the average yearly damage *increases*.

It seems that most people do not extrapolate from experienced small hazards to a possibility of large risks; rather, the past experience of small hazards sets a perceived upper bound on risks. A society well protected against minor hazards will take no action against major risks (building on flood plains once the regular minor floods are eliminated). A society subject to regular minor hazards will treat those minor hazards as an upper bound on the size of the risks (guarding against regular minor floods but not occasional major floods).

Risks of human extinction may tend to be underestimated since, obviously, humanity has never yet encountered an extinction event.[1]

5.3 Hindsight bias

Hindsight bias is when subjects, after learning the eventual outcome, give a much higher estimate for the *predictability* of that outcome than subjects who predict the outcome without advance knowledge. Hindsight bias is sometimes called the I-knew-it-all-along effect.

Fischhoff and Beyth (1975) presented students with historical accounts of unfamiliar incidents such as a conflict between the Gurkhas and the British in 1814. Given the account as background knowledge, five groups of students were asked what they would have predicted as the *probability* for each of four outcomes: British victory, Gurkha victory, stalemate with a peace settlement, or stalemate with no peace settlement. Four experimental groups were, respectively, told that these four outcomes were the historical outcome. The fifth, control group was not told any historical outcome. In every case, a group told an outcome assigned substantially higher probability to that outcome, than did any other group or the control group.

Hindsight bias is important in legal cases, where a judge or jury must determine whether a defendant was legally negligent in failing to foresee a hazard (Sanchiro, 2003). In an experiment based on an actual legal case, Kamin and Rachlinski (1995) asked two groups to estimate the probability of flood damage caused by blockage of a city-owned drawbridge. The control

[1] Milan M. Ćirković points out that the Toba supereruption (~73,000 BCE) may count as a near-extinction event. The blast and subsequent winter killed off a super majority of humankind; genetic evidence suggests there were only a few thousand survivors, perhaps even less (Ambrose, 1998). Note that this event is not in our *historical* memory – it predates writing.

group was told only the background information known to the city when it decided not to hire a bridge watcher. The experimental group was given this information, plus the fact that a flood had actually occurred. Instructions stated that the city was negligent if the foreseeable probability of flooding was greater than 10%. As many as 76% of the control group concluded the flood was so unlikely that no precautions were necessary; 57% of the experimental group concluded the flood was so likely that failure to take precautions was legally negligent. A third experimental group was told the outcome and also explicitly instructed to avoid hindsight bias, which made no difference: 56% concluded the city was legally negligent. Judges cannot simply instruct juries to avoid hindsight bias; that debiasing manipulation has no significant effect.

When viewing history through the lens of hindsight, we vastly *underestimate* the cost of preventing catastrophe. In 1986, the space shuttle *Challenger* exploded for reasons eventually traced to an O-ring losing flexibility at low temperature (Rogers et al., 1986). There were warning signs of a problem with the O-rings. But preventing the *Challenger* disaster would have required, not attending to the problem with the O-rings, but attending to *every* warning sign which seemed as severe as the O-ring problem, *without benefit of hindsight.*

5.4 Black Swans

Taleb (2005) suggests that hindsight bias and availability bias bear primary responsibility for our failure to guard against what he calls *Black Swans*. Black Swans are an especially difficult version of the problem of the fat tails: sometimes *most of* the variance in a process comes from exceptionally rare, exceptionally huge events. Consider a financial instrument that earns \$10 with 98% probability, but loses \$1000 with 2% probability; it is a poor net risk, but it looks like a steady winner. Taleb (2001) gives the example of a trader whose strategy worked for 6 years without a single bad quarter, yielding close to \$80 million – then lost \$300 million in a single catastrophe.

Another example is that of Long-term Capital Management (LTCM), a hedge fund whose founders included two winners of the Nobel Prize in Economics. During the Asian currency crisis and Russian bond default of 1998, the markets behaved in a literally *unprecedented* fashion, assigned a negligible probability by LTCM's historical model. As a result, LTCM began to lose \$100 million per day, day after day. On a single day in 1998, LTCM lost more than \$500 million (Taleb, 2005).

The founders of LTCM later called the market conditions of 1998 a '10-sigma event'. But obviously it was *not* that improbable. Mistakenly believing that the past was predictable, people conclude that the future is predictable. As

Fischhoff (1982) puts it:

> When we attempt to understand past events, we implicitly test the hypotheses or rules we use both to interpret and to anticipate the world around us. If, in hindsight, we systematically underestimate the surprises that the past held and holds for us, we are subjecting those hypotheses to inordinately weak tests and, presumably, finding little reason to change them.

The lesson of history is that swans happen. People are surprised by catastrophes lying outside their anticipation, beyond their historical probability distributions. Then why are we so taken aback when Black Swans occur? Why did LTCM borrow a leverage of $125 billion against $4.72 billion of equity, almost ensuring that *any* Black Swan would destroy them?

Because of hindsight bias, we learn *overly specific* lessons. After September 11, the US Federal Aviation Administration prohibited box-cutters on airplanes. The hindsight bias rendered the event too predictable in retrospect, permitting the angry victims to find it the result of 'negligence' – such as intelligence agencies' failure to distinguish warnings of Al Qaeda activity amid a thousand *other* warnings. We learned not to allow hijacked planes to overfly our cities. We did not learn the lesson: 'Black Swans do occur; do what you can to prepare for the unanticipated.'

Taleb (2005) writes:

> It is difficult to motivate people in the prevention of Black Swans ... Prevention is not easily perceived, measured, or rewarded; it is generally a silent and thankless activity. Just consider that a costly measure is taken to stave off such an event. One can easily compute the costs while the results are hard to determine. How can one tell its effectiveness, whether the measure was successful or if it just coincided with no particular accident? ... Job performance assessments in these matters are not just tricky, but may be biased in favor of the observed 'acts of heroism'. History books do not account for heroic preventive measures.

5.5 The conjunction fallacy

Linda is thirty-one year old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Rank the following statements from most probable to least probable:

1. Linda is a teacher in an elementary school.
2. Linda works in a bookstore and takes Yoga classes.
3. Linda is active in the feminist movement.
4. Linda is a psychiatric social worker.
5. Linda is a member of the League of Women Voters.

6. Linda is a bank teller.
7. Linda is an insurance salesperson.
8. Linda is a bank teller and is active in the feminist movement.

Among the eighty-eight undergraduate subjects, 89% ranked statement (8) as more probable than (6) (Tversky and Kahneman, 1982). Since the given description of Linda was chosen to be similar to a feminist and dissimilar to a bank teller, (8) is more *representative* of Linda's description. However, ranking (8) as more *probable* than (6) violates the conjunction rule of probability theory which states that $p(A \& B) \leq p(A)$. Imagine a sample of 1000 women; surely more women in this sample are bank tellers than are feminist bank tellers.

Could the conjunction fallacy rest on subjects interpreting the experimental instructions in an unanticipated way? Perhaps subjects think that by 'probable' is meant the probability of Linda's description given statements (6) and (8), rather than the probability of (6) and (8) given Linda's description. It could also be that subjects interpret (6) to mean 'Linda is a bank teller and is not active in the feminist movement'. Although many creative alternative hypotheses have been invented to explain away the conjunction fallacy, the conjunction fallacy has survived all experimental tests meant to disprove it; see, for example, Sides et al. (2002) for a summary. For example, the following experiment excludes both of the alternative hypotheses proposed earlier.

Consider a regular six-sided die with four green faces and two red faces. The die will be rolled 20 times and the sequence of greens (G) and reds (R) will be recorded. You are asked to select one sequence, from a set of three, and you will win $ 25 if the sequence you chose appears on successive rolls of the die. Please check the sequence of greens and reds on which you prefer to bet.

1. RGRRR
2. GRGRRR
3. GRRRRR

A total of 125 undergraduates at University of British Columbia (UBC) and Stanford University played this gamble with real pay-offs. Among them, 65% of subjects chose sequence (2) (Tversky and Kahneman, 1983). Sequence (2) is most *representative* of the die, since the die is mostly green and sequence (2) contains the greatest proportion of green faces. However, sequence (1) *dominates* sequence (2) because (1) is strictly included in (2) – to get (2) you must roll (1) *preceded* by a green face.

