Information Hazards: A Typology of Potential Harms from Knowledge

Nick Bostrom

Faculty of Philosophy & Oxford Martin School

Oxford University

[Published in *Review of Contemporary Philosophy*, Vol. 10 (2011): pp. 44-79]

www.nickbostrom.com

Abstract

Information hazards are risks that arise from the dissemination or the potential dissemination of true information that may cause harm or enable some agent to cause harm. Such hazards are often subtler than direct physical threats, and, as a consequence, are easily overlooked. They can, however, be important. This paper surveys the terrain and proposes a taxonomy.

1. Introduction

There is, these days, a commonly held presumption in favor of knowledge, truth, and the uncovering and dissemination of information. It is rare to find somebody who self-identifies as an obscurantist or who openly espouses obscurantism as a legitimate policy instrument of wide utility.

Even reactionaries rarely object to this general favoring of information. Consider some particularly intransigent creationist who opposes the teaching of evolution theory in public schools. He does not constitute a counterexample. For he does not believe that evolution theory is a truth to be concealed. Rather, he believes evolution theory an error that ought to be replaced with more accurate information. Therefore, although he happens unwittingly to stand in the way of truth, he need not disagree with the claim that the truth should be promoted. The creationist, too, is a truth-lover, albeit one whose affections are unreciprocated.

Although nobody makes a brief for ignorance generally, there are many special cases in which ignorance is cultivated—in order, for example, to protect national security, sexual innocence, jury impartiality; to preserve anonymity for patients, clients, reviewers, and voters; to create suspense in films and novels; to protect trade secrets; to measure the placebo effect and avoid various

research biases; and to create mental challenges for gaming and study. These cases are commonly accepted exceptions to the general rule of knowledge-favoring.[1]

In this paper, we will not be concerned with postmodernist critiques of the idea of objective truth nor with skeptical doubts about the possibility of knowledge. I shall assume some broad commonsensical understanding according to which there are truths and we humans sometimes manage to know some of these truths.

This paper will also not discuss the ways in which harm can be caused by *false* information. Many of those ways are obvious. We can be harmed, for instance, by false information that misleads us into believing that some carcinogenic pharmaceutical is safe; or, alternatively, that some safe pharmaceutical is carcinogenic. We will limit our investigation to the ways in which the discovery and dissemination of *true* information can be harmful.

Let us define

> *Information hazard*: A risk that arises from the dissemination or the potential dissemination of (true) information that may cause harm or enable some agent to cause harm.[2]

Relative to their significance, and compared to many direct physical dangers, some types of information hazard are unduly neglected. It will therefore be useful to distinguish some different types of information hazard. This will serve to draw attention to some easily-overlooked risks and will help us create a vocabulary for discussing them.

The aim of this paper is to catalogue some of the various possible ways in which information can cause harm. We will not here seek to determine how common and serious these harms are or how they stack up against the many benefits of information—questions that would need to be engaged before one could reach a considered position about potential policy implications. It may be worth stressing, however, that even if one has an extremely strong intellectual commitment to truth-seeking and public education, one can still legitimately and in good conscience explore the question of how some knowledge might be harmful. In fact, this very commitment demands that one does not shy away from such an exploration or from reporting openly on the findings.

2. Six information transfer modes

[1]The notion of dangerous or forbidden knowledge is also a common trope in literature and in many mythologies; see e.g. Shattuck 1996.

[2] We set aside the trivial way in which utterances can cause harm via their performative functions; cf. Austin 1962. Thus, a dictator who proclaims "Invade that country!" disseminates information than can obviously cause a lot of harm, but we shall not count this possibility as an information hazard.

We can distinguish several different information formats, or "modes" of information transfer. Each can be associated with risk. Perhaps most obviously, we have:

> *Data hazard*: Specific data, such as the genetic sequence of a lethal pathogen or a blueprint for making a thermonuclear weapon, if disseminated, create risk.[3]

But also:

> *Idea hazard*: A general idea, if disseminated, creates a risk, even without a data-rich detailed specification.

For example, the idea of using a fission reaction to create a bomb, or the idea of culturing bacteria in a growth medium with an antibiotic gradient to evolve antibiotic resistance, may be all the guidance a suitably prepared developer requires; the details can be figured out. Sometimes the mere demonstration that something (such as a nuclear bomb) is possible provides valuable information which can increase the likelihood that some agent will successfully set out to replicate the achievement.

Even if the relevant ideas and data are already "known", and published in the open literature, an increased risk may nonetheless be created by drawing attention to a particularly potent possibility.

> *Attention hazard*: The mere drawing of attention to some particularly potent or relevant ideas or data increases risk, even when these ideas or data are already "known".

Because there are countless avenues for doing harm, an adversary faces a vast search task in finding out which avenue is most likely to achieve his goals. Drawing the adversary's attention to a subset of especially potent avenues can greatly facilitate the search. For example, if we focus our concern and our discourse on the challenge of defending against viral attacks, this may signal to an adversary that viral weapons—as distinct from, say, conventional explosives or chemical weapons—constitute an especially promising domain in which to search for destructive applications. The better we manage to focus our defensive deliberations on our greatest vulnerabilities, the more useful our conclusions may be to a potential adversary.

It would be a mistake to suppose that because some idea is already in the public domain there can be no further harm in discussing the idea and referring to it in other publications. Such further discussions can create or aggravate an attention hazard by increasing the idea's salience. One index of how much mileage there is in an idea is the amount "buzz" surrounding it.

Clumsy attempts to suppress discussion often backfire. An adversary who discovers an attempt to conceal an idea may infer that the idea could be of great value. Secrets have a special allure. The

[3] The term "data hazard" also has a narrow technical meaning in the context of computer processor design, which is not relevant here.

adversary may invest more in understanding and exploring an idea that he knows that his enemy is trying hard to keep secret. A book that is censored often becomes more widely read.[4]

It is possible that efforts to contemplate some risk area—say, existential risk—will do more harm than good. One might suppose that *thinking* about a topic should be entirely harmless, but this is not necessarily so. If one gets a good idea, one will be tempted to share it; and in so doing one might create an information hazard. Still, one likes to believe that, on balance, investigations into existential risks and most other risk areas will tend to reduce rather than increase the risks of their subject matter.

Sometimes it is right to harm. If information enables an agent to harm rightly, this can be a good thing; and the possibility of that happening should therefore not be classified as a risk. For example, the police's obtainment of certain information might harm some criminal by enabling the police to arrest him; and this can be good. However, we could say that from the criminal's point of view there is an information hazard. He faces a risk that his whereabouts will be reported.

Not all types of information transfer are best thought of in terms of data, ideas, or attention. We can also consider implicit forms of information, such as processes or organizational structures, which can give one firm an advantage over another, and which might be imitated or replicated by a competitor.[5] Similarly, individuals often learn, and shape their own personality, by "emulating" some other person. Such emulation can happen unintentionally and even without awareness that emulation is taking place. A bad role model can pose a template hazard.

> *Template hazard*: The presentation of a template enables distinctive modes of information transfer and thereby creates risk.

We can also register as a distinct mode of communication social signaling, where the focus is not on the content that is being transmitted but on what this content—and the fact that it is being communicated—says about the sender. Non-verbal actions can also serve a social signaling role, if they are aimed not only at directly achieving the some practical outcome to which they are nominally geared but also to signal some hidden quality of the agent that performs the action. For example, one could give alms to the poor not only because one wishes to satisfy their needs but also because one wishes to be seen by one's peers as a kind and generous person. One might utter patriotic statements not only to convey to the listener information about various attributes of one's nation, but also to signal one's status as a loyal citizen, or one's affiliation with some political grouping.

> *Signaling hazard*: Verbal and non-verbal actions can indirectly transmit information about some hidden quality of the sender, and such social signaling creates risk.

---

[4] A good example of this is the rather dull *Spycatcher* by Peter Wright, which became a bestseller in the 1980s after Thatcher tried to ban it; see Zuckerman 1987.

[5] Nelson and Winter 1990 and Porter 2004.

Some topics are especially attractive to crackpots. Serious academics might shy away from such topics because they fear that working on those topics signals intellectual flakiness. At least two signaling hazards arise in this context. One is the risk to individual thinkers who might suffer undeserved reputational damage merely for working in an area which also happens to attract lots of crackpots. Another is the risk to society that important areas of research will remain uncultivated because the ablest researchers (and their sponsors) protect their reputations either by shunning those areas in favor of more socially acceptable, high-status fields of study or by adopting relatively less effective means of exploration, such as hypertrophic formalism and expensive technical apparatus, which are harder for crackpots to mimic.

Finally, we also count as a distinct mode the transfer of information contained in the particular way some content is formulated and expressed. A vivid description of some event, for example, can activate psychological processes that lie dormant when the same event is recounted in dry prose.

> *Evocation hazard*: There can be a risk that the particular mode of presentation used to convey some content can activate undesirable mental states and processes.

Each of these information transfer modes—data, idea, attention, template, signaling, and evocation—can play a role in creating various kinds of risk. The latter four, especially, are easily overlooked.

The following five sections introduce another categorization scheme which, when superimposed upon the division into information transfer modes, renders a more fine-grained picture of the ways in which information can be hazardous (summarized in table 1).

