Luusua, Aale; Ylipulli, Johanna

Artificial intelligence and risk in design

Published in:
DIS 2020 - Proceedings of the 2020 ACM Designing Interactive Systems Conference

DOI:
10.1145/3357236.3395491

Published: 03/07/2020

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Luusua, A., & Ylipulli, J. (2020). Artificial intelligence and risk in design. In *DIS 2020 - Proceedings of the 2020 ACM Designing Interactive Systems Conference* (pp. 1235-1244). ACM.
https://doi.org/10.1145/3357236.3395491

Artificial Intelligence and Risk in Design

Aale Luusua
University of Oulu
Oulu, Finland
aale.luusua@oulu.fi

Johanna Ylipulli
Aalto University
Espoo, Finland
johanna.ylipulli@aalto.fi

ABSTRACT

As artificial intelligence (AI) technologies are more and more integrated into everyday lives, both scholarly and popular discourses on AI's often revolve around charting the various risks that may be associated with them. The manner and magnitude of risk that various researchers identify and foresee varies; however, what is common between them is, undoubtedly, the concept of risk itself. This concept, we argue, has been largely taken for granted by the fields involved in the research on AI's; in other words, “risk” has been employed with an everyday sensibility without due critical examination. In this paper, we address risk as a concept directly, by examining interdisciplinary theories and literatures on risk to discuss examples of AI technologies. Through this work, we aim to begin a critical discussion of the importance of theorising risk within design research and practice, and within the development of emerging technologies.

Author Keywords

Artificial intelligence; general AI; narrow AI; urban AI; risk; subjective risk; objective risk; experience; theory.

CSS Concepts

•**Computing methodologies**; Artificial intelligence; Philosophical/theoretical foundations of artificial intelligence •**Applied computing**; Law, social and behavioral sciences; Anthropology; Sociology •**Human-centered computing~Human computer interaction (HCI)**; Interaction design; Interaction design theory, concepts and paradigms

INTRODUCTION

Societal discourse around artificial intelligence (AI) technologies has once again been ignited after two periods of relative uninterest, or “AI winters” [9, 10]. This is due to the recent rapid advances that have been made in the research of AIs, most notably the advances in new neural networks that are capable of learning autonomously from data.

AI can be described as a technology or technologies that mimic human intelligence, and may also theoretically surpass it [7, 9, 24, 28]. A more sober way of looking at AI might be to look at it as simply as algorithms that are capable of autonomous adaptation and decision-making; this is our view here. The newfound buzz around AIs has also created a slew of scholarly and popular literatures that often revolve around charting the various risks that may be associated with these technologies. Within AI research literature, however, we can identify a notable polarity between opposite approaches that we might here call “the philosophical approach to AI” and “the engineering approach to AI”. Most notably, the philosopher Nick Boström [9] has argued that (general) AI is a potentially extremely risky technology that may develop with a sudden, exponential burst, surpassing human intelligence and capabilities. The opposite of this viewpoint is that of Andrew Ng [19, 42] who maintains that there is still a lot of work to be done to make AIs even remotely ‘intelligent’. We acknowledge that many researchers fall in between these polar opposites, but we deem it important to identify and highlight that prominent researchers have such differing views on AI—views that are also echoed by the media and transmitted to the general public. These two approaches seem to hold fundamentally differing views of the potency and the associated risks of AI. While Boström tends to view the phenomenon from a bird’s eye view – taking a longer time scale as his standpoint – Ng seems to focus on the technology’s current state and extrapolates from these existing conditions, with an eye on the next few years.

While not all AI scholars and engineers can be described in this way, these two ends of the spectrum nevertheless result in wildly varying risk assessments: at the one end, AI philosophers such as Boström [9] wish to warn humanity of impending doom and subjugation by a superhuman *general AI*, whereas engineers such as Ng [19, 42] often see this as preposterous, since AI systems are still in their infancy, and building any additional functionality into these existing systems seems a toilsome project. Ng, then, seems to talk mostly about *narrow AI*, because that is what exists, at least currently. Thus, he sees the development of any human-like AI an unlikely occurrence. The risks related to AIs, then, are also seen as much lower through this approach.

Need for an interdisciplinary view

It is important to stress that both of the abovementioned positions have merit; we need the philosophical approach as well as the engineering approach in the research on AI.

DIS '20, July 6–10, 2020, Eindhoven, Netherlands.