In the above-mentioned task, the exact probabilities for each event could in principle have been calculated by the students. However, rather than go to the effort of a numerical calculation, it would seem that (at least 65% of) the students made an intuitive guess, based on which sequence seemed most 'representative' of the die. Calling this 'the representativeness heuristic' does not imply that students deliberately decided that they would estimate probability by estimating similarity. Rather, the representativeness heuristic is what produces the intuitive sense that sequence (2) 'seems more likely' than sequence (1). In other words, the 'representativeness heuristic' is a built-in feature of the brain for producing rapid probability judgements, rather than a consciously adopted procedure. We are not *aware* of substituting judgement of representativeness for judgement of probability.

The conjunction fallacy similarly applies to futurological forecasts. Two independent sets of professional analysts at the Second International Congress on Forecasting were asked to rate, respectively, the probability of 'a complete suspension of diplomatic relations between the United States and the Soviet Union sometime in 1983' or 'a Russian invasion of Poland, and a complete suspension of diplomatic relations between the United States and the Soviet Union sometime in 1983.' The second set of analysts responded with significantly higher probabilities (Tversky and Kahneman 1983).

In a study by Johnson et al. (1993), MBA students at Wharton were scheduled to travel to Bangkok as part of their degree programme. Several groups of students were asked how much they were willing to pay for terrorism insurance. One group of subjects was asked how much they were willing to pay for terrorism insurance covering the flight *from* Thailand *to* the United States. A second group of subjects was asked how much they were willing to pay for terrorism insurance covering the round-trip flight. A third group was asked how much they were willing to pay for terrorism insurance that covered the complete trip to Thailand. These three groups responded with average willingness to pay of $17.19, $13.90, and $7.44, respectively.

According to probability theory, adding a detail to a hypothesis *must* render the hypothesis less probable. It is less probable that Linda is a feminist bank teller than that she is a bank teller, since all feminist bank tellers are necessarily bank tellers. Yet human psychology seems to follow the rule that *adding a detail can make the story more plausible.*

People might pay more for international diplomacy intended to prevent nanotechnological warfare *by China* than for an engineering project to defend against nanotechnological attack *from any source*. The second threat scenario is less vivid and alarming, but the defence is more useful *because* it is more vague. More valuable still would be strategies which make humanity harder to extinguish without being specific to nanotechnological threats – such as colonizing space, or see Chapter 15 (this volume) on Artificial Intelligence. Security expert Bruce Schneier observed (both before and after the 2005 hurricane in New Orleans) that the US government was guarding *specific* domestic targets against 'movie-plot scenarios' of terrorism, at the cost of taking away resources from emergency-response capabilities that could respond to *any* disaster (Schneier, 2005).

Overly detailed reassurances can also create false perceptions of safety: 'X is *not* an existential risk and you don't need to worry about it, because of A, B, C, D, and E'; where the failure of any *one* of propositions A, B, C, D, or E potentially extinguishes the human species. 'We don't need to worry about nanotechnological war, because a UN commission will initially develop the technology and prevent its proliferation until such time as an active shield is developed, capable of defending against all accidental and malicious outbreaks that contemporary nanotechnology is capable of producing, and this condition will persist indefinitely.' Vivid, specific scenarios can inflate our probability estimates of security, as well as misdirecting defensive investments into needlessly narrow or implausibly detailed risk scenarios.

More generally, people tend to overestimate conjunctive probabilities and underestimate disjunctive probabilities (Tversky and Kahneman, 1974). That is, people tend to overestimate the probability that, for example, seven events of 90% probability will *all* occur. Conversely, people tend to underestimate the probability that *at least one* of seven events of 10% probability will occur. Someone judging whether to, for instance, incorporate a new start-up, must evaluate the probability that many individual events will *all* go right (there will be sufficient funding, competent employees, customers will want the product) while also considering the likelihood that *at least one* critical failure will occur (the bank refuses a loan, the biggest project fails, the lead scientist dies). This may help explain why only 44% of entrepreneurial ventures[2] survive after four years (Knaup, 2005).

Dawes (1988, p. 133) observes: 'In their summations lawyers avoid arguing from disjunctions ('either this or that or the other could have occurred, all of which would lead to the same conclusion') in favor of conjunctions. Rationally, of course, disjunctions are *much* more probable than are conjunctions.'

The scenario of humanity going extinct in the next century is a disjunctive event. It could happen as a result of any of the existential risks discussed in this book – or some other cause which none of us foresaw. Yet for a futurist, disjunctions make for an awkward and unpoetic-sounding prophecy.

5.6 Confirmation bias

Peter Wason (1960) conducted a now-classic experiment that became known as the '2–4–6' task. Subjects had to *discover a rule*, known to the experimenter but not to the subject – analogous to scientific research. Subjects wrote three numbers, such as '2–4–6' or '10–12–14', on cards, and the experimenter told them whether the triplet *fit* the rule or *did not fit* the rule. Initially subjects were given the triplet 2–4–6, and told that this triplet fitted the rule. Subjects could

[2] Note that the figure 44% is for all new businesses, including small restaurants, rather than, say, dot-com start-ups.

continue testing triplets until they felt sure they knew the experimenter's rule, at which point the subject announced the rule.

Although subjects typically expressed high confidence in their guesses, only 21% of Wason's subjects guessed the experimenter's rule, and replications of Wason's experiment usually report success rates of around 20%. Contrary to the advice of Karl Popper, subjects in Wason's task try to *confirm* their hypotheses rather than *falsifying* them. Thus, someone who forms the hypothesis 'Numbers increasing by two' will test the triplets 8–10–12 or 20–22–24, hear that they fit, and confidently announce the rule. Someone who forms the hypothesis X–2X–3X will test the triplet 3–6–9, discover that it fits, and then announce that rule. In every case the *actual* rule is the same: the three numbers must be in ascending order. In some cases subjects devise, 'test', and announce rules far more complicated than the actual answer.

Wason's 2–4–6 task is a 'cold' form of *confirmation bias*; people seek confirming but not falsifying evidence. 'Cold' means that the 2–4–6 task is an emotionally neutral case of confirmation bias; the belief held is logical, not emotional. 'Hot' refers to cases where the belief is emotionally charged, such as political argument. Unsurprisingly, 'hot' confirmation biases are stronger – larger in effect and more resistant to change. Active, effortful confirmation biases are labelled *motivated cognition* (more ordinarily known as 'rationalization'). As stated by Brenner et al. (2002, p. 503) in 'Remarks on Support Theory':

> Clearly, in many circumstances, the desirability of believing a hypothesis may markedly influence its perceived support ... Kunda (1990) discusses how people who are motivated to reach certain conclusions attempt to construct (in a biased fashion) a compelling case for their favored hypothesis that would convince an impartial audience. Gilovich (2000) suggests that conclusions a person does not want to believe are held to a higher standard than conclusions a person wants to believe. In the former case, the person asks if the evidence *compels* one to accept the conclusion, whereas in the latter case, the person asks instead if the evidence *allows* one to accept the conclusion.

When people subject disagreeable evidence to more scrutiny than agreeable evidence, this is known as *motivated* scepticism or *disconfirmation bias*. Disconfirmation bias is especially destructive for two reasons: (1) Two biased reasoners considering the *same* stream of evidence can shift their beliefs in *opposite* directions – both sides selectively accepting only favourable evidence. Gathering more evidence may not bring biased reasoners to agreement. (2) People who are more skilled sceptics – who know a larger litany of logical flaws – but apply that skill *selectively*, may change their minds more slowly than *unskilled* reasoners.

Taber and Lodge (2000) examined the prior attitudes and attitude changes of students – when exposed to political literature for and against gun control and affirmative action, two issues of particular salience in the political life of the United States. The study tested six hypotheses using two experiments:

1. *Prior attitude effect.* Subjects who feel strongly about an issue – even when encouraged to be objective – will evaluate supportive arguments more favourably than contrary arguments.
2. *Disconfirmation bias.* Subjects will spend more time and cognitive resources denigrating contrary arguments than supportive arguments.
3. *Confirmation bias.* Subjects free to choose their information sources will seek out supportive rather than contrary sources.
4. *Attitude polarization.* Exposing subjects to an apparently balanced set of pro and con arguments will exaggerate their initial polarization.
5. *Attitude strength effect.* Subjects voicing stronger attitudes will be more prone to the above biases.
6. *Sophistication effect.* Politically knowledgeable subjects, because they possess greater ammunition with which to counter-argue incongruent facts and arguments, will be more prone to the above biases.