TYPOLOGY OF INFORMATION HAZARDS		
I. By information transfer mode		
	Data hazard	
	Idea hazard	
	Attention hazard	
	Template hazard	
	Signaling hazard	
	Evocation hazard	
II. By effect		
	TYPE	SUBTYPE
ADVERSARIAL RISKS	*Competiveness hazard*	*Enemy hazard*
		Intellectual property hazard
		Commitment hazard
		Knowing-too-much hazard
RISKS TO SOCIAL ORGANIZATON AND MARKETS	*Norm hazard*	*Information asymmetry hazard*
		Unveiling hazard
		Recognition hazard
RISKS OF IRRATIONALITY	*Ideological hazard*	

AND ERROR	*Distraction and temptation hazards*	
	Role model hazard	
	Biasing hazard	
	De-biasing hazard	
	Neuropsychological hazard	
	Information-burying hazard	
RISKS TO VALUABLE STATES AND ACTIVITIES	*Psychological reaction hazard*	*Disappointment hazard*
		Spoiler hazard
		Mindset hazard
	Belief-constituted value hazard	
	(mixed)	*Embarrassment hazard*
RISKS FROM INFORMATION TECHNOLOGY SYSTEMS	*Information system hazard*	*Information infrastructure failure hazard*
		Information infrastructure misuse hazard
		Artificial intelligence hazard
RISKS FROM DEVELOPMENT	*Development hazard*	

Table 1

3. Adversarial risks

Previous examples focused on adversarial situations in which some foe is intent on causing us harm. A burglar who knows where we keep our money and when we will return home is in a stronger position to succeed with his crime.

> *Enemy hazard*: By obtaining information our enemy or potential enemy becomes stronger and this increases the threat he poses to us.

One paradigmatic context for this type of hazard is national security. Within the defense sector, activities aimed at reducing enemy information hazard range from counter-intelligence work to the application of camouflage to conceal troops in the field.

Enemy hazard depends on the existence of valuable information that an enemy might obtain. Indirectly, therefore, our own activities can be hazardous if they contribute to the production of such information. Military research offers many examples. We invest in research and development of new weapons and new tactics. This activity produces information that is valuable to our enemy. The enemy observes our improved tactics. His spies obtain the blueprints of our improved technology. Or the relevant information leaks out in other ways, perhaps in the form of ideas, attention, or templates. As a consequence, our enemy soon replicates our achievement. When hostilities erupt, we battle our own inventions.

Rational strategy for military research would give significant consideration to such effects. The United States, for example, might refrain from aggressively pursuing development of electromagnetic pulse weapons. Because of the country's unusually heavy reliance on electronics, the existence of effective EMP weapons would differentially benefit its adversaries.

Conversely, an aggressive approach to research could help protect a country against its enemies. A country might pursue military research to catch up with the technological leader. A leader in military technology might invest in research in order to maintain its lead. The leader might reason that, while its discoveries will eventually leak out and benefit its enemies, it can produce a steady stream of discoveries and continually keep a few steps ahead.

There are situations, though, in which a country is better off retarding its progress in military technology, even when the new technologies would not differentially benefit its enemies and even if considerations such as cost and foregone opportunities for building international trust are set aside. Suppose some country has great superiority in military power and military technology, and as a consequence faces little serious threat. By investing heavily in military research, it could increase its lead and thereby further enhance its security somewhat. Yet such investment might undermine security in the longer term. The rate of information leakage might be a function of the size of the technological gap such that increasing the gap increases the rate of the leakage. Perhaps weapons systems come in "generations" and it may be infeasible to keep secret more than about one generation beyond the enemy's level. If so, introducing new generations at a faster rate might not increase the technological lead, but serve only to accelerate both countries' ascent of the military technology tree, perhaps to levels where wars become more destructive. If you are already utterly superior in conventional weapons, then rushing to invent a fission bomb long before your enemies could have got there might be counterproductive. Similarly, hastening the introduction of the fusion bomb might be bad strategy if you could have been confident of remaining superior in fission bombs.

Accelerating the ascent of the technology tree could also be bad if the tree is of finite height, so that at some point the leader runs out of opportunities for innovation. Some weapons systems might reach a level of perfection from which further improvements are difficult or impossible. (In the category of weapons of mass destruction for deterrent use, for instance, the hydrogen bomb might represent a near-maximum.) Eventually everybody may plateau at this level, and the previous leader will lose its advantage. In order to maintain a technological lead for as long as possible, the leader might wish to push the technological frontier at the slowest possible pace that is consistent with maintaining an acceptable lead at every point in time until technological maturity is reached.

The military illustration shows how information hazards arise in some situations from one party's (potential) intent to inflict harm on another. However, information risks stemming from an adversarial relationship arise much more widely. In competitive situations, one person's information can cause harm to another even if no intention to cause harm is present. Example: The rival job applicant knew more and got the job.

> *Competiveness hazard*: There is a risk that, by obtaining information, some competitor of ours will become stronger, thereby weakening our competitive position.

Exclusive possession of information is central to the business model of many firms. A competitor can gain valuable information by observing the production and marketing methods of a rival firm, reverse-engineering its products, or headhunting its employees.[6] Firms go to great lengths to protect their intellectual capital, relying on a wide variety of methods including patenting, copyright, non-disclosure agreements, physical security, in-house production instead of outsourcing, compensation schemes that discourage employee turnover, and so forth. We can identify threat to intellectual property as a special case of competitiveness hazard:

> *Intellectual property hazard*: *A* faces the risk that some other firm *B* will obtain *A*'s intellectual property, thereby weakening *A*'s competitive position.

Another type of adversarial hazard arises when an agent's own possession of some information has the potential to render her less able to prevail in some competitive situation. In order for a blackmailer to be successful, his target must believe that he has some incriminating or embarrassing information, information that he could threaten to release. So long as the target remains unaware, no blackmail can take place. When she learns about the threat, she is delivered into the extortionist's clutches. Similarly, in the game of chicken: Two drivers speed towards one another from opposite directions; the first to swerve loses. If one driver could credibly commit to not swerving, he would win, since his opponent (it is assumed) would rather lose the game than crash. A game theorist engaging in this pastime could have himself blindfolded, preventing himself from acquiring information about the relative distance of the two cars, thus rendering himself incapable of reliably swerving at the last possible moment; and thereby convincing his (hopefully rational) counterpart to yield the road.

> *Commitment hazard*: There is a risk that the obtainment of some information will weaken one's ability credibly to commit to some course of action.

Commitment hazards are sometimes instances of enemy hazard and sometimes of competitiveness hazards. (Commitment hazards can also arise in a single-agent context, as we shall see later.)

In some situations it can be advantageous to make a probabilistic threat, a "threat that leaves something to chance" in the terminology of Thomas Shelling.[7] A threat, to be effective, must be credible. Yet the reason for issuing a threat is deterrence *ex ante*, not revenge *ex post*; and carrying out a threat is often costly. Consider some possible punitive action that is so costly to carry out that a threat to do so would scarcely be credible, such as a nuclear first strike by one major power on another. A nuclear arsenal could nevertheless be used to make a threat. Side *A* can threaten that unless the side *B* makes some concession, *A* will take some action that *increases the risk* of nuclear war. For instance, *A* could threaten to initiate skirmishes with conventional weapons that would create some risk of escalation. Alternatively, *A* could threaten to make its own nuclear control and command system less safe against accidental launch, for instance by creating a crisis and putting its forces on high alert. The idea here is that it is much less costly for *A* to carry out a threat to

[6] Porter 2004.

[7] Schelling 1981.

moderately increase the risk of a nuclear war than it would be for *A* to actually launch a nuclear attack. The probabilistic threat can therefore be more credible and more effective.

If, however, new information came to light that dispelled the requisite uncertainty—uncertainty, for example, about how key actors would react during a crisis—then the ability to make probabilistic threats would be undermined. The possibility of such information being released can thus constitute a kind of information hazard. During the Cold War, kindred considerations may have led the superpowers to maintain some ambiguity in their strategic postures. This kind of information hazard might involve a combination of enemy hazard and commitment hazard.[8]

We can also identify another type of risk that can arise from our own knowledge when there is a possibility that somebody else will become our adversary because of this knowledge:

> *Knowing-too-much hazard*: Our possessing some information makes us a potential target or object of dislike.

Nadezhda Sergeevna Alliluyeva, Stalin's second wife, was known to have misgivings about the Communist Party purges and the concomitant famine. Following a spat with Stalin in 1932, she was found dead in her bedroom, apparently having taken her own life.[9] The suicide could be interpreted as a kind of desperate rebuke of Stalin and his policies; and since that would be politically embarrassing, the death was officially attributed to appendicitis. The individuals who had discovered the body and who knew the real cause of death found themselves in grave danger. In a later allegedly unrelated trial, two doctors who had declined to sign the false death certificate were sentenced to death and executed.

In the witch hunts of the Early Modern period in Europe, a woman's alleged possession of knowledge of the occult or of birth control methods may have put her at increased risk of being accused of witchcraft.[10] In the genocide perpetrated by Pol Pot and the Khmer Rouge regime, the entire intellectual class was slated for extermination.[11] Some cultures place a high value on sexual innocence, particularly in girls, and a young woman might find her marriage prospects dimmed if she appears to know too much about sex or if her general education intimidates prospective

---

[8] If side A knew how A would behave in a crisis; and if side B, while not knowing how A would behave but knowing that A knew how A would behave; then A could become less able to issue an effective probabilistic threat. B could reason that if A knew that A would launch a nuclear attack in a crisis then A would be less likely to threaten to create a crisis (assuming that B knew that A was uncertain as to whether B would yield to A's threat). Thus, B could infer that if A does in fact issue a threat to create a crisis, it would likely be because A knew that a crisis would not escalate into a nuclear war. This would make B less likely to yield to the threat.

[9] Montefiore 2005.

[10] Levack 1987.

[11] Fawthrop and Jarvis 2005. In any actual large-scale historical episode, of course, multiple causes are likely to have been involved, possession of dangerous knowledge being at most one contributing factor.

husbands.[12] In many schools, "nerdy" children who appear to have too much academic knowledge are ostracized and bullied. Knowing-too-much hazards, of varying degrees of severity, seem to arise in many different contexts.

4. Risks to social organization and markets

We have so far focused on the possibility of an adversary gaining an advantage as a result of information obtained by either the adversary or ourselves. The adversary might then harm us deliberately and directly, as in a military attack; or indirectly and perhaps unwittingly by weakening our competitive position.

Yet there are many other types of information hazard. In particular, information can sometimes damage parts of our social environment, such as cultures, norms, and markets. Such damage can harm some agents without necessarily strengthening or benefitting their adversaries or anybody else.