ACM ISBN 978-1-4503-6974-9/20/07...$15.00.
DOI: https://doi.org/10.1145/3357236.3395491

However, both seem to be incomplete viewpoints on their own, and also strangely incompatible despite their usefulness. We suggest, then, that a middle scale view must be established into AI literature. Traditionally, this has been the realm and purpose of design: to take engineering abilities and technologies and ask *what, in what manner, and for whom*, should be build; it also questions what *could and should* be built. Design is an interdisciplinary field by nature, melding perspectives from art, engineering and the humanities and social sciences into theories, practices and products. This viewpoint is naturally relevant in the design of interactive systems, and AI, we could argue is among the most challenging systems that humans have ever attempted to make. It is important, then, that the design research community begins to grapple with this phenomenon. In this paper, we combine the design perspective with that of the social sciences, which traditionally have asked *why* humans do what they do, and *what does it mean* for individuals and society. By utilizing various viewpoints then, we argue for a an integrated, interdisciplinary understanding of AIs. For these purposes, in this paper we discuss AI technologies in the light of social science theories in order to inform AI design and research. In this endeavour, we utilise our expertise and separate backgrounds in architecture and urban design, and cultural anthropology. These experiences necessarily inform our approach.

Similarly, our approach is also heavily influenced by shared backgrounds in urban technology research; and indeed, as evidenced by the centrally important types of AI applications we address in this article (autonomous vehicles and facial recognition in addition to general AI), AI seems to intertwine with urban places and spaces. As computing systems in general are now a part of the lived environment [25], including homes, workplaces and public places [33], we argue that these existing urban technologies act as a gateway for the introduction of AIs. These developments are connected also with the Smart City agenda, which specifically aims to integrate more and more digital technology into urban environments. [21] Applications of these "urban AIs", as we might call them, may be found everywhere: in mobile, personal or infrastructural computing. Importantly, they have already made their way into commercial end-user applications, such as personal assistant applications (Siri), (semi)autonomous vehicles (Tesla's autopilot and various manufacturers' parking-assistant systems), and increasingly into everyday homes (Amazon Alexa, Google Home). Contemporary travel practices are fundamentally informed by AIs. Without adaptive and autonomous algorithms, we would not take the routes we now often take via car, air, rail or even by foot and bicycle; we would not use the accommodation services we use when we get there; and, we might not even visit the places that we do visit without online sites' recommendations. In short, these technologies mold our experiences of the environment in a manner never seen before. Thus, on a general level, these applications can now be said to orchestrate [6] human lives to an increasing extent. The development of AIs is also raising many ethical issues, [9,10], and as AI applications are developed further, these become intertwined with questions of city-making ethics, namely, who has the right to design and live in human environments [20]. Whatever happens with AI's, then, happens to everyday individuals. On this basis, we deem it important to study applications of AI from an interdisciplinary point of view.

The larger goal of our work is to contribute to the research of AIs, design theory, and urban research. To further these goals, the direct contributions of this paper are to (1) identify ***risk*** as an emerging theme in AI research literature, (2) argue that the ***concept*** of risk in this literature has been taken mostly for granted; i.e., it has been used with an everyday sensibility rather than with scholarly scrutinisation; (3) and explore how the concept might inform the design and research of AIs as part human lives. Through these contributions, we aim to increase designers' and researchers' understanding to inform the design of interactive systems and to start a conversation on risk in design theory at large.

UNDERSTANDING RISK

We consider it important to understand the precarious nature of design in general; any design venture must by necessity grapple with the fact that it aims to bring forth novelty, and thus, the realm where design happens is always to some degree unknown. Thus, design is inherently risky, and designers cannot but know this, as each new project brings forth a series of unknowns. Yet, a brief foray into interdisciplinary design literature reveals very scant results into the subject of risk as a concept and as a phenomenon. It seems that risk is severely under-theorised in several, if not most, design fields.

In HCI, the only contribution that we are aware of is by Klemmer, Hartmann and Takayama [23] who argue that risk is a fundamental theme for interaction design (along with four other themes they suggest). In their mind, the theme of "*Risk* explores how the uncertainty and risk of physical co-presence shapes interpersonal and human-computer interactions." [23] We agree with this, as risk is an unavoidable part of human experience; it cannot but affect also our experiences with technology and should receive much more attention as a facet of interaction design. Klemmer at al. present four aspects of experience that are affected by risk:

(1) Physical Action
(2) Sense of trust and commitment
(3) Personal responsibility
(4) Attention

All these aspects, they argue, are affected by our notion and subjective valuations of risk. Klemmer at al. base their argument on the work of Dreyfus [16]: "But where there is no risk and every commitment can be revoked without consequences, choice becomes arbitrary and meaningless."