Ironically, Taber and Lodge's experiments confirmed all six of the authors' prior hypotheses. Perhaps you will say: 'The experiment only reflects the beliefs the authors started out with – it is just a case of confirmation bias.' If so, then by making you a more sophisticated arguer – by teaching you another bias of which to accuse people – I have actually harmed you; I have made you slower to react to evidence. I have given you another opportunity to fail each time you face the challenge of changing your mind.

Heuristics and biases are widespread in human reasoning. Familiarity with heuristics and biases can enable us to detect a wide variety of logical flaws that might otherwise evade our inspection. But, as with *any* ability to detect flaws in reasoning, this inspection must be applied *even-handedly* – both to our own ideas and the ideas of others; to ideas which discomfort us and also to ideas which comfort us. Awareness of human fallibility is dangerous knowledge, if you remind yourself of the fallibility of those who disagree with you. If I am selective about *which* arguments I inspect for errors, or even *how hard* I inspect for errors, then every new rule of rationality I learn, every new logical flaw I know how to detect, makes me that much stupider. Intelligence, to be useful, must be used for something other than defeating itself.

You cannot 'rationalize' what is not rational to begin with – as if lying were called 'truthization'. There is no way to obtain more truth for a proposition by bribery, flattery, or the most passionate argument – you can make more people *believe* the proposition, but you cannot make it more *true*. To improve the truth of our beliefs we *must* change our beliefs. Not every change is an improvement, but every improvement is necessarily a change.

Our beliefs are more swiftly determined than we think. Griffin and Tversky (1992) discreetly approached twenty-four colleagues faced with a choice between two job offers and asked them to estimate the probability that they would choose each job offer. The average confidence in the choice assigned the greater probability was a modest 66%. Yet only one of twenty-four respondents chose the option initially assigned the lower probability, yielding an overall accuracy of 96% (one of few reported instances of human *under*confidence).

The moral may be that *once you can guess what your answer will be* – once you can assign a greater probability to your answering one way than another – you have, in all probability, already decided. And if you were honest with yourself, you would often be able to guess your final answer within seconds of hearing the question. We change our minds less often than we think. How fleeting is that brief unnoticed moment when we cannot yet guess what our answer will be, the tiny fragile instant when there is a chance for intelligence to act – in questions of choice, as in questions of fact.

Thor Shenkel said: 'It ain't a true crisis of faith unless things could just as easily go either way.'

Norman R.F. Maier said: 'Do not propose solutions until the problem has been discussed as thoroughly as possible without suggesting any.' Robyn Dawes, commenting on Maier, said: 'I have often used this edict with groups I have led – particularly when they face a very tough problem, which is when group members are most apt to propose solutions immediately.'

In computer security, a 'trusted system' is one that you *are in fact trusting*, not one that is in fact trustworthy. A 'trusted system' is a system which, if it is untrustworthy, can cause a failure. When you read a paper which proposes that a potential global catastrophe is impossible, or has a specific annual probability, or can be managed using some specific strategy, then you trust the rationality of the authors. You trust the authors' ability to be driven from a comfortable conclusion to an uncomfortable one, even in the absence of overwhelming experimental evidence to prove a cherished hypothesis wrong. You trust that the authors did not unconsciously look just a little bit harder for mistakes in equations that seemed to be leaning the wrong way, before you ever saw the final paper.

However, if authority legislates that the mere suggestion of an existential risk is enough to shut down a project, or if it becomes a de facto truth of the political process that no possible calculation can overcome the burden of a suggestion once made, no scientist will ever again make a suggestion, which is worse. I do not know how to solve this problem. But I think it would be well for estimators of existential risks to know something about heuristics and biases in general, and disconfirmation bias in particular.

5.7 Anchoring, adjustment, and contamination

An experimenter spins a 'Wheel of Fortune' device as you watch, and the Wheel happens to come up pointing to (version one) the number 65 or (version two) the number 15. The experimenter then asks you whether the percentage of African countries in the United Nations is above or below this number. After you answer, the experimenter asks you your estimate of the percentage of African countries in the United Nations.

Tversky and Kahneman (1974) demonstrated that subjects who were first asked if the number was above or below fifteen, later generated substantially lower percentage estimates than subjects first asked if the percentage was above or below 65. The groups' median estimates of the percentage of African countries in the United Nations were twenty-five and forty-five, respectively. This, even though the subjects had watched the number being generated by an apparently random device, the Wheel of Fortune, and hence believed that the number bore no relation to the actual percentage of African countries in the United Nations. Payoffs for accuracy did not change the magnitude of the effect. Tversky and Kahneman hypothesized that this effect was due to *anchoring and adjustment;* subjects took the initial uninformative number as their starting point, or *anchor*, and then *adjusted* the number up or down until they reached an answer that sounded plausible to them; then they stopped adjusting. The result was under-adjustment from the anchor.

In the example that opens this chapter, we *first* asked the expert on Substance P to guess the actual value for the strength of radio signal that would detonate Substance P, and only *afterwards* asked for confidence bounds around this value. This elicitation method leads people to adjust upwards and downwards *from their starting estimate,* until they reach values that 'sound improbable' and stop adjusting. This leads to under-adjustment and too-narrow confidence bounds.

Following the study by Tversky and Kahneman (1974), continued research showed a wider range of anchoring and pseudo-anchoring effects. Anchoring occurred even when they represented utterly implausible answers to the question; for example, asking subjects to estimate the year Einstein first visited the United States, after considering anchors of 1215 or 1992. These implausible anchors produced anchoring effects just as large as more plausible anchors such as 1905 or 1939 (Strack and Mussweiler, 1997). Walking down the supermarket aisle, you encounter a stack of cans of canned tomato soup, and a sign saying 'Limit 12 per customer'. Does this sign actually prompt people to buy more cans of tomato soup? According to empirical experiment, it does (Wansink et al., 1998).

Such generalized phenomena became known as *contamination* effects, since it turned out that almost *any* information could work its way into a cognitive judgement (Chapman and Johnson, 2002). Attempted manipulations to eliminate contamination include paying subjects for correct answers (Tversky and Kahneman, 1974), instructing subjects to avoid anchoring on the initial quantity (Quattrone et al., 1981), and facing real-world problems (Wansink et al., 1998). These manipulations did not decrease, or only slightly decreased, the magnitude of anchoring and contamination effects. Furthermore, subjects asked whether they had been influenced by the contaminating factor typically did not believe they had been influenced, when experiment showed they had been (Wilson et al., 1996).

A manipulation which consistently *increases* contamination effects is placing the subjects in cognitively 'busy' conditions such as rehearsing a word-string while working (Gilbert et al., 1988) or asking the subjects for quick answers (Gilbert and Osborne, 1989). Gilbert et al. (1988) attribute this effect to the extra task interfering with the ability to *adjust* away from the anchor; that is, less adjustment was performed in the cognitively busy condition. This decreases adjustment, hence increases the under-adjustment effect known as anchoring.

To sum up, information that is *visibly* irrelevant still anchors judgements and contaminates guesses. When people start from information known to be irrelevant and adjust until they reach a plausible-sounding answer, they under-adjust. People under-adjust more severely in cognitively busy situations and other manipulations that make the problem harder. People deny they are anchored or contaminated, even when experiment shows they are. These effects are not diminished or only slightly diminished by financial incentives, explicit instruction to avoid contamination, and real-world situations.

Now consider how many media stories on Artificial Intelligence cite the *Terminator* movies as if they were documentaries, and how many media stories on brain-computer interfaces mention *Star Trek's* Borg.