> *Norm hazard*: Some social norms depend on a coordination of beliefs or expectations among many subjects; and a risk is posed by information that could disrupt these expectations for the worse.

Behavior in some given domain can be guided by different norms, with the result of different social equilibria being instantiated. Norms are sometimes formulated and imposed from above, with legal backing; for example, a norm that one must drive on the right side of the road. But even if there had been no such law, motorists might have spontaneously developed the norm of driving on the right side, just as there is a norm of extending the right hand in a handshake.

With regard to which side to drive on, there is no intrinsic benefit to left or right, so long as there is some clear rule that everybody follows. In other domains, however, different possible social equilibria can have widely divergent consequences for human welfare. In a society with low corruption, individuals might prosper most by being honest, trusting, and playing by the official rules; while in a high-corruption society, individuals following those strategies would be suckers. The optimal strategy for one individual depends on the strategies pursued by others who chose their strategies on the basis their expectations about how others will react. Information that alters these expectations can therefore change behavior. This can lead to a change in norms that moves a group or a whole society from one equilibrium state to another. The possibility of moving to a worse social equilibrium can be characterized as a norm hazard.

Locally suboptimal policies are sometimes justified from a wider perspective as a price worth paying to protect norms that serve to block a slide into a worse social equilibrium. A socially conservative outlook might be based on the belief that such slides are a major danger and that strict

[12] Schlegel 1991.

enforcement of existing norms is a necessary countermeasure.[13] Even calling into question a particular norm, or making small adjustments of some norm, might undermine the authority of—and thereby weaken—the overall structure of extant norms, increasing the risk of moral decay or social unraveling.[14] Similarly, one can object to some judicial decisions because of the legal precedents they set; to some foreign policy decisions because of their effect on credibility; and so forth.[15]

If we take the word "norm" in its wide sense, we can also think of money as a norm or a bundle of norms. The functions that money serves in the economy depend on people having certain expectations about other people's beliefs and attitudes towards money and its specific forms, such as cash. Counterfeiting and excessive money-printing can undermine a currency, destroying its ability to serve as a medium of exchange and a store of value. This is another example of norm hazard.

It is obvious how some kinds of false information can damage beneficial norms. But norms can also be damaged by true information. We have already alluded to the phenomenon of *self-fulfilling prophesies*—people acting more honestly if they believe themselves to be in a low-corruption society, and vice versa; drivers driving on the right side if they believe that others will make the same choice. Another phenomenon in which true information can damage norms is *information cascades*. Information cascades can arise when agents make choices sequentially, and each agent has, in addition to some noisy private information, the ability to observe the choices (but not the information) of some of the agents in front of her in the queue.[16] It has been suggested that

---

[13] Hirschman 1991.

[14] Cf. Schelling's concept of a focal point (Schelling 1960).

[15] Rizzo and Whitman 2003; Volokh 2003.

[16] Suppose that hundreds of rock fans are driving to the Glastonbury music festival. At some point each driver reaches an intersection where the road signs have been vandalized. As a result, there is uncertainty as to whether to turn left or right. Each driver has some private information, perhaps a dim drug-clouded recollection from the previous year, which gives her a 2/3 chance of picking the correct direction. The first car arrives at the intersection, and turns right. The second car arrives, and also turns right. The driver in the third car has seen the first two cars turn right, and although his private intuition tells him to turn left, he figures it is more likely that his own intuition is wrong (1/3) than that both the preceding cars went the wrong way (1/9); so he turns right as well. A similar calculation is performed by each subsequent driver who can see at least two cars ahead. Every car ends up turning right.

In this scenario, there is a 1/9 chance that all the rock fans get lost. Let us suppose that if that happens, the festival is cancelled. Had there been a dense fog, preventing each driver from seeing the car in front (thus reducing information), then, almost certainly, approximately 2/3 of all the fans would have reached Glastonbury, enabling the festival to take place. Once the festival starts, any lost fan can hear the music from afar and find their way there. — We could thus have a situation in which reducing information available to each driver increases the chance that he will reach his destination. Clear weather creates an informational cascade that leads to an inefficient search pattern.

information cascades play an important explanatory role in accounting for faddish behavior in many domains, including finance, zoology, politics, medical practice, peer influence and stigma.[17] Informational cascading might also contribute to a Matthew ("the rich get richer") effect. For example, eminent scientists tend to get more credit than unknown researchers for similar contributions.[18] Part of the reason might be that when there is uncertainty as to who made the bigger contribution, it is, *ceteris paribus*, more likely to have been made by the more eminent scientist, who consequently gets the credit; but with the result that the fame of the already slightly famous can snowball to undeserved proportions while others are unfairly ignored.

Another important way in which true information can damage social organization is through *information asymmetries*. When one party has information that others lack, the information asymmetry sometimes prevents mutually beneficial transactions from taking place.

> *Information asymmetry hazard*: When one party to a transaction has the potential to gain information that the others lack, a market failure can result.

Economic models of adverse selection and moral hazard illustrate some of the possibilities. In the market for used automobiles, the seller often has more information about the quality of the car than the prospective buyer. Owners of bad cars, "lemons", are more willing to part with their vehicle than owners of good cars. Buyers, knowing this, suspect that the car being offered them is a lemon, and are willing to pay accordingly. This buy price is too low to interest potential sellers of good cars, with the result that high-quality used cars are withheld from the market, leaving predominantly lemons. The information asymmetry inhibits the market in high-quality used cars. This helps explain why the value of a brand new vehicle plummets the moment it disembarks the dealership.[19]

Insurance offers many illustrations of the potential for negative effects of information asymmetry. For example, in countries with private health care, consider a scenario in which the availability of genetic testing combined with vastly improved knowledge about how interpret the tests provide buyers of health insurance with a wealth of new information about their personal risk profile. If privacy legislation prohibited insurance companies from accessing the same information, the resulting information asymmetry could undermine the insurance market. Adverse selection would lead the subjects with the highest risk profiles to buy more insurance. Insurance companies, anticipating this, would raise premiums. The higher premiums would deter more low-risk subjects, amplifying the adverse selection effect—until, in an extreme scenario, the health insurance market collapses.[20] Relative to such a scenario, both buyers and sellers of insurance might better off if

---

[17] Bikhchandani, Hirshleifer and Welch 1992.

[18] Merton 1968.

[19] Akerlof 1970. Here, as throughout this paper, we are not concerned to give a detailed account of some particular empirical phenomenon; our goal is to illuminate some features of the conceptual landscape.

[20] It is unrealistic to suppose genetic information to produce such an extreme consequence since much of the variance in health outcomes is due to non-genetic variables and chance.

neither obtains the extra information. The possibility of release of new information to one party of a potential transaction can thus, under certain circumstances, be a hazard.

Although asymmetric information is particularly corrosive, insurance markets can also collapse because of *symmetric information*, information that is shared between all parties. Insurance is predicated on uncertainty. It makes no sense for you to insure against a loss that you are certain you will *not* incur, and it makes no sense for an insurance company to offer you insurance against a loss that it knows that you *will* incur at some known date; the premium would have to exceed the coverage.

If the *only* useful role of insurance were to reduce uncertainty about future revenue or welfare, then information that increased predictability would remove the need for insurance at the same time as it removed the possibility of insurance: no harm would be done. However, insurance serves other functions as well. One is redistributive justice. In insurance, the fortunate subsidize the unfortunate. Selfish agents join the scheme because they do not know, *ex ante*, to which group they belong.

Entire political philosophies have been constructed around the notion of insurance. For example, in John Rawls' theory of justice, the just social order is defined with reference to what people would hypothetically choose from behind a "veil of ignorance", i.e. if they were ignorant about which social role they themselves occupy.[21] A Rawlsian might attribute many of the practical difficulties in getting this conception of justice implemented to the fact that voters and political decision-makers are in reality not behind a veil of ignorance. Selfish people who know their own circumstances—their socio-economic class, race, occupation, and so forth—may favor policies that promote their self-interest rather than the allegedly fairer policies that they would have chosen had they been ignorant about their own actual situation. Knowledge of one's present and future situation, though, is a matter of degree. One can think of scenarios in which increasing the availability of information about these things would make the implementation of a just social order more difficult. For instance, elite support for a social safety net might slacken if elites could know with certainty that neither they nor their children or grandchildren would ever need to use it.[22] Support for protection of freedom of speech and minority rights might weaken if most individuals could be sure that they would never find themselves in a prosecuted minority and that their opinions would never be among the ones that the censors would silence.

The possibility of such effects of symmetric information can be viewed as a risk:

> *Unveiling hazard*: The functioning of some markets, and the support for some social policies, depends on the existence of a shared "veil of ignorance"; and the lifting of which veil can undermine those markets and policies.

[21] Rawls 2005.

[22] A similar point is made in Kavka 1990. Kavka also argues that intense social conflict would arise if those individuals and groups that would suffer (possibly non-compensable) harm from some proposed policy could know this ex ante.

This phenomenon can also be instantiated in the iterated prisoner's dilemma, where agents face a choice between cooperating and defecting in an unknown number of repeat encounters. Agents might cooperate in each round in order to secure the other player's cooperation in the following rounds. Yet cooperation can unravel if players know how many rounds there will be. When they know they are in the final round—hence facing the equivalent of a traditional one-shot prisoner's dilemma—they both face incentives to defect. Worse, in the penultimate round they can foresee that they will next be in the final round in which they will both defect; so incentives favor defecting in the penultimate round too—and so on, all the way back to the first round. The opportunity for long-term mutually beneficial cooperation could thus be ruined through the loss of ignorance about the future duration of the relationship.