To put this simply, if there is no pay, there can be no pay-off. Physical actions, trust and commitment, personal responsibility and attention rely on us taking some sort of a risk, either physical or social. Remove this risk (as many technologies attempt to do) and the experience of life is fundamentally altered. Thus, we argue that more empirical research and analysis should be conducted with the notion of risk in mind.

However, we must bear in mind that the view that is presented in Klemmer at al's formulation is limited to a very specific notion of risk; risk as it pertains to micro-level interaction design. This is important, but in order to understand risk and AI, we must incorporate various scales and realms of human experience. To unravel the concept of risk, we look at emerging technologies and risk as a set of social and cultural structures that vary over time and are approached differently among different cultures. This includes also subcultures of experts coming from different fields, as we argue below. We draw from sociological and anthropological theories of risk, and notably from the sociologist Ulrich Beck, who takes a macro-level view or risk, and the anthropologist Åsa Boholm, whose focus is on the micro-level of society; these conceptualizations are central when we attempt to open up the concept of risk.

Technology as a human response to risk

Humans have always grappled with uncertainty in their lives in some way. This is true regardless of time, place and culture. Risk, then, is a part of the human condition. However, societal risk management seems to be intimately tied to commonly held worldviews. As a result, over the course of human history, managing risks have included taboos, rituals and magic to ward off evil or maintain a balance [8]. In the Western world, modernism and its techno-scientific project has been a major watershed in our relationship to risk. On a macro level, this has been identified by sociologist Ulrich Beck as the birth of what he has termed *risk society* [5, 40]. The central idea behind the concept of risk society is that, as a consequence of modernity, man [sic] has sought to gain control of nature and himself [sic]. The flip side of this coin, naturally, is the acceptance of responsibility. We no longer consider negative experiences to be 'acts of God'; rather, they are events in the real world that we might avoid; thus, they become risks which must be managed. [40]. An excellent, while tragic, example of a risk society issue is the COVID-19 pandemic, which is being experienced globally at the time this article is being written. Pandemics (as opposed to epidemics) are exacerbated by modern travel technologies, which have given us the ability to traverse the globe in a matter of hours. The pandemic, then, is the dark side of this technological progress, and as a result, we are burdened by the demand of remaining constantly vigilant to control novel diseases.

This fundamental observation is at the root of Beck's risk society theory – that we cannot only consider the positive outcomes of the modern project and ignore the negatives. Thus, risk management becomes an integral part of a techno-scientific project. As a result, risk is very much enmeshed with technology. In fact, we deem it useful to offer an alternative explanation of *technology as the (non-magical) practice of attempting to mitigate or remove risk.* While this working formulation does not negate other ways of viewing technology, we can use it here to underscore an important *motivation* for the adoption of technologies, even when we do not even really want them, as might be the case with, for example, nuclear weapons. This understanding of technology from the viewpoint of risk renders it possible to see how important risk is for human experience. While the desire to mitigate risk drives technological development, it also poses new threats. Ironically, this drive for technological novelty has also produced novel technological risks: the introduction of nuclear power, air travel and telecommunications have also given rise to fears of borderless, uncontainable disasters such as nuclear fallout, spread of pandemics; and now, superintelligent AI. [6, 40]

Producing risk: Objective and subjective notions

To discuss risk as a concept further, we must have an important distinction between *subjective* and *objective risk* [34]. Objective risk refers to the statistical, quantified likelihood of an event; it is seen as being based on highly rational, scientific assessment. As Åsa Boholm writes, risk in this view is seen "as the statistical probability of an outcome in combination with severity of the effect construed as a 'cost' that could be estimated in terms of money, deaths or cases of ill health." [8] However, as humans, we manage various risks in our everyday lives all the time, without the ability to calculate risks statistically—and even if we did have such a capability, that sort of a rigid approach to life would paralyze us. Thus, subjective, intuitive valuations are still an important part of human life and risk assessment in everyday life and practice. We refer to *subjective risk*, then, when we talk about this phenomenon.