If briefly presenting an anchor has a substantial effect on subjects' judgements, how much greater an effect should we expect from reading an entire book or watching a live-action television show? In the ancestral environment, there were no moving pictures; whatever you saw with your own eyes was true. People do seem to realize, so far as conscious thoughts are concerned, that fiction is fiction. Media reports that mention *Terminator* do not *usually* treat Cameron's screenplay as a prophecy or a fixed truth. Instead the reporter seems to regard Cameron's vision as something that, having happened before, might well happen again – the movie is recalled (is *available*) as if it were an illustrative historical case. I call this mix of anchoring and availability the *logical fallacy of generalization from fictional evidence.*[3]

Storytellers obey strict rules of narrative unrelated to reality. Dramatic logic is not logic. Aspiring writers are warned that *truth is no excuse*: you may not justify an unbelievable event in your fiction by citing an instance of real life. A good story is painted with bright details, illuminated by glowing metaphors; a storyteller must be concrete, as hard and precise as stone. But in forecasting,

[3] A related concept is the *good-story bias* hypothesized in Bostrom (2001). Fictional evidence usually consists of 'good stories' in Bostrom's sense. Note that not all good stories are presented as fiction.

every added detail is an extra burden! Truth is hard work and not the kind of hard work done by storytellers. We should avoid not only being *duped* by fiction – failing to expend the mental effort necessary to 'unbelieve' it – but also being *contaminated* by fiction, letting it anchor our judgements. And we should be aware that we are not always aware of this contamination. Not uncommonly in a discussion of existential risk, the categories, choices, consequences, and strategies derive from movies, books, and television shows. There are subtler defeats, but this is outright surrender.

5.8 The affect heuristic

The *affect heuristic* refers to the way in which subjective impressions of 'goodness' or 'badness' can act as a heuristic, capable of producing fast perceptual judgements, and also systematic biases.

In a study by Slovic et al. (2002), two groups of subjects evaluated a scenario in which an airport had to decide whether to spend money to purchase new equipment, while critics argued that money should be spent on other aspects of airport safety. The response scale ranged from zero (would not support at all) to twenty (very strong support). A measure that was described as 'Saving 150 lives' had mean a support of 10.4, whereas a measure that was described as 'Saving 98% of 150 lives' had a mean support of 13.6. Even 'Saving 85% of 150 lives' had higher support than simply 'Saving 150 lives'. The hypothesis motivating the experiment was that saving 150 lives sounds diffusely good and is therefore only weakly evaluable, whereas saving 98% of something is clearly very good because it is so close to the upper bound on the percentage scale.

Finucane et al. (2000) wondered if people conflated their assessments of the *possible benefits* of a technology such as nuclear power, and their assessment of *possible risks*, into an overall good or bad feeling about the technology. Finucane et al. tested this hypothesis by providing four kinds of information that would increase or decrease perceived risk or perceived benefit. There was no logical relation between the information provided (e.g., about risks) and the non-manipulated variable (e.g., benefits). In each case, the manipulated information produced an inverse effect on the affectively inverse characteristic. Providing information that increased perception of risk decreased perception of benefit. Similarly, providing information that decreased perception of benefit increased perception of risk. Finucane et al. also found that time pressure greatly *increased* the inverse relationship between perceived risk and perceived benefit – presumably because time pressure increased the dominance of the affect heuristic over analytical reasoning.

Ganzach (2001) found the same effect in the realm of finance: analysts seemed to base their judgements of risk and return for *unfamiliar* stocks upon a global affective attitude. Stocks perceived as 'good' were judged to have low risks and high return; stocks perceived as 'bad' were judged to have low return and high risks. That is, for unfamiliar stocks, perceived risk and perceived return were negatively correlated, as predicted by the affect heuristic.[4] For *familiar* stocks, perceived risk and perceived return were positively correlated; riskier stocks were expected to produce higher returns, as predicted by ordinary economic theory. (If a stock is safe, buyers pay a premium for its safety and it becomes more expensive, driving down the expected return.)

People typically have sparse information in considering future technologies. Thus it is not surprising that their attitudes should exhibit affective polarization. When I first began to think about such matters, I rated biotechnology as having relatively smaller benefits compared to nanotechnology, *and* I worried more about an engineered supervirus than about misuse of nanotechnology. Artificial Intelligence, from which I expected the largest benefits of all, gave me not the least anxiety. Later, after working through the problems in much greater detail, my assessment of relative benefit remained much the same, but my worries had inverted: the more powerful technologies, with greater anticipated benefits, now appeared to have correspondingly more difficult risks. In retrospect this is what one would expect. But analysts with scanty information may rate technologies affectively, so that information about perceived benefit seems to mitigate the force of perceived risk.

5.9 Scope neglect

Migrating birds (2000/20,000/200,000) die each year by drowning in uncovered oil ponds, which the birds mistake for water bodies . These deaths could be prevented by covering the oil ponds with nets. How much money would you be willing to pay to provide the needed nets?

Three groups of subjects considered three versions of the above question, asking them how high a tax increase they would accept to save 2,000, 20,000, or 200,000 birds. The response – known as Stated Willingness-to-Pay (SWTP) – had a mean of $80 for the 2000-bird group, $78 for 20,000 birds, and $88 for 200,000 birds (Desvousges et al., 1993). This phenomenon is known as *scope insensitivity* or *scope neglect*.

Similar studies have shown that Toronto residents would a pay little more to clean up all polluted lakes in Ontario than polluted lakes in a particular region of Ontario (Kahneman, 1986); and that residents of four western US states

[4] Note that in this experiment, *sparse information* played the same role as cognitive business or time pressure in increasing reliance on the affect heuristic.

would pay only 28% more to protect all fifty-seven wilderness areas in those states than to protect a single area (McFadden and Leonard, 1995).

The most widely accepted explanation for scope neglect appeals to the affect heuristic. Kahneman et al. (1999, pp. 212–213) write:

> The story constructed by Desvouges et al. probably evokes for many readers a mental representation of a prototypical incident, perhaps an image of an exhausted bird, its feathers soaked in black oil, unable to escape. The hypothesis of valuation by prototype asserts that the affective value of this image will dominate expressions of the attitude to the problem – including the willingness to pay for a solution. Valuation by prototype implies extension neglect.

Two other hypotheses accounting for scope neglect include *purchase of moral satisfaction* (Kahneman and Knetsch, 1992) and *good cause dump* (Harrison, 1992). 'Purchase of moral satisfaction' suggests that people spend enough money to create a 'warm glow' in themselves, and the amount required is a property of the person's psychology, having nothing to do with birds. 'Good cause dump' suggests that people have some amount of money they are willing to pay for 'the environment', and *any* question about environmental goods elicits this amount.

Scope neglect has been shown to apply to human lives. Carson and Mitchell (1995) report that increasing the alleged risk associated with chlorinated drinking water from 0.004 to 2.43 annual deaths per 1000 (a factor of 600) increased stated willingness to pay from $3.78 to $15.23 (a factor of four). Baron and Greene (1996) found no effect from varying lives saved by a factor of ten.

Fetherstonhaugh et al. (1997), in a paper titled 'Insensitivity to the value of human life: a study of psychophysical numbing', found evidence that our perception of human deaths, and valuation of human lives, obeys Weber's Law – meaning that we use a *logarithmic* scale. And indeed, studies of scope neglect in which the quantitative variations are huge enough to elicit any sensitivity at all, show small *linear* increases in Willingness-to-Pay corresponding to *exponential* increases in scope. Kahneman et al. (1999) interpret this as an additive effect of scope affect and prototype affect – the prototype image elicits most of the emotion, and the scope elicits a smaller amount of emotion which is *added* (not multiplied) with the first amount.

Albert Szent-Györgyi, famous Hungarian physiologist and the discoverer of vitamin C, said: 'I am deeply moved if I see one man suffering and would risk my life for him. Then I talk impersonally about the possible pulverization of our big cities, with a hundred million dead. I am unable to multiply one man's suffering by a 100 million.' Human emotions take place within an analogous brain. The human brain cannot release enough neurotransmitters to feel emotion a 1000 times as strong as the grief of one funeral. A prospective risk going from 10,000,000 deaths to 100,000,000 deaths does not multiply by ten the strength of our determination to stop it. It adds one more zero on paper for our eyes to glaze over, an effect so small that one must usually jump several orders of magnitude to detect the difference experimentally.