We have already discussed *intellectual property hazard* as an example of adversarial risk. Intellectual property theft, though, is a problem not only for individual firms that risk losing out to their competitors. Threats to intellectual property can undermine entire sectors of the economic system by making it harder for firms and individuals to internalize the benefits of their research and product development. The legal system provides only partial protection and imposes big administrative, transaction, and enforcement costs which can themselves impede innovation. Defense of intellectual assets therefore tends to depend also on various forms of secrecy and physical barriers to access and copying of sensitive data. The potential for developments that would reduce these obstacles, when that would have negative consequences, constitutes an unveiling hazard.[23]

Consider, finally

> *Recognition hazard*: Some social fiction depends on some shared knowledge not becoming common knowledge or not being publicly acknowledged; but public release of information could ruin the pretense.

Two gentlemen, *A* and *B*, are in a small room when *A* breaks wind. Each knows what has happened. Each might also know that the other knows. Yet they can collude to prevent an embarrassing incident. First, *B* must pretend not to have noticed. Second, *A* might, without letting on that he knows that *B* knows, provide *B* with some excuse for escaping or opening the window; for example, *A* could casually remark, after a short delay, that the room seems to be rather overheated.[24] The recognition hazard consists in the possibility of dissemination of some information that would constitute or force a public acknowledgement; only then would the flatus become a socially painful faux pas.

[23] The claim here is not that the easier it is to protect intellectual assets, the better. In some areas there might for example be an inefficiently high level of legal protection. Developments that make intellectual property theft easier to carry out, and harder to detect and punish, could then be socially beneficial.

[24] Goffman 1959.

5. Risks of irrationality and error

Then there are information hazards which, by contrast to those mentioned above, depend on either irrationality or false beliefs. This dependency, of course, does not consign the corresponding hazards to a marginal status.

Consider

> *Ideological hazard*: An idea might, by entering into an ecology populated by other ideas, interact in ways which, in the context of extant institutional and social structures, produce a harmful outcome, even in the absence of any intention to harm.

Suppose that Bob believes that scripture *S* consists of exclusively literal truths, and that he is committed to doing whatever *S* says ought to be done. Suppose, furthermore, that *S* contains the (presumably false) moral statement "Thou shalt drink sea water", but that Bob is unaware of this. The potential dissemination to Bob of this part of the content of *S* constitutes an information hazard. The information could harm Bob by inducing him to drink sea water. (Note that the conveyance of *true* information harms Bob here; in this case, the true information that *S* calls for drinking sea water.)

In the preceding example, the hazard posed by the relevant information is tightly coupled to Bob's idiosyncratic belief system. It is true that the idea of a nuclear bomb is also a hazard only when coupled with a larger belief system—for instance, beliefs about physics and technology required to bring a bomb into existence. Yet it seems possible and useful to distinguish this kind of instrumental information hazard from ideological information hazard. Ideological hazard, we might say by way of explication, refers to the possibility that that somebody will be mislead to head in some bad direction because of the way that some information interacts with false beliefs or incomplete knowledge.

That bad ideologies can be extremely dangerous is amply evidenced by twentieth century history. What is less clear is how ideological hazard can best be reduced. Part of the reason why this is a difficult problem is that ideology can also be a force for good. The ideology of the American civil rights movement, for example, helped push back racial discrimination in the U.S. In a wide sense, ideology is perhaps an inevitable part of the human condition, and the problem of distinguishing good from bad ideology may be no easier to solve than the problem of distinguishing good from bad policy: no simple, generally acceptable algorithm exists. Moreover, while radical ideologies may be especially dangerous, they may also—depending on what the status quo is relative to which the alternatives they present are "radical"—be especially appropriate for the situation. If the status quo is slavery and religious prosecution, then it would be a radical ideology that proposes not merely amelioration of the working conditions for slaves and reduction of the penalties for heresy, but complete abolition and unlimited religious freedom.

Next we turn to the fact that human beings are not perfectly rational nor do we have perfect self-control. We can be distracted against our will and we can succumb to temptation against our better judgment. Exposure to information can have effects on us other than simply improving the accuracy of our representations of the world.

Some information is distracting. It involuntarily draws our attention to some idea or theme when we would prefer to focus our minds elsewhere. An advertizing jingle might loop in our minds and distract us from something we would rather be thinking about. One technique we use to fight temptation is to put something out of our mind; yet information about the tempting object can undermine our effort and make us more likely to cave. A recovering alcoholic can be harmed by exposure to a vivid account of the attributes of Chateau Petrus Pomerol 1990.

> *Distraction and temptation hazards*: Information can harm us by distracting us or presenting us with temptation.

In most individual cases the damage done by distracting or tempting information is perhaps minor. Yet it is not unreasonable to wonder whether the ready availability of certain *kinds* of information might potentially cause damage on a wider scale. Perhaps it could be argued that television has an aggregate effect on the contemporary human condition not too dissimilar from that which would be produced by the widespread recreational use of opiate drugs. In the future, even more compellingly presented information and hyper-stimuli might become available and prove enormously addictive; for example, new forms of highly immersive or interactive virtual reality environments. Drug-like effects on our psyches can be produced not only through injection, ingestion, and inhalation but also through the intake of information presented in certain manners to our senses.

We can also be harmed by exposure to (the template hazard of) bad role models. Even when we know that a model is bad, and we would prefer not to be influenced by it, prolonged exposure can nevertheless be detrimental because of a kind of social osmosis. Someone who aspires to a good writing style might be well advised to avoid reading too much trash. One who seeks to cultivate a lofty sentiment might want to avoid the company of the mean and petty. And those who hope that their children will become upright citizens should keep them away from delinquent peers.[25] Recent studies indicate that subjective well-being and even body mass are significantly influenced by our associates.[26] Thus,

> *Role model hazard*: We can be corrupted and deformed by exposure to bad role models.

One example of this is the "Werther" effect, named after the wave of suicides among young men which swept Europe after the publication in 1774 of Goethe's novel *Die Leiden des jungen Werthers*. Several studies have corroborated the existence of such an effect, finding a link between media reporting of high-profile cases and ensuing copycat suicides.[27]

---

[25] Other things being equal, that is; which of course they seldom are. When deciding what to do, one should also take into account that exposure to a wide range of role models could provide more opportunities for choice; and that one can become wiser by also knowing something about the dark side. When excessive, the fear of contamination by bad influences is stultifying. In its extreme forms, a love of "purity" can produce dangerous intolerance and bigotry.

[26] Halliday and Kwak 2007.

[27] See e.g. Phillips 1982; Stack 1996; Jonas 1992.

Information risks arise out of our susceptibility to various cognitive biases that can be aggravated by the provision of certain kinds of data. Anchoring bias results from application of the "anchoring and adjustment" heuristic in which people estimate some unknown quantity by first anchoring on some figure that happens to come to mind and then adjusting this preliminary estimate either up or down in an attempt to reflect their total information. This leads to bias when people initially anchor on an irrelevant quantity and then under-adjust in the adjustment phase. In one study subjects were asked to estimate the number of countries in Africa. Before producing their estimate, a wheel of fortune was spun. Subjects who observed a larger number on the wheel tended to give a higher estimate of the number of African countries, despite the transparent irrelevance of the former fact. The extra piece of true information about the number on the fortune wheel diminished the accuracy of geographical judgment.[28]

Many people overestimate their own virtues and abilities. Suppose such a person receives some additional weak cue of their supposed excellence, such as a good score on a trivia quiz. This bit of evidence, which we can suppose to be true and in a very limited way informative, could aggravate their self-overestimation and conceitedness.[29]

Even knowledge of human biases and critical philosophy can lead the unwary deeper into error, and reduce his ability to learn, by arming him with clever arguments with which to rebut objections and rationalize inconvenient facts.[30] A special kind of fool is born when intelligence thus outwits itself.

> *Biasing hazard*: When we are biased, we can be led further away from the truth by exposure to information that triggers or amplifies our biases.

Methodology, such as double-blinding in drug trials, can help reduce the risk of biases entering uninvited into our thinking and acting. For similar precautionary reasons, the gullible often have reason to avoid the highly persuasive. And if one plans to experience transports and ecstasies that will temporarily increase one's susceptibility to dangerous illusions and impulses, one should first have oneself tied to the mast.

Conversely, information could also harm us by *reducing* our biases insofar as our biases serve some useful purpose. For example, a tendency to overestimate our own abilities might not only make us feel happier and confident; a strong belief in our own ability might also signal competence and lead others to ally with us, promote us, or vote for us. Information that helps us see ourselves for what we really are could deprive us of these benefits. It is also possible that society benefits from excess individual risk-taking in some disciplines; for example if entrepreneurs, inventors, and young academics overestimate their own chances of success. If these occupations have net positive

---

[28] Tversky and Kahneman 1974.

[29] Ditto for those who underestimate their own virtues and abilities: feedback that confirms this tends to be internalized while feedback that contradicts it tends to be ignored (Brown, Dutton et al. 2007).

[30] Yudkowsky 2008.

externalities, it could be beneficial that biases and unrealistic expectations of fame, fortune, or high achievement seduce additional entrants into these fields.

> *De-biasing hazard*: When our biases have individual or social benefits, harm could result from information that erodes these biases.

There is also a wider phenomenon of which role model influence is but a special case. Our brains are constantly reshaped by what we learn and experience. Information gleaned is not simply stored away as inert data packages, as though it were new volumes superadded to some internal bookshelf. Rather, the incoming information interacts complexly with preexisting cognitive structures in ways that are not always easy to characterize in folk psychological terms. New concepts might form; boundaries of extant concepts might change; neuronal wiring patterns are altered; some cortical areas might expand, causing other areas to contract; and so forth. There is a risk that some of these changes will be for the worse.

> *Neuropsychological hazard*: Information might have negative effects on our psyches because of the particular ways in which our brains are structured, effects that would not arise in more "idealized" cognitive architectures.