However, we must note that also statistical conceptualizations of risk are eventually based on *values*. These, in turn, vary considerably within different cultures and even among different communities belonging to the same ethnic or national group [8]. If we do not first decide what is valuable (e.g. human life) we cannot calculate "a risk". For example, if we want to assess how harmful urban pollution is for people suffering from asthma, we have already made a judgement that the lives of asthmatic persons are valuable. A mere likelihood of something occurring is not the same as risk. That something is "at risk" is an inherently human valuation that does not exist outside of human culture. We can say, then, that notions of risk are culturally produced. This was most famously argued in *Risk and Culture* [13], by anthropologist Mary Douglas and political scientist Aaron Wildavsky. They argue that 'risk' is created within culture. According to them, risk is constituted by collectively shared representations and thus, it is always socially and culturally constructed. Cultural theory thus brought a somewhat subversive perspective into risk research: prevailing theories had leaned on assumptions that individuals estimate risks based on rationality and maximization of utility. Cultural theory, in turn, does not intend to analyze risk as something individual, developing (only) within individuals' minds, but as formed by social and cultural processes. Thus, according to Douglas, for instance, risk can be construed as being fundamentally subjective and intersubjective, not inherently objective.

To produce notions of risk, a broad set of social and cultural values, perceptions and beliefs come into play. These enable us to judge situations or occurrences without a quantified or objective lens. In other words, we follow shared cultural rules, as argued by, for instance, Boholm: "Decisions about risk and management of risk are socially embedded, shaped by culturally based notions about the state of the world, what the world consists of and how it works" [8, 36, 13]. What is important here is that the objective, statistical definition of risk cannot explain how risks related to real-life, complex phenomena (such as AI) are approached and understood in practice.

Risk is neither subjective nor objective

However, Boholm [8] also takes a more in-between approach and maintains that there has been an unfruitful dichotomy between culturally and/or psychologically constructed risk and conceptualizations of risk as purely objective and calculable. Boholm argues we need to find some middle ground between "the absolute relativism of psychologically constituted 'subjective' risk – or the culturally 'constructed' risk of cultural theory" – and "the technical 'objectivist' notion of risk in terms of 'pure' de-contextualized calculations" [8]. As a more fruitful alternative to this dichotomy she offers a definition presented by sociologist Gene Rosa who concludes that risks are neither objective nor subjective. According to Rosa, risk can be defined as "a situation or event where something of human value [including humans themselves, as Boholm adds] has been put at stake and where the outcome is uncertain" [38]. The key concept here is uncertainty; without uncertainty, there is no risk. Arriving at the notion of uncertainty, Boholm adds "*This analytical perspective on risk (...) as referring to a domain of uncertainty about values and assets, and a fundamental experiential realm of human existence, for both individuals and collective, should serve as a starting point for any theory aspiring to account for the social and cultural dimensions of risk. From such a definition one can ask how people identify, understand and manage uncertainty in terms of knowledge of consequences and probabilities of events.*" [8]

We can conclude that the fundamentality of uncertainty and risk at the core of human decision-making and experience in general, is a concept that has reached widespread agreement across disciplines. We can begin to operationalise this in the following manner. In pre-modern societies, taboos, rituals and other methods were accepted ways of perceiving and understanding risk. After modernism, *modes of risk knowledge* are considerably different, as argued by Boholm [8]. In short, *everyday knowledge* is shared by everyone in everyday conversations; *scientific scenarios* are created by scientists, designers and engineers; finally, *collective narratives* are produced by the media [8]. We would like to emphasize the role of fiction writing, especially science fiction, as a part of cultural narratives. [8, 32].

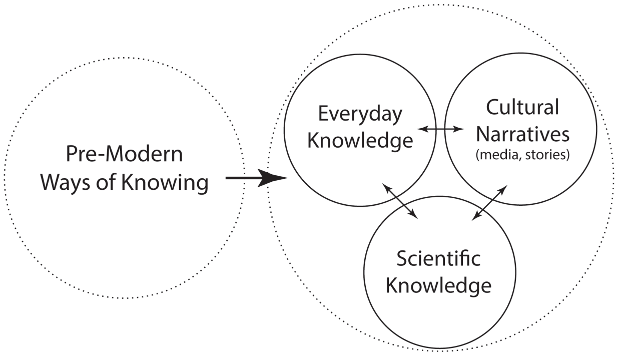

Figure 1. Combining various ways of knowing about risk. There is an interplay between contemporary ways of knowing and older ways of knowing.

We can add to this that we believe that these notions are affected by older beliefs and knowledge on all levels (Fig. 1). Cultural taboos, rituals and other ways of knowing influence our everyday lives, as not every decision we make in our lives can go through a scientific