5.10 Calibration and overconfidence

What confidence do people place in their erroneous estimates? In Section 5.2 on Availability, I discussed an experiment on perceived risk, in which subjects overestimated the probability of newsworthy causes of death in a way that correlated to their selective reporting in newspapers. Slovic et al. (1982, p. 472) also observed:

> A particularly pernicious aspect of heuristics is that people typically have great confidence in judgments based upon them. In another followup to the study on causes of death, people were asked to indicate the odds that they were correct in choosing the more frequent of two lethal events (Fischhoff, Slovic, and Lichtenstein, 1977) – In Experiment 1, subjects were reasonably well calibrated when they gave odds of 1:1, 1.5:1, 2:1, and 3:1. That is, their percentage of correct answers was close to the appropriate percentage correct, given those odds. However, as odds increased from 3:1 to 100:1, there was little or no increase in accuracy. Only 73% of the answers assigned odds of 100:1 were correct (instead of 99.1%). Accuracy 'jumped' to 81% at 1000:1 and to 87% at 10,000:1. For answers assigned odds of 1,000,000:1 or greater, accuracy was 90%; the appropriate degree of confidence would have been odds of 9:1 – In summary, subjects were frequently wrong at even the highest odds levels. Moreover, they gave many extreme odds responses. More than half of their judgments were greater than 50:1. Almost one-fourth were greater than 100:1 – 30% of the respondents in Experiment 1 gave odds greater than 50:1 to the incorrect assertion that homicides are more frequent than suicides.

This extraordinary-seeming result is quite common within the heuristics and biases literature, where it is known as *overconfidence*. Suppose I ask you for your best guess as to an uncertain quantity, such as the number of 'Physicians and Surgeons' listed in the Yellow Pages of the Boston phone directory, or total US egg production in millions. You will generate some value, which surely will not be *exactly* correct; the true value will be more or less than your guess. Next I ask you to name a *lower bound* such that you are 99% confident that the true value lies *above* this bound and an *upper bound* such that you are 99% confident the true value lies *beneath* this bound. These two bounds form your 98% *confidence interval*. If you are *well calibrated*, then on a test with 100 such questions, around two questions will have answers that fall outside your 98% confidence interval.

Alpert and Raiffa (1982) asked subjects a collective total of 1000 general knowledge questions like those described above; 426 of the true values lay outside the subjects' 98% confidence intervals. If the subjects were properly calibrated there would have been approximately twenty surprises. Put another way: Events to which subjects assigned a probability of 2% happened 42.6% of the time.

Another group of thirty-five subjects was asked to estimate 99.9% confident upper and lower bounds. They received 40% surprises. Another thirty-five subjects were asked for 'minimum' and 'maximum' values and were surprised 47% of the time. Finally, a fourth group of thirty-five subjects were asked for 'astonishingly low' and 'astonishingly high' values; they recorded 38% surprises.

In a second experiment, a new group of subjects was given a first set of questions, scored, provided with feedback, told about the results of previous experiments, had the concept of calibration explained to them at length, and then asked to provide 98% confidence intervals for a new set of questions. The post-training subjects were surprised 19% of the time, a substantial improvement over their pre-training score of 34% surprises, but still a far cry from the well-calibrated value of 2% surprises.

Similar failure rates have been found for experts. Hynes and Vanmarke (1976) asked seven internationally known geotechnical engineers to predict the height of an embankment that would cause a clay foundation to fail and to specify confidence bounds around this estimate that were wide enough to have a 50% chance of enclosing the true height. None of the bounds specified by the engineers enclosed the true failure height. Christensen-Szalanski and Bushyhead (1981) reported physician estimates for the probability of pneumonia for 1531 patients examined because of a cough. At the highest calibrated bracket of stated confidences, with average verbal probabilities of 88%, the proportion of patients actually having pneumonia was less than 20%.

In the words of Alpert and Raiffa (1982, p. 301): 'For heaven's sake, *Spread Those Extreme Fractiles!* Be honest with yourselves! Admit what you don't know!'

Lichtenstein et al. (1982) reviewed the results of fourteen papers on thirty-four experiments performed by twenty-three researchers studying human calibration. The *overwhelmingly* strong result was that people are overconfident. In the modern field, overconfidence is no longer noteworthy; but it continues to show up, in passing, in nearly any experiment where subjects are allowed to assign extreme probabilities.

Overconfidence applies forcefully to the domain of planning, where it is known as the *planning fallacy*. Buehler et al. (1994) asked psychology students to predict an important variable – the delivery time of their psychology honours thesis. They waited until students approached the end of their year-long projects, then asked the students when they realistically expected to submit their thesis and also when they would submit the thesis 'if everything went as poorly as it possibly could'. On average, the students took fifty-five days to complete their thesis, twenty-two days longer than they had anticipated, and seven days longer than their *worst-case* predictions.

Buehler et al. (1995) asked students for times by which they were 50% sure, 75% sure, and 99% sure that they would finish their academic project. Only 13% of the participants finished their project by the time assigned a 50% probability level, only 19% finished by the time assigned a 75% probability, and 45% finished by the time of their 99% probability level. Buehler et al. (2002) wrote: 'The results for the 99% probability level are especially striking: Even when asked to make a highly conservative forecast, a prediction that they felt virtually certain that they would fulfill, students' confidence in their time estimates far exceeded their accomplishments.'

Newby-Clark et al. (2000) found that asking subjects for their predictions based on realistic 'best guess' scenarios and asking subjects for their hoped-for 'best case' scenarios produced indistinguishable results. When asked for their 'most probable' case, people tend to envision everything going exactly as planned, with no unexpected delays or unforeseen catastrophes: the same vision as their 'best case'. *Reality, it turns out, usually delivers results somewhat worse than the 'worst case'.*

This chapter discusses overconfidence *after* discussing the confirmation bias and the sub-problem of the disconfirmation bias. The calibration research is dangerous knowledge – so tempting to apply selectively. 'How foolish my opponent is, to be so certain of his arguments! Doesn't he know how often people are surprised on their certainties?' If you realize that expert opinions have less force than you thought, you had better also realize that your own thoughts have *much* less force than you thought, so that it takes less force to compel you away from your preferred belief. Otherwise you become slower to react to incoming evidence. You are left worse off than if you had never heard of calibration. That is why – despite frequent great temptation – I avoid discussing the research on calibration unless I have previously spoken of the confirmation bias, so that I can deliver this same warning.

Note also that a confidently expressed expert opinion is quite a different matter from a calculation made *strictly* from actuarial data, or *strictly* from a *precise, precisely confirmed* model. Of all the times an expert has ever stated, even from strict calculation, that an event has a probability of 10^{-6}, they have undoubtedly been wrong more often than one time in a million. But if combinatorics could not correctly predict that a lottery ticket has a 10^{-8} chance of winning, ticket sellers would go broke.

5.11 Bystander apathy

My last bias comes, not from the field of heuristics and biases, but from the field of social psychology. A now-famous series of experiments by Latane and Darley (1969) uncovered the *bystander effect,* also known as *bystander apathy,* in which larger numbers of people are less likely to act in emergencies – not only individually, but also collectively. Among subjects alone in a room, on noticing smoke entering from under a door, 75% of them left the room to report it. When three naïve subjects were present, the smoke was reported only 38% of the time. A naive subject in the presence of two confederates who purposely ignored the smoke, even when the room became hazy, left to report the smoke only 10% of the time. A college student apparently having an epileptic seizure was helped 85% of the time by a single bystander and 31% of the time by five bystanders.

The bystander effect is usually explained as resulting from *diffusion of responsibility* and *pluralistic ignorance*. Being part of a group reduces individual responsibility. Everyone hopes that someone else will handle the problem instead, and this reduces the individual pressure to the point that no one does anything. Support for this hypothesis is adduced from manipulations in which subjects believe that the victim is especially dependent on them; this reduces the bystander effect or negates it entirely. Cialdini (2001) recommends that if you are ever in an emergency, you single out *one* bystander, and ask that person to help – thereby overcoming the diffusion.