Too much knowledge can be bad for some types of memory.[31] Perhaps some mental illnesses result from inappropriate cross-talk between cognitive modules designed to operate as more encapsulated units—a kind of undesirable internal information dissemination. A recurring idea in literature and mythology is the "motif of harmful sensation", where a person suffers mental or physical harm merely by experiencing what should normally be a benign sensation (the myth of Medusa, beliefs about the "evil eye" etc.). A real world example of harmful sensation is photosensitive epilepsy which can be triggered in some sensitive individuals flickering lights or specific geometric patterns.[32]

Irrelevant information can make valuable information harder to find. This fact is used in steganography, the cryptographic technique of hiding secret messages within representations that appear to be of something else so that even the existence of covert text is concealed. For example, some of the pixels in an image file can be subtly modified so as to encode a verbal message in what looks like an ordinary tourist picture. In a similar vein, legal defense teams sometimes conceal incriminating documentation that has been subpoenaed by the prosecution by overwhelming it with such massive amounts of archival material that the relevant documents cannot be located in time for the trial.

---

[31] Robinson and Sloutsky 2007.

[32] When the cartoon episode Dennō Senshi Porygon aired in Japan in 1997, one scene featuring an explosion rendered with strobe lighting effect caused mild symptoms in 5-10% of the viewers (though some of these might instead have been afflicted with epidemic hysteria) and sent 685 children to hospital in ambulance. No long term damage was reported. There has also been at least one malicious attempt to deliberately trigger photosensitive epilepsy online, but it appears not to have been very successful. See Radford and Bartholemew 2001; Takada, Aso et al. 1999.

Information-burying hazard: Irrelevant information can make relevant information harder to find, thereby increasing search costs for agents with limited computational resources.[33]

On a grander scale, an overabundance of informational affordances might deflect our thinking from topics that are more central to us and relatively more worthy our contemplation, so that we shall live, as in T. S. Eliot's characterization of the modern predicament, "Distracted from distraction by distraction".[34] This kind of possibility leads us to the next section.

6. Risks to valuable states and activities

We have looked at how information can cause harm by affecting behavior. Information can also harm through its direct psychological effects, for example by causing disappointment. Moreover, according to at least some accounts of well-being, information can cause harm even aside from psychological spillover effects by affecting some part of some epistemic or attentional state that plays a constitutive role in some subject's well-being. Thus we can define

Psychological reaction hazard: Information can reduce well-being by causing sadness, disappointment, or some other psychological effect in the receiver.

And we can distinguish this from the following more "philosophically intricate" notion:

Belief-constituted value hazard: If some component of well-being depends constitutively on epistemic or attentional states, then information that alters those states might thereby directly impact well-being.

Consider first the obvious example of a psychological reaction hazard: bad news, the receipt of which makes us sad.

Disappointment hazard: Our emotional well-being can be adversely affected by the receipt of bad news.

In some cases, if something goes wrong, we are bound to hear of it eventually. In such cases, the disappointment is in a sense already "committed" when the adverse event takes place, even though it might take a while before the effect is known and felt.

In other cases, however, there is a real chance that if a subject avoids hearing of her misfortune now, she will remain ignorant and will be spared the disappointment that the bad news would occasion. Such cases make it easier to disentangle the disappointment hazard from other possible harms that might be involved. Take the case of a mother on her deathbed, whose only son is fighting in some faraway war. The mother faces at least two distinct risks. First, there is the risk

[33] And potentially result in worse solutions; for a discussion of how excessive knowledge can degrade performance in some artificial intelligence systems, see Markovitch and Scott 1988.

[34] Eliot 2001.

that her son will be killed or injured; this is not necessarily an information risk. Suppose that the son is in fact killed. Then there is a second risk, which is that the mother will find out about her loss. Suppose that the news is contained in a letter, which might reach her quickly or it might be delayed. If it reaches her quickly, she will spend her last days in agony and despair; if it is delayed, she will die in peace. Here we might say that the mother is exposed to a severe disappointment hazard.

Spoilers constitute a special kind of disappointment. Many forms of entertainment depend on the marshalling of ignorance. Hide-and-seek would be less fun if there were no way to hide and no need to seek. For some, knowing the day and the hour of their death long in advance might cast shadow over their existence.

Before his retirement, my father would sometimes miss a pivotal televised soccer game that took place during working hours. Planning to watch the reprise later, he would meticulously avoid any news source that might disclose the results. His design, however, was thwarted by my grandfather, who had watched the game live and who invariably found himself unable to refrain from making not-quite-subtle-enough allusions to the match, enabling my father to guess who had won.

> *Spoiler hazard*: Fun that depends on ignorance and suspense is at risk of being destroyed by premature disclosure of truth.

Knowledge can also exert more general effects on our psyches and personalities. Perhaps an unwanted cynicism is promoted by an excess of knowledge about the dark side of human affairs and motivations. Nietzsche warned of the misuses of history: how historical knowledge, approached and valued in a certain way, can sap our zest for life and inhibit artistic and cultural authenticity and innovation. The danger Nietzsche pointed to was not the effects of any one particular piece of information but rather the consequences of a certain "excess of history" which can cause living to crumble away: "es gibt einen Grad, Historie zu treiben, und eine Schätzung derselben, bei der das Leben verkümmert und entartet" ("there is a degree of doing history and valuing of it through which life atrophies and degenerates").[35] If Nietzsche is right about this, and if the dissemination of (various kinds of) information about the past can, under unfavorable circumstances, contribute to such an atrophy of spirit, then we have here an example of another type of psychological reaction hazard, namely

> *Mindset hazard*: Our basic attitude or mindset might change in undesirable ways as a consequence of exposure to information of certain kinds.

Along similar lines, some people worry that scientific reductionism, akin to strip-mining in an ancient forest, despoils life of its mystery and wonder.

Let us turn to belief-constituted value hazard. In practice, this category can be difficult to distinguish from psychological reaction hazard.

---

[35] German quotation taken from Nietzsche 1984; English translation taken from Nietzsche 2007.

Consider again the example of the mother on her deathbed who risks hearing that her son has been killed. There is clearly one respect in which hearing this would be bad for her: it would cause her extreme psychological distress. This is sufficient for there to be a psychological reaction hazard. It does not require that it would be bad for the mother, all things considered, to hear of her son's death.

There are several reasons for this. First, there are of course various practical matters that would need to be considered in an actual situation like this: Perhaps the mother needs to know so that she can alter her will? Perhaps concealment of unpleasant news would tend to erode social trust? But even aside from such pragmatic considerations, there is a second type of reason why it might be better for the mother to know of her son's death despite the suffering this knowledge would cause her. Such knowledge, according to some moral theories, can be a component of a person's well-being ("the good for a person") even if it affects the subjective component of well-being for the worse. One might hold that a life is made worse, other things equal, by its being lived in ignorance of important facts about the central concerns of that life. Life in a fool's paradise, even if it scores high on the hedonic dimension, might nevertheless score quite low in overall well-being on such a theory.

Just as one might hold that there is some knowledge the possession of which is an important constituent of a good life, one might also hold that there is knowledge (at least for some people, in some circumstances) that makes a direct negative contribution to their well-being. This can most obviously be seen to be the case according to a preference-satisfaction account of well-being; for there we can generate examples trivially simply by supposing somebody to have a basic preference against knowing about some particular subject matter. But many other accounts of well-being might also permit of examples of such directly burdensome knowledge. Innocence might be valued for its own sake and might be ruined by knowledge. We might treasure our privacy and find it infringed by other people's knowing things about us that we would rather have kept to ourselves or shared exclusively with chosen intimates. Or we might be better off not knowing some personal details about others, not just because such knowledge might expose us to risk of worse treatment from others (as in knowing-too-much hazard) but also because our awareness of these details would stand in the way of our conceiving of others in manners that are more appropriate or more to our liking. With regard to our friends' bowels and our parents' bedrooms, the less we know the better.[36]

One commonly feared risk from information is

> *Embarrassment hazard*: We may suffer psychological distress or reputational damage as a result of embarrassing facts about ourselves being disclosed.

Embarrassment hazards (which often take the form of signaling hazard) commonly combine elements of psychological reaction hazard, belief-constituted value hazard, and competitiveness hazard. We may even fear to embarrass ourselves to ourselves, perhaps because self-esteem is not

[36] And of course Bismarck claimed of laws and sausages that it is better not to see them being made.

a wholly private matter but is also a social signal that influences other's opinions of us.[37] Some psychologists believe that a concern to protect self-esteem from undermining by self-relevant failures can lead individuals to engage in self-handicapping behavior.[38] This could help account for some instances of phenomena such as procrastination, hypochondria, substance abuse, and practice-avoidance.[39] Suppose that thinking of yourself as intelligent is important for your self-esteem and that you have an important exam coming up. If you practice hard and fail on the exam, your sense of self-competence will take a hit. But if you put off practicing until the night before the exam, your risk is reduced; even smart people can do poorly on exams when they have not studied enough. And if despite the handicap of insufficient preparation you still manage to get a high mark, why then you must be truly brilliant. Such perception management can impose significant costs.

Risk of embarrassment can suppress frank discussion. A study on deliberation in the Federal Reserve's Federal Open Market Committee found evidence that a newly adopted policy of transparency involving the publication of detailed transcripts from monetary policy meetings stifled the voicing of dissenting opinions and seemed to reduce the quality of debate.[40]

Intangible assets, such as reputation and brand name, constitute a large part of the value of many firms. Embarrassments that negatively impact these assets can cause billions of dollars in damage. For an example on an even grander scale, consider the Cold War superpower rivalry, in which both contenders were engaged in status contest as well as a military arms race. The Apollo project was a direct response to the embarrassment caused to the United States by the Soviet Union's launch of Sputnik 1, an accomplishment that challenged the America's claim to technological superiority. The Vietnam and the Afghan wars were both prolonged because of reluctance to suffer the reputational damage that leaders believed would result from admitting defeat.

7. Risks from information technology systems

It is not only animate beings that process and disseminate information; our information technological systems do so as well. Distinctive information hazards arise in relation to our computers and networks.

Information technology systems are vulnerable to unintentionally disruptive input sequences or system interactions as well as to attacks by determined hackers. Here we consider only risk occasioned by informational effects—unanticipated system interactions, worms, viruses, Trojan horses, denial of service attacks, and so forth. This means we exclude risks arising from the possibility of flooding, power outages, and somebody attacking your computer with a sledge

[37] Hobden 1997.

[38] Berglas and Jones 1978.

[39] Smith, Snyder and Perkins 1983; Stone 2002; Thompson and Richardson 2001.

[40] Meade and Stasavage 2008.

hammer—except in so far as a risk consists in the possibility of informational amplification of the effects of some such non-informational trauma. Thus, the risk that you might drop your brand new laptop on a hard floor so that it breaks and you incur the cost of buying a replacement is not an information hazard. Nor is the risk that some critical information system might go down necessarily an information hazard as defined here. The mere cessation of functioning of some useful information-providing system is not enough unless the cause of the cessation is informational or the harm arises from some kind of undesirable propagation of information.

> *Information system hazard*: The behavior of some (non-human) information system can be adversely affected by some informational inputs or system interactions.

This category can be subdivided in various ways: one could, for example, distinguish computer hazards from network hazards; or single out threats to critical information infrastructure; or one could make a separation between scenarios involving loss of data, corruption of data, dissemination of data to the wrong parties; and so forth. Quite a lot of attention is already given to information system hazards, and much of this attention is focused on what we may term

> *Information infrastructure failure hazard*: There is a risk that some information system will malfunction, either accidentally or as result of cyber attack; and as a consequence, the owners or users of the system may be inconvenienced, or third parties whose welfare depends on the system may be harmed, or the malfunction might propagate through some dependent network, causing a wider disturbance.

Risks of this type can be quite severe when some complex system or network is used to coordinate important human activities. For instance, a corruption of the software that undergirds important financial systems could have serious consequences.

A different type of information system hazard is that some information system will in fact function as intended, but by doing so it will cause harm or amplify some risk of harm.

A privacy advocate might object to some government database project that will amass vast quantities of information about the citizenry, not only because of the risk that the system might malfunction or be hacked, but also because of the risk that it will perform to specification and thereby strengthen the state's ability to monitor the activities of its people and—should the government one day see a need to do so—to take action against elements deemed undesirable or disloyal. Even if it were admitted that the government that builds the system can be trusted to use it only for good, one might fear that later governments which inherit the system cannot be so trusted, or that some more pernicious government elsewhere will see in the system an inspiring precedent (cf., idea hazard, and attention hazard) or justification (cf., norm hazard) for building its own comparable system and applying it to its own nefarious ends.

Similar concerns can apply to private firms, such as Google, that collect personal information about hundreds of millions of users. Consider how useful it would have been for a twenty-first century Stalin to be able to have his security service data mine the citizenry's email correspondence and

search engine queries—not least text written before his ascent to power and at a time when his enemies might have communicated their most incriminating thoughts unguardedly.[41]

> *Information infrastructure misuse hazard*: There is a risk that some information system, while functioning according to specifications, will service some harmful purpose and will facilitate the achievement of said purpose by providing useful information infrastructure.

A system can also be dangerous by presenting an easy opportunity for *unintentional* misuse. Consider a poorly designed email program that makes is too easy for the unwary user accidentally to forward an email reply to all the addressees in her contact list; an embarrassment waiting to blush. This hazard lies on the border between information infrastructure failure and information infrastructure misuse, it being unclear whether such an email program is functioning according to its intended specifications and arguable how apportion blame between the system's designers and its users.

For comparison, we may also note two other types of hazard potentially arising out of information technology (but which are typically not information system hazards) where the harm is not so much a consequence of the general information infrastructure services that a system provides or fails to provide but instead is more directly related to the "agency" of the system itself:

> *Robot hazard*: There are risks that derive substantially from the physical capabilities of a robotic system.

An autonomous vehicle, loaded with explosive missiles, and able to launch on its own initiative, could constitute a robot hazard. We can contrast this with

> *Artificial intelligence hazard*: There could be computer-related risks in which the threat would derive primarily from the cognitive sophistication of the program rather than the specific properties of any actuators to which the system initially has access.

An artificial intelligence would need to be very advanced in order to pose any significant threat in virtue of its own ingenuity and agency. The creation of artificial general intelligence, with general powers of reasoning exceeding those of human beings, would be associated with a serious, indeed existential, risk.[42] A superintelligence, even if initially restricted to interacting with human gatekeepers via a text interface, might hack or talk its way out of its confinement. It could then gain control over effectors to conduct operations in the external world—for example, by using its powers of persuasion to get human beings to do its biddings, or by assuming control of robotic manipulators. It could use these effectors to develop new technologies and to secure a more comprehensive grasp of its physical surroundings. The threat posed by a sufficiently advanced artificial intelligence may depend much more on its cognitive capabilities and its goal architecture than on the physical capabilities with which it is initially endowed.

---

[41] Of course there are big potential upsides too; e.g., a good government could subpoena this information for use in a good cause.

[42] Bostrom 2002; Yudkowsky 2008.

Not all risks related to robots or artificial intelligences are to be classified as information system hazards. A risk would count as such a hazard if, for example, it arose from the possibility of a computer virus infecting the operating system for a robot or an AI. But aside from such special cases, we shall not count robot hazards and artificial intelligence hazards as information system hazards.[43]

There is, however, another way for robot- and AI-related risks to enter our information hazard taxonomy. They can enter it in the same ways as any risk relating to potentially dangerous technological development.

8. Risks from development

Both technological innovation and economic development more broadly arise from the accumulation of information, ideas, and insights; and this can result in a range of risks that we can group together under the rubric of development hazards.

> *Development hazard*: Progress in some field of knowledge can lead to enhanced technological, organizational, or economic capabilities, which can produce negative consequences (independently of any particular extant competitive context).

When the mushroom clouds rose over Hiroshima and Nagasaki, physicists, many of whom had entered their profession for the sheer joy of discovery, found themselves complicit in the deaths of perhaps 200,000 people.[44] If the cold war had ended in an all-out nuclear exchange between NATO and the Soviet Union, as it might easily have done, then more than a billion civilians could have died as a fairly direct consequence of the development of nuclear weapons.[45]

---

[43] There is, of course, a sense in which both robots and advanced machine intelligences are information systems. There is also a sense in which the human brain is an information system. Yet the risks that arise from intelligence in general, or from the physical equipment of some robot, are extremely heterogeneous; wherefore it would seem not very illuminating to construct an "information system hazard" category that lumped them all together.

[44] The Atomic Archive estimates the deaths in Hiroshima and Nagasaki immediately following the bombings at 105,000, with a further 94,000 injured (The Manhattan Engineer District 1946). Many have later died of cancer or birth defects caused by radiation exposure, but the exact figures are a subject of debate.

[45] President Kennedy is said to have at one point during the Cuban missile crisis estimated the probability of a nuclear war between the U.S. and the USSR to be "somewhere between one out of three and even" Kennedy 1968; Leslie 1996. John von Neumann, who as chairman of the Air Force Strategic Missiles Evaluation Committee was one of the architects of early U.S. nuclear strategy, is reported to have said it was "absolutely certain (1) that there would be a nuclear war; and (2) that everyone would die in it" (Putnam 1979, 114). See also Cirincione 2008.

Robert Oppenheimer, the scientist who had spearheaded the Manhattan project, acknowledged afterwards that "the physicists have known sin; and this is a knowledge which they cannot lose."[46] Of course, reaching a moral verdict on the scientists who worked on the Manhattan project is not as simple as toting up the number of deaths that were later caused by the weapon they invented. Many of these scientists devoted themselves to the project because they feared that Hitler might get the bomb first—a fear which, although it turned out to be unfounded, was reasonable given the information available when the project began. Richard Feynman, another physicist who later reflected on his involvement, regarded his initial decision to participate as morally justified for just this reason; but he thought that he had failed morally in not reconsidering his involvement after it became clear that Hitler had been unable to acquire the bomb and that Germany could be defeated without it. Furthermore, the decision to use the bomb was not taken by physicists but by President Truman, who may have acted on a variety of motives in a complex strategic situation; and so forth.[47]

The point here is not that some particular past action was or was not justified, but that this kind of consequence can result from the information-gathering work of physicists—including basic research such as the earlier work in quantum and particle physics that laid the theoretical foundations for the Manhattan project. To proceed blithely and without scruple, as though nothing very bad could come from such research, was perhaps excusable naïveté back then.[48] For our own generation, which is able to observe more historical precedent, such negligence would more likely amount to culpable abrogation of moral responsibility.

What was true of physics in the decades leading up to the Second World War may be true of other academic disciplines today. Biology and biotechnology, while providing urgently needed munitions for use in our battle against disease, malnourishment, and age-related debility, also threaten to arm the human species with weapons of mass destruction that might be deployed against our own kind.

Recent developments point to disturbing possibilities down the road. Consider the steadily improving capacity and availability of DNA synthesis machines. This trend is worrisome when one considers that the genomes of many extremely dangerous pathogens reside in the public domain, including Ebola, Marburg, smallpox, and the Spanish flu virus (believed to have killed more than 50 million people during 1918-1919). The knowledge and technology required to genetically modify microorganisms so as to enhance their pathogenicity and their resistance to countermeasures are also advancing. Technological barriers to the production of superbugs are being steadily lowered while biotechnological know-how and equipment diffuse ever more widely.[49]

---

[46] Oppenheimer 1947.

[47] For one attempt at a moral assessment, see Glover 2001.

[48] Although Leo Szilard's example suggests that much of this naïveté was avoidable at least as early as 1933. Rhodes 1995.

[49] See e.g. Nouri and Chyba 2008. Of course, there are also risk-mitigating benefits from such research, for example better prophylactics and therapeutics, and better knowledge of our own vulnerabilities.