Pluralistic ignorance is a more subtle effect. Cialdini (2001, p. 114) writes:

> Very often an emergency is not obviously an emergency. Is the man lying in the alley a heart-attack victim or a drunk sleeping one off? ... In times of such uncertainty, the natural tendency is to look around at the actions of others for clues. We can learn from the way the other witnesses are reacting whether the event is or is not an emergency. What is easy to forget, though, is that everybody else observing the event is likely to be looking for social evidence, too. Because we all prefer to appear poised and unflustered among others, we are likely to search for that evidence placidly, with brief, camouflaged glances at those around us. Therefore everyone is likely to see everyone else looking unruffled and failing to act.

The bystander effect is not about individual selfishness or insensitivity to the suffering of others. Alone subjects *do* usually act. Pluralistic ignorance can explain, and individual selfishness cannot explain, subjects failing to react to a room filling up with smoke. In experiments involving apparent dangers to either others or the self, subjects placed with non-reactive confederates frequently glance at the non-reactive confederates.

I am sometimes asked: 'If *existential risk X* is real, why aren't more people doing something about it?' There are many possible answers, a few of which I have touched on here. People may be overconfident and over-optimistic. They may focus on overly specific scenarios for the future, to the exclusion of all others. They may not recall any past extinction events in memory. They may overestimate the predictability of the past and hence underestimate the surprise of the future. They may not realize the difficulty of preparing for emergencies without benefit of hindsight. They may prefer philanthropic gambles with higher pay-off probabilities, neglecting the value of the stakes. They may conflate positive information about the benefits of a technology as negative information about its risks. They may be contaminated by movies, where the world ends up being saved. They may purchase moral satisfaction more easily by giving to other charities. Otherwise, the extremely unpleasant prospect of human extinction may spur them to seek arguments that humanity will *not* go extinct, without an equally frantic search for reasons why we *would*.

But if the question is, specifically, 'Why aren't more people doing something about it?', one possible component is that people are asking that very question – darting their eyes around to see if anyone else is reacting to the emergency, meanwhile trying to appear poised and unflustered. If you want to know why others are not responding to an emergency, before you respond yourself, you may have just answered your own question.

5.12 A final caution

Every true idea which discomforts you will seem to match the pattern of at least one psychological error.

Robert Pirsig said: 'The world's biggest fool can say the sun is shining, but that doesn't make it dark out.' If you believe someone is guilty of a psychological error, then demonstrate your competence by first demolishing their consequential factual errors. If there are no factual errors, then what matters the psychology? The temptation of psychology is that, knowing a little psychology, we can meddle in arguments where we have no technical expertise.

If someone wrote a novel about an asteroid strike destroying modern civilization, then the reader might criticize that novel as dystopian, apocalyptic, and symptomatic of the author's naïve inability to deal with a complex technological society. We should recognize this as a literary criticism and not a scientific one; it is about good or bad novels and not about good or bad hypotheses. To quantify the annual probability of an asteroid strike *in real life*, one *must* study astronomy and the historical record (while avoiding the case-specific biases; see Chapters 1 and 11, this volume). No amount of literary criticism can put a number on it. Garreau (2005) seems to hold that a scenario of a mind slowly increasing in capability, is more *mature* and *sophisticated* than a scenario of extremely rapid intelligence increase. But that is a technical question, not a matter of taste; no amount of psychologizing can tell you the exact slope of that curve.

It is harder to abuse heuristics and biases than psycho-analysis. Accusing someone of conjunction fallacy leads naturally into listing the specific details that you think are burdensome and drive down the joint probability. Even so, do not lose track of the real-world facts of primary interest; do not let the argument become *about* psychology.

Despite all dangers and temptations, it is better to know about psychological biases than not to know. Otherwise we will walk directly into the whirling helicopter blades of life. But be very careful not to have *too much fun* accusing others of biases. That is the road that leads to becoming a sophisticated arguer – someone who, faced with any discomforting argument, finds at once a bias in it; the one whom you must watch above all is yourself.

Jerry Cleaver said: 'What does you in is not failure to apply some high-level, intricate, complicated technique. It's overlooking the basics. Not keeping your eye on the ball.'

Analyses should finally *centre* on testable real-world assertions. Do not take your eyes off the ball.

5.13 Conclusion

Why should there be an organized body of thinking about global catastrophic and existential risks? Falling asteroids are not like engineered superviruses; physics disasters are not like nanotechnological wars. Why not consider each of these problems separately?

If someone proposes a physics disaster, then the committee convened to analyse the problem must obviously include physicists. But someone on that committee should also know how terribly dangerous it is to have an answer in your mind before you finish asking the question. Someone on that committee should remember the reply of Enrico Fermi to Leo Szilard's proposal that a fission chain reaction could be used to build nuclear weapons. (The reply was 'Nuts!' – Fermi considered the possibility so remote as to not be worth investigating.) Someone should remember the history of errors in physics calculations: the Castle Bravo nuclear test that produced a 15-megaton explosion, instead of four to eight, because of an unconsidered reaction in lithium-7: they correctly solved the wrong equation, failed to think of all the terms that needed to be included, and at least one person in the expanded fallout radius died. Someone should remember Lord Kelvin's careful proof, using multiple, independent quantitative calculations from well-established theories, that the Earth could not possibly have existed for as much as 40 million years. Someone should know that when an expert says the probability is 'a million to one' without using actuarial data or calculations from a precise, precisely confirmed model, the calibration is probably more like 20 to 1 (although this is not an exact conversion).

Any existential risk evokes problems that it shares with all other existential risks, *in addition to* the domain-specific expertise required for the *specific* existential risk (see more details in Chapter 1, this book). Someone on the physics-disaster committee should know what the term 'existential risk' means and should possess whatever skills the field of existential risk management has accumulated or borrowed. For maximum safety, that person should also be a physicist. The domain-specific expertise and the expertise pertaining to existential risks should combine in one person. I am sceptical that a scholar of heuristics and biases, unable to read physics equations, could check the work of physicists who knew nothing of heuristics and biases.

Once upon a time I made up overly detailed scenarios, without realizing that *every* additional detail was an extra burden. I really did think that I could say there was a 90% chance of Artificial Intelligence being developed between 2005 and 2025, with the peak in 2018. This statement now seems to me like complete gibberish. Why did I *ever* think I could generate a tight probability distribution over a problem like that? Where did I even get those numbers in the first place?

I once met a lawyer who had made up his own theory of physics. I said to the lawyer: 'You cannot invent your own physics theories without knowing math and studying for years; physics is hard.' He replied: 'But if you really understand physics you can explain it to your grandmother, Richard Feynman told me so.' And I said to him: 'Would you advise a friend to argue his own court case?' At this he fell silent. He knew abstractly that physics was difficult, but I think it had honestly never occurred to him that physics might be as difficult as lawyering.

One of many biases *not* discussed in this chapter describes the biasing effect of not knowing what we do not know. When a company recruiter evaluates his own skill, he recalls in his mind the performance of candidates he hired, many of whom subsequently excelled; therefore the recruiter thinks highly of his skill. But the recruiter never sees the work of candidates *not* hired. Thus I must warn that this paper touches upon only a small subset of heuristics and biases; for when you wonder how much you have already learned, you will recall the few biases this chapter *does* mention, rather than the many biases it does not. Brief summaries cannot convey a sense of the field, the larger understanding which weaves a set of memorable experiments into a unified interpretation. Many highly relevant biases, such as *need for closure*, I have not even mentioned. The purpose of this chapter is not to teach the knowledge needful to a student of existential risks but to intrigue you into learning more.

Thinking about existential risks falls prey to all the same fallacies that prey upon thinking in general. But the stakes are much, much higher. A common result in heuristics and biases is that offering money or other incentives does not eliminate the bias. (Kachelmeier and Shehata [1992] offered subjects living in the People's Republic of China the equivalent of three months' salary.) The subjects in these experiments do not make mistakes on purpose; they make mistakes because they do not know how to do better. Even if you told them the survival of humankind was at stake, they still would not thereby know how to do better. (It might increase their need for closure, causing them to do worse.)

It is a terribly frightening thing, but people do not become any smarter, *just* because the survival of humankind is at stake.