Dangerous information could also arise from other fields of inquiry. Advanced future forms of molecular nanotechnology might be used to build weapons system even more powerful than hydrogen bombs and supergerms.[50] Artificial intelligence might one day surpass biological intelligence and thereby become extremely powerful.[51]

Technologies for monitoring and modifying human behavior might advance on several fronts such as ubiquitous surveillance systems, automated face and voice recognition software, effective lie detection, psychopharmacology, genetic engineering, or neural implants. Social science might make progress on understanding and predicting the triggers of social unrest and insurrection. Such capabilities could be used for good or ill. In a worst case scenario they could facilitate the emergence of new and permanent forms of totalitarianisms, possibly on a global scale.

The possibilities referred to above constitute some of the most significant *existential risks* that may confront humanity in the future.[52] Other potential technological developments—some foreseeable, others perhaps not—may also create existential risks. Because of the extreme values at stake in existential risks, they can deserve substantial concern even if they could be shown to be both very unlikely and very remote—neither of which is clearly the case for the risks just mentioned.[53]

These technoscientific areas do not function in isolation. Bioweapons engineers would draw on data and techniques developed by a wide range of researchers in fields such as virology, medicine, genetics, and biochemistry. Nanotechnologists draw on draw on fields such as materials science, chemistry, protein engineering, biotechnology, and systems engineering. Artificial intelligence pioneers may benefit from advances in neuroscience, cognitive science, computer science, foundations of probability theory, and semi-conductor physics. Furthermore, all of these areas are influenced to some extent by general economic growth, which tends to lead to increased funding for research, better supporting infrastructure, and a more educated workforce.

Development hazards thus arise in many areas, and they range in severity from trivial to existential. It is important to recognize that development hazards are not confined to especially sinister or "Promethean" technological breakthroughs. Even something as innocent as medical or agricultural advances that help reduce infant mortality can pose significant development hazards, such as the risk of overpopulation and potentially negative knock-on effects on conflict, per capita income, and the environment. (Obviously, the fact that some potential development is associated with some risk does not entail that this development would on be on balance bad or that it ought not be vigorously pursued.)

---

[50] Drexler 1987; Freitas 2000; Gubrud 1997.

[51] Moravec 2000; Bostrom 1998; Vinge 1993; Kurzweil 2006; Bostrom and Sandberg 2008. A self-enhancing general intelligence that became superintelligent would become extremely powerful and would, unless rightly designed, constitute a severe threat to humanity. Bostrom 2003; Yudkowsky 2008.

[52] An existential risk is one that threats to cause the extinction of Earth-originating intelligent life or to permanently and drastically destroy its potential; see Bostrom 2002; Rees 2004.

[53] Bostrom 2003; Matheny 2007; Leslie 1996; Posner 2005.

The distinction between development hazard and the various hazards listed above as adversarial risks is somewhat vague. Development hazards, by contrast to adversarial risks, are not tied to any particular extant competitive context. For example, a risk of some technological development that consists in the potential harm to us that could result from the differential strengthening of our enemy or rival should be classified as an enemy hazard or a competitiveness hazard rather than a development hazard. But a risk of some technological development that consists in the potential for harm that arises from the fact that this technology would be likely to cause some severe accident or would generally lend itself to abuses by a wide range of individuals, groups, or states would pose a development hazard. Some technological developments might pose both adversarial and developmental risks.

9. Discussion

The considerations adduced above, although not on their own determinative of what is to be done in any particular actual case, can help inform our choices by highlighting the sometimes subtle ways in which even true information can have harmful as well as beneficial effects.

There are many ways of responding to information hazards. In many cases, the best response is no response, i.e., to proceed as though no such hazard existed. The benefits of information may so far outweigh its costs that even when information hazards are fully accounted for, we still under-invest in the gathering and dissemination of information. Moreover, ignorance carries its own dangers which are oftentimes greater than those of knowledge. Information risks might simply be tolerated. In some contexts they could be insured or hedged against using a variety of financial instruments.[54]

When mitigation is called for, it need not take the form of an active attempt to suppress information through measures such as bans, censorship, disinformation campaigns, encryption, or secrecy. One response option is simply to invest less in discovering and disseminating certain kinds of information. Somebody who is worried about the spoiler hazard of learning about the ending of a movie can simply refrain from reading reviews and plot summaries.

Sometimes, such as in the cases of some ideological hazards and some information asymmetry hazards, the danger lies in partial information. The best response may then be to provide more information, not less. Some problems can be solved through policy measures—the problem of asymmetries in health information can be obviated, for example, by instituting publicly funded universal health care. In other cases, such as distraction hazard and some biasing hazards, the solution may be to carefully select an appropriate format and context for the information that is to be presented.

When contemplating the adoption of some policy designed to restrict information, it is worth reflecting that historically such policies have often served special interests. In "The Weapon of

[54] See e.g. Petratos 2007.

Openness", a short essay on role of secrecy and openness in national security, Arthur Kantrowitz wrote:

> [S]ecrecy insiders come from a culture where access to deeper secrets conveys higher status. Those who "get ahead" in the culture of secrecy understand its uses for personal advancement. Knowledge is power, and for many insiders access to classified information is the chief source of their power. It is not surprising that secrecy insiders see the publication of technological information as endangering national security.[55]

Outsiders often have reason for skepticism when insiders insist that their inner dealings must be protected from public scrutiny. Secrecy breeds corruption. Kantrowitz argued that even with respect to the narrow criterion of military strength, a unilateral policy of openness (at least in peacetime) leads to better results.

At the same time, however, we should recognize that knowledge and information frequently have downsides. Future scientific and technological advances, in particular, may create information which, misused, would cause tremendous harm—including, potentially, existential catastrophe. If we add in the many lesser hazards that can be created by such advances, for example by technologies that facilitate commercial fraud or that introduce insidious new chemicals into the human body, the range and complexity of potential information hazards grows even greater. If we further expand our purview and consider the many indirect and reciprocal influences between, for instance, scientific information and economic growth, and if, moreover, we also give attention to the numerous ways, catalogued in preceding sections, in which information outside the realms of science and technology can cause harm—then we shall come to appreciate that information hazards are ubiquitous, potentially serious, and often non-obvious.

An analysis of the policy implications of this result is beyond the scope of this paper.[56] By way of conclusion, though, we may adumbrate two contrasting potential responses. Given the complexity of the issues involved, and their entanglement with many strategic, philosophical, and political considerations, it is not trivial to ascertain which of these responses has the most to recommend it.[57]

One possible response, then, would be to take to heart the manifold ways in which the discovery and dissemination of information can have negative effects.[58] We could accept the need to qualify the fawning admiration and unquestioning commitment to the pursuit of truth that currently constitutes official policy—if not always consistent practice—in most universities and learned

---

[55] Kantrowitz 1992.

[56] I hope to address some of these issues in a companion paper.

[57] We may be likely to overlook at least one "crucial consideration"; see Bostrom 2006.

[58] One fairly recent and well-known attempt to argue this is Bill Joy's article in which he advocates selective relinquishment of research in certain fields within artificial intelligence, nanotechnology, and biotechnology because of dangers he foresees in the future if such research is pursued (Joy 2000).

bodies. A motto like Harvard's "*Veritas!*" could be viewed as naïve and reckless. Instead, one might conclude, we ought to think more carefully and open-mindedly about which particular areas of knowledge deserve to be promoted, which should be let be, and which should perhaps even be actively impeded.

Since scholars are very likely to be biased in favor of thinking that their own field deserves to be promoted, outsiders who are less prejudiced should be brought in to participate in these deliberations. The old Enlightenment model of scientific research, which pictures science as a goose that lays golden eggs but only if allowed full autonomy and if shielded from external social control, would perhaps have to be replaced with a different model in which, for example, democratic processes and preferences are allowed greater influence over research directions and priorities.

Another response would note the great benefits that historically have come from the pursuit of knowledge and enlightenment, and fasten on the dangers inherent in any attempt to curtail free inquiry or to yoke scientific research to some preconceived notion of the social good. Those inclined to give this response need not deny that true information can in many instances be harmful or hazardous; they need only maintain that on balance we are better off as loyal subjects to the cause of enlightenment. It can also be hoped that new information technologies will bring about a vastly more transparent society, in which everybody (the watchmen included) are under constant surveillance; and that this universal transparency will prevent the worst potential misuses of the new technological powers that humanity will develop.[59]

Even if our best policy is to form an unyielding commitment to unlimited freedom of thought, virtually limitless freedom of speech, an extremely wide freedom of inquiry, we should realize not only that this policy has costs but that perhaps the strongest reason for adopting such an uncompromising stance would itself be based on an information hazard; namely, norm hazard: the risk that precious yet fragile norms of truth-seeking and truthful reporting would be jeopardized if we permitted convenient exceptions in our own adherence to them or if their violation were in general too readily excused.

It is said that a little knowledge is a dangerous thing. It is an open question whether more knowledge is safer. Even if our best bet is that more knowledge is on average good, we should recognize that there are numerous cases in which more knowledge makes things worse.[60]

References

[59] Cf. Brin 1999.

[60] For comments and discussions I am grateful to Stuart Armstrong, Allen Buchanan, HRH Prince Constanijn of the Neatherlands, Tyler Cowen, Tom Douglas, Robin Hanson, Toby Ord, Pythagoras Petratos, Rebecca Roache, Anders Sandberg, Nick Shackel, and Walter Sinnott-Armstrong. I would also like to thank Nancy Patel and Rachel Woodcock for assistance.

1. Akerlof, G. A. 1970. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics* 84 (3): 488-500.

2. Austin, J. L. 1962. In J. O. Urmson, ed. *How to Do Things with Words: The William James Lectures Delivered at Harvard University in 1955.* Oxford: Clarendon.

3. Berglas, S. and E. E. Jones. 1978. Drug choice as a self-handicapping strategy in response to noncontingent success. *Journal of Personality and Social Psychology* 36 (4): 405-417.

4. Bikhchandani, S., Hirshleifer, D. and Welch, I. 1992. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *The Journal of Political Economy* 100 (5): 992.