In addition to standard biases, I have personally observed what look like harmful modes of thinking specific to existential risks. The Spanish flu of 1918 killed 25–50 million people. World War II killed 60 million people; 10^7 is the order of the largest catastrophes in humanity's written history. Substantially larger numbers, such as 500 million deaths, and *especially* qualitatively different scenarios such as the extinction of the entire human species, seem to trigger a *different mode of thinking* – enter into a 'separate magisterium'. People who would never dream of hurting a child hear of an existential risk, and say, 'Well, maybe the human species doesn't really deserve to survive.'

There is a saying in heuristics and biases that people do not evaluate events, but descriptions of events – what is called non-extensional reasoning. The *extension* of humanity's extinction includes the death of yourself, your friends, your family, your loved ones, your city, your country, your political fellows. Yet people who would take great offence at a proposal to wipe the country of Britain from the map, to kill every member of the Democratic Party in the United States, to turn the city of Paris to glass – who would feel still greater horror on hearing the doctor say that their child had cancer – these people will discuss the extinction of humanity with perfect calm. The phrase 'extinction of humanity', as words on paper, appears in fictional novels or is discussed in philosophy books – it belongs to a different context compared to the Spanish flu. We evaluate descriptions of events, not extensions of events. The cliché phrase *end of the world* invokes the magisterium of myth and dream, of prophecy and apocalypse, of novels and movies. The challenge of existential risks to rationality is that, the catastrophes being so huge, people snap into a different mode of thinking. Human deaths are suddenly no longer bad, and detailed predictions suddenly no longer require any expertise, and whether the story is told with a happy ending or a sad ending is a matter of personal taste in stories.

But that is only an anecdotal observation of mine. I thought it better that this essay should focus on mistakes well documented in the literature – the general literature of cognitive psychology, because there is not yet experimental literature specific to the psychology of existential risks. There should be.

In the mathematics of Bayesian decision theory there is a concept of *information value* – the expected utility of knowledge. The value of information emerges from the value of whatever it is *about*; if you double the stakes, you double the value of information about the stakes. The value of rational thinking works similarly – the value of performing a computation that integrates the evidence is calculated much the same way as the value of the evidence itself (Good, 1952; Horvitz et al., 1989).

No more than Albert Szent-Györgyi could multiply the suffering of one human by a 100 million can I truly understand the value of clear thinking about global risks. Scope neglect is the hazard of being a biological human, running on an analogous brain; the brain cannot multiply by 6 billion. And the stakes of existential risk extend beyond even the 6 billion humans alive today, to all the stars in all the galaxies that humanity and humanity's descendants may some day touch. All that vast potential hinges on our survival here, now, in the days when the realm of humankind is a single planet orbiting a single star. I cannot feel our future. All I can do is try to defend it.

Acknowledgement

I thank Michael Roy Ames, Olie Lamb, Tamas Martinec, Robin Lee Powell, Christian Rovner, and Michael Wilson for their comments, suggestions, and criticisms. Needless to say, any remaining errors in this chapter are my own.

Suggestions for further reading

Dawes, R. (1988). *Rational Choice in an Uncertain World: The Psychology of Intuitive Judgment* (San Diego, CA: Harcourt, Brace, Jovanovich). First edition 1988 by Dawes and Kagan, second edition 2001 by Dawes and Hastie. This book aims to introduce heuristics and biases to an intelligent general audience. (For example, Bayes's Theorem is explained, rather than assumed, but the explanation is only a few pages.) A good book for quickly picking up a sense of the field.

Kahneman, D., Slovic, P., and Tversky, A. (eds.) (1982). *Judgment Under Uncertainty: Heuristics and Biases* (New York: Cambridge University Press). This is the edited volume that helped establish the field, written with the outside academic reader firmly in mind. Later research has generalized, elaborated, and better explained the phenomena treated in this volume, but the basic results given are still standing strong.

Kahneman, D. and Tversky, A. (eds.) (2000). *Choices, Values, and Frames* (Cambridge: Cambridge University Press). Gilovich, T. Griffin, D. and Kahneman, D. (2003). *Heuristics and Biases.* These two edited volumes overview the field of heuristics and biases in its current form. They are somewhat less accessible to a general audience.

References

Alpert, M. and Raiffa, H. (1982). A progress report on the training of probability assessors. In Kahneman, D., Slovic, P., and Tversky, A. (eds.), *Judgement Under Uncertainty: Heuristics and Biases*, pp. 294–305 (Cambridge: Cambridge University Press).

Ambrose, S.H. (1998). Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans. *J. Human Evol.*, **34**, 623–651.

Baron, J. and Greene, J. (1996). Determinants of insensitivity to quantity in valuation of public goods: contribution, warm glow, budget constraints, availability, and prominence. *J. Exp. Psychol.: Appl.*, **2**, 107–125.

Bostrom, N. (2001). Existential risks: analyzing human extinction scenarios. *J. Evol. Technol.*, **9**.

Brenner, L.A., Koehler, D.J., and Rottenstreich, Y. (2002). Remarks on support theory: recent advances and future directions. In Gilovich, T., Griffin, D., and Kahneman, D. (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*, pp. 489–509 (Cambridge: Cambridge University Press).

Buehler, R., Griffin, D., and Ross, M. (1994). Exploring the 'planning fallacy': why people underestimate their task completion times. *J. Personal. Social Psychol.*, **67**, 366–381.

Buehler, R., Griffin, D., and Ross, M. (1995). It's about time: optimistic predictions in work and love. *Eur. Rev. Social Psychol.*, **6**, 1–32.

Buehler, R., Griffin, D., and Ross, M. (2002). Inside the planning fallacy: the causes and consequences of optimistic time predictions. In Gilovich, T., Griffin, D., and Kahneman, D. (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*, pp. 250–270 (Cambridge: Cambridge University Press).

Burton, I., Kates, R., and White, G. (1978). *Environment as Hazard* (New York: Oxford University Press).

Carson, R.T. and Mitchell, R.C. (1995). Sequencing and nesting in contingent valuation surveys. *J. Environ. Econ. Manag.*, **28**(2), 155–173.

Chapman, G.B. and Johnson, E.J. (2002). Incorporating the irrelevant: anchors in judgments of belief and value. In Gilovich, T., Griffin, D., and Kahneman, D. (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*, pp. 120–138 (Cambridge: Cambridge University Press.

Christensen-Szalanski, J.J.J. and Bushyhead, J.B. (1981). Physicians' use of probabilistic information in a real clinical setting. *J. Exp. Psychol. Human Percept. Perf.*, **7**, 928–935.

Cialdini, R.B. (2001). *Influence: Science and Practice* (Boston, MA: Allyn and Bacon).

Combs, B. and Slovic, P. (1979). Causes of death: Biased newspaper coverage and biased judgments. *Journalism Quarterly*, **56**, 837–843.

Dawes, R.M. (1988). *Rational Choice in an Uncertain World* (San Diego, CA: Harcourt, Brace, Jovanovich).

Desvousges, W.H., Johnson, F.R., Dunford, R.W., Boyle, K.J., Hudson, S.P., and Wilson, N. (1993). Measuring natural resource damages with contingent valuation: tests of validity and reliability. In Hausman, J.A. (ed.), *Contingent Valuation: A Critical Assessment*, pp. 91–159 (Amsterdam: North Holland).

Fetherstonhaugh, D., Slovic, P., Johnson, S., and Friedrich, J. (1997). Insensitivity to the value of human life: a study of psychophysical numbing. *J. Risk Uncertainty*, **14**, 238–300.

Finucane, M.L., Alhakami, A., Slovic, P., and Johnson, S.M. (2000). The affect heuristic in judgments of risks and benefits. *J. Behav. Decision Making*, **13**(1), 1–17.

Fischhoff, B. (1982). For those condemned to study the past: heuristics and biases in hindsight. In Kahneman, D., Slovic, P., and Tversky, A. (eds.), *Judgement Under*

Uncertainty: Heuristics and Biases, pp. 306–354 (Cambridge: Cambridge University Press).

Fischhoff, B. and Beyth, R. (1975). I knew it would happen: remembered probabilities of once-future things. *Organ. Behav. Human Perf.*, **13**, 1–16.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1977). Knowing with certainty: the appropriateness of extreme confidence. *J. Exp. Psychol Human Percept. Perf.*, **3**, 522–564.

Ganzach, Y. (2001). Judging risk and return of financial assets. *Organ. Behav. Human Decision Processes*, **83**, 353–370.

Garreau, J. (2005). *Radical Evolution: The Promise and Peril of Enhancing Our Minds, Our Bodies – and What It Means to Be Human* (New York: Doubleday).

Gilbert, D.T. and Osborne, R.E. (1989). Thinking backward: Some curable and incurable consequences of cognitive busyness. *J. Person. Social Psychol.*, **57**, 940–949.

Gilbert, D.T., Pelham, B.W., and Krull, D.S. (1988). On cognitive busyness: when person perceivers meet persons perceived. *J. Person. Social Psychol.*, **54**, 733–740.

Gilovich, T. (2000). *Motivated Skepticism and Motivated Credulity: Differential Standards of Evidence in the Evaluation of Desired and Undesired Propositions*. Presented at the 12th Annual Convention of the American Psychological Society, Miami Beach, FL.

Gilovich, T., Griffin, D., and Kahneman, D. (eds.) (2003). *Heuristics and Biases: The Psychology of Intuitive Judgment* (Cambridge: Cambridge University Press).

Good, I.J. (1952). Rational decisions. *J. Royal Statist. Soc., Series B*, **14**, 107–114.

Griffin, D. and Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cogn. Psychol.*, **24**, 411–435.

Harrison, G.W. (1992). Valuing public goods with the contingent valuation method: a critique of Kahneman and Knestch. *J. Environ. Econ. Manag.*, **23**, 248–257.

Horvitz, E.J., Cooper, G.F., and Heckerman, D.E. (1989). Reflection and action under scarce resources: theoretical principles and empirical study. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pp. 1121–1127 (Detroit, MI).

Hynes, M.E. and Vanmarke, E.K. (1976). Reliability of embankment performance predictions. *Proceedings of the ASCE Engineering Mechanics Division Specialty Conference* (Waterloo, Ontario: University of Waterloo Press).

Johnson, E., Hershey, J., Meszaros, J., and Kunreuther, H. (1993). Framing, probability distortions and insurance decisions. *J. Risk Uncertainty*, **7**, 35–51.

Kachelmeier, S.J. and Shehata, M. (1992). Examining risk preferences under high monetary incentives: experimental evidence from the People's Republic of China. *Am. Econ. Rev.*, **82**, 1120–1141.

Kahneman, D. (1986). Comments on the contingent valuation method. In Cummings, R.G., Brookshire, D.S., and Schulze, W.D. (eds.), *Valuing Environmental Goods: A State of the Arts Assessment of the Contingent Valuation Method*, pp. 185–194 (Totowa, NJ: Roweman and Allanheld).

Kahneman, D. and Knetsch, J.L. (1992). Valuing public goods: the purchase of moral satisfaction. *J. Environ. Econ. Manag.*, **22**, 57–70.

Kahneman, D., Ritov, I., and Schkade, D.A. (1999). Economic preferences or attitude expressions?: An analysis of dollar responses to public issues. *J. Risk Uncertainty*, **19**, 203–235.

Kahneman, D., Slovic, P., and Tversky, A. (eds.) (1982). *Judgment Under Uncertainty: Heuristics and Biases* (New York: Cambridge University Press).

Kahneman, D. and Tversky, A. (eds.) (2000). *Choices, Values, and Frames* (Cambridge: Cambridge University Press).

Kamin, K. and Rachlinski, J. (1995). Ex post ≠ ex ante: determining liability in hindsight. *Law Human Behav.*, **19**(1), 89–104.

Kates, R. (1962). Hazard and choice perception in flood plain management. Research Paper No. 78. Chicago, IL: University of Chicago, Department of Geography.

Knaup, A. (2005). Survival and longevity in the business employment dynamics data. *Monthly Labor Rev.*, **May** 2005, 50–56.

Kunda, Z. (1990). The case for motivated reasoning. *Psychol. Bull.*, **108**(3), 480–498.

Latane, B. and Darley, J. (1969). Bystander 'Apathy'. *Am. Scientist*, **57**, 244–268.

Lichtenstein, S., Fischhoff, B., and Phillips, L.D. (1982). Calibration of probabilities: the state of the art to 1980. In Kahneman, D., Slovic, P., and Tversky, A. (eds.),*Judgment Under Uncertainity: Heuristics and Biases*, pp. 306–334 (New York: Cambridge University Press).

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., and Combs, B. (1978). Judged frequency of lethal events. *J. Exp. Psychol.: Human Learn. Memory*, **4**(6), 551–578.

McFadden, D. and Leonard, G. (1995). Issues in the contingent valuation of environmental goods: methodologies for data collection and analysis. In Hausman, J.A. (ed.), *Contingent Valuation: A Critical Assessment*, pp. 165–208 (Amsterdam: North Holland).

Newby-Clark, I.R., Ross, M., Buehler, R., Koehler, D.J., and Griffin, D. (2000). People focus on optimistic and disregard pessimistic scenarios while predicting their task completion times. *J. Exp. Psychol. Appl.*, **6**, 171–182.

Quattrone, G.A., Lawrence, C.P., Finkel, S.E., and Andrus, D.C. (1981). Explorations in anchoring: the effects of prior range, anchor extremity, and suggestive hints. Manuscript, Stanford University.

Rasmussen, N.C. (1975). *Reactor Safety Study: An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants.* NUREG-75/014, WASH-1400 (Washington, DC: U.S. Nuclear Regulatory Commission)

Rogers, W.P., Armstrong, N., Acheson, D.C., Covert, E.E., Feynman, R.P., Hotz, R.B., Kutyna, D.J., Ride, S.K., Rummel, R.W., Suffer, D.F., Walker, A.B.C., Jr., Wheelon, A.D., Yeager, C., Keel, A.G., Jr. (1986). Report of the Presidential Commission on the Space Shuttle Challenger Accident. *Presidential Commission on the Space Shuttle Challenger Accident.* Washington, DC.

Sanchiro, C. (2003). Finding Error. *Mich. St. L. Rev.* 1189.

Schneier, B. (2005). Security lessons of the response to hurricane Katrina. http://www.schneier.com/blog/archives/2005/09/security_lesson.html. Viewed on 23 January 2006.

Sides, A., Osherson, D., Bonini, N., and Viale, R. (2002). On the reality of the conjunction fallacy. *Memory Cogn.*, **30**(2), 191–198.

Slovic, P., Finucane, M., Peters, E., and MacGregor, D. (2002). Rational actors or rational fools: implications of the affect heuristic for behavioral economics. *J. Socio-Econ.*, **31**, 329–342.

Slovic, P., Fischoff, B., and Lichtenstein, S. (1982). Facts versus fears: understanding perceived risk. In Kahneman, D., Slovic, P., and Tversky, A. (eds.), *Judgment Under Uncertainity: Heuristics and Biases*, pp. 463–492 (Cambridge: Cambridge University Press).

Strack, F. and Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: mechanisms of selective accessibility. *J. Person. Social Psychol.*, **73**, 437–446.

Taber, C.S. and Lodge, M. (2000). Motivated skepticism in the evaluation of political beliefs. Presented at the 2000 meeting of the American Political Science Association.

Taleb, N. (2001). *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*, pp. 81–85 (New York: Textre).

Taleb, N. (2005). *The Black Swan: Why Don't We Learn that We Don't Learn?* (New York: Random House).

Tversky, A. and Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cogn. Psychol.*, **4**, 207–232.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, **185**, 251–284.

Tversky, A. and Kahneman, D. (1982). Judgments of and by representativeness. In Kahneman, D., Slovic, P., and Tversky, A. (eds.), *Judgement Under Uncertainty: Heuristics and Biases*, pp. 84–98 (Cambridge: Cambridge University Press).

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.*, **90**, 293–315.

Wansink, B., Kent, R.J., and Hoch, S.J. (1998). An anchoring and adjustment model of purchase quantity decisions. *J. Market. Res.*, **35**(February), 71–81.

Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly J. Exp. Psychol.*, **12**, 129–140.

Wilson, T.D., Houston, C., Etling, K.M., and Brekke, N. (1996). A new look at anchoring effects: basic anchoring and its antecedents. *J. Exp. Psychol.: General*, **4**, 387–402.