5. Bostrom, N. and A. Sandberg. 2008. *Whole Brain Emulation: A Roadmap.* Oxford: Future of Humanity Institute.

6. Bostrom, N. 2006. Technological Revolutions: Ethics and Policy in the Dark. In *Nanoscale: Issues and Perspectives for the Nano Century*, eds. N. M. D. S. Cameron and M. E. Mitchell, 129-152. Hoboken, N.J.: Wiley.

7. Bostrom, N. 2003. Astronomical Waste: The Opportunity Cost of Delayed Technological Development. *Utilitas* 15 (3): 308-314.

8. Bostrom, N. 2003. Ethical Issues in Advanced Artificial Intelligence. *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence* 2: 12-17.

9. Bostrom, N. 2002. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology* 9.

10. Bostrom, N. 1998. How Long before Superintelligence? *International Journal of Future Studies* 2.

11. Brin, D. 1999. *The Transparent Society: Will Technology Force Us to Choose Between Privacy and Freedom?* New York: Perseus Books.

12. Brown, J. D., K. A. Dutton and K. E. Cook. 2001. From the Top Down: Self-Esteem and Self-Evaluation. *Cognition and Emotion* 15: 615-631.

13. Cirincione, J. 2008. The Continuing Threat of Nuclear War. In *Global Catastrophic Risks*, ed. N. Bostrom and M. Cirkovic, 381-401. Oxford: Oxford University Press.

14. Drexler, E. 1987. Chapter 3:11 Engines of Destruction. In *Engines of Creation: The Coming Era of Nanotechnology*, 171-190. Garden City: Anchor.

15. Eliot, T. S. 2001. *Four Quartets.* London: Faber and Faber.

16. Fawthrop, T. and H. Jarvis. 2005. *Getting Away With Genocide: Cambodia's Long Struggle Against the Khmer Rouge.* Sydney: UNSW Press.

17. Freitas, R. 2000. "Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations." *Foresight Institute*. http://www.foresight.org/nano/Ecophagy.html.

18. Glover, J. 2001. *Humanity: A Moral History of the Twentieth Century.* New Haven: Yale University Press.

19. Goffman, E. 1959. *The Presentation of Self in Everyday Life.* Garden City*:* Doubleday Anchor.

20. Gubrud, M. 1997. "Nanotechnology and International Security." *Foresight Institute*. http://www.foresight.org/Conferences/MNT05/Papers/Gubrud/index.html.

21. Halliday, T. J. and S. Kwak. 2007. "Identifying Endogenous Peer Effects in the Spread of Obesity", Working Paper 7-27. University of Hawaii at Manoa, Department of Economics. http://www.economics.hawaii.edu/research/workingpapers/WP_07-27.pdf.

22. Hirschman, A. O. 1991. *The Rhetoric of Reaction: Perversity, Futility, Jeopardy.* Cambridge, M.A.: Belknap Press.

23. Hobden, K. L. 1997. *Behavioural Versus Claimed Self-Handicapping: Underlying Motivations and Attributions Following Failure.* Toronto: University of Toronto doctoral thesis.

24. Hurka, T. 1993. *Perfectionism.* Oxford: Clarendon Press.

25. Jonas K. 1992. Modelling and suicide: a test of the Werther effect. *Br J Soc Psychol.*, Dec 31 (Pt 4): 295-306.

26. Joy, B. 2000. Why the Future Doesn't Need Us. *Wired* 8.04, April: 238-245, 248-263.

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

28. Kantrowitz, A. 1992. The Weapon of Openness. In *Nanotechnology Research and Perspectives*, eds. B. C. Crandall and J. Lewis, 303-311. Cambridge, M.A.: MIT Press.

29. Kavka, G. S. 1990. Some Social Benefits of Uncertainty. *Midwest Studies in Philosophy* 15 (1): 311-326.

30. Kennedy, R. 1968. *13 Days.* London: Macmillan.

31. Kurzweil, R. 2006. *The Singularity Is Near: When Humans Transcend Biology.* New York: Penguin.

32. Leslie, J. 1996. *The End of the World: The Science and Ethics of Human Extinction.* London: Routledge.

33. Levack, B.P. 1987. *The Witch-Hunt in Early Modern Europe.* London: Longman.

34. Markovitch, S. and P. D. Scott. 1988. Knowledge Considered Harmful, Research Paper #030788. Ann Arbor, Michigan: Center for Machine Intelligence.

35. Matheny, J. G. 2007. Reducing the Risk of Human Extinction. *Risk Analysis* 27 (5): 1335-1344.

36. Meade, E. E. and D. Stasavage. 2008. Publicity of Debate and the Incentive to Dissent: Evidence from the Us Federal Reserve. *Economic Journal* 118 (528): 695-717.

37. Merton, R. K. 1968. The Matthew Effect in Science. *Science* 159 (3810): 56-63.

38. Montefiore, S. S. 2005. *Stalin: The Court of the Red Tsar.* New York: Vintage.

39. Moravec, H. 2000. *Robot: Mere Machine to Transcendent Mind.* New York: Oxford University Press.

40. Nelson, R. and G. Winter. 1990. *An Evolutionary Theory of Economic Change.* Cambridge, M.A.: Harvard University Press.

41. Nietzsche, F. 1984. *Vom Nutzen und Nachteil der Historie für das Leben.* Zurich: Diogenes Verlag.

42. Nietzsche, F. 2007. *On the Use and Abuse of History for Life.* Sioux Falls, S.D.: NuVision Publications.

43. Nouri, A. and C. F. Chyba. 2008. Biotechnology and Biosecurity. In *Global Catastrophic Risks* , eds. N. Bostrom and M. Cirkovic , 450-480. Oxford: Oxford University Press.

44. Oppenheimer, R. 1947. Physics in the Contemporary World, *Lecture at M.I.T.*, November 25.

45. Petratos, P. 2007. Weather, Information Security, and Markets. *IEEE Security and Privacy* 5 (6): 54-57.

46. Phillips, D. 1982. The Impact of Fictional Television Stories on U.S. Adult Fatalities: New Evidence on the Effect of the Mass Media on Violence". *The American Journal of Sociology* 87 (6): 1340–59.

47. Porter, M. E. 2004. *Competitive Advantage.* New York: Free Press.

48. Posner, R. 2005. *Catastrophe: Risk and Response.* New York: Oxford University Press.

49. Putnam, H. 1979. The place of facts in a world of values. In *The Nature of the Physical Universe*, eds. D. Huff and O. Prewett, 113–140. New York: John Wiley.

50. Radford, B. and R. Bartholomew. 2001. Pokémon Contagion: Photosensitive Epilepsy or Mass Psychogenic Illness? *Southern Medical Journal* 94 (2): 197-204.

51. Rawls, J. 2005. *A Theory of Justice.* Cambridge, M.A.: Belknap Press.

52. Rees, M. 2004. *Our Final Century: Will the Human Race Survive the Twenty-First Century?* Arrow Books Ltd.

53. Rhodes, R. 1995. *The Making of the Atomic Bomb.* Simon & Schuster.

54. Rizzo, M. J. and D. G. Whitman. 2003. The Camel's Nose is in the Tent: Rules, Theories and Slippery Slopes. *UCLA Law Review* 51: 539-592

55. Robinson, C. W. and V. M. Sloutsky. 2007. Linguistic Labels and Categorization in Infancy: Do Labels Facilitate or Hinder? *Infancy* 11 (3): 233-253.

56. Schelling, T. C. 1981. *The Strategy of Conflict*, pp. 187ff. Cambridge, M.A.: Harvard University Press.

57. Schlegel, A. 1991. Status, Property, and the Value on Virginity. *American Ethnologist* 18 (4): 719-734.

58. Shattuck, R. 1996. *Forbidden Knowledge: From Prometheus to Pornography.* New York: St.

Martin's Press.

59. Smith, T. W., C. R. Snyder and S. C. Perkins. 1983. The self-serving function of hypochondriacal complaints: physical symptoms as self-handicapping strategies. *Journal of Personality and Social Psychology* 44 (4): 787-797.

60. Stack S. 1996. The effect of the media on suicide: evidence from Japan, 1955-1985. *Suicide Life Threat Behav.* 26 (2) :132-42.

61. Stone, J. 2002. Battling Doubt by Avoiding Practice: The Effects of Stereotype Threat on Self-Handicapping in White Athletes. *Personality and Social Psychology Bulletin* 28 (12): 1667-1678.

62. Takada, H., K. Aso, K. Watanabe, A. Okumura, T. Negoro and T. Ishikawa. 1999. Epileptic Seizures Induced by Animated Cartoon, "Pocket Monster." *Epilepsia* 40 (7): 997-1002.

63. The Manhattan Engineer District. 1946. "Total Casualties." *The Atomic Bombings of Hiroshima and Nagasaki,* June 29*.* http://www.atomicarchive.com/Docs/MED/med_chp10.shtml.

64. Thompson, T. and A. Richardson. 2001. Self-handicapping status, claimed self-handicaps and reduced practice effort following success and failure feedback. *British Journal of Educational Psychology* 71 (1): 151-70.

65. Tversky, A. and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185 (4157): 1124-1131

66. Vinge, V. 1993. The coming technological singularity: How to survive in the post-human era. *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace* 11-22. United States: NASA Lewis Research Center.

67. Volokh, E. 2003. The Mechanisms of the Slippery Slope. *Harvard Law Review* 116: 1026-1134.

68. Yudkowsky, E. 2008. Artificial Intelligence as a Positive and Negative Factor in Global Risk. In *Global Catastrophic Risks*, eds. N. Bostrom and M. Cirkovic, 308-345. Oxford: Oxford University Press.

69. Yudkowsky, E. 2008. Cognitive Biases Potentially Affecting Judgement of Global Risks. In *Global Catastrophic Risks,* eds. N. Bostrom and M. Cirkovic, 91-119. Oxford: Oxford University Press.

70. Zuckerman, L., P. Hofheinz and S. Naushad. 1987. How Not to Silence a Spy. *Time*, August 17. http://www.time.com/time/magazine/article/0,9171,965233,00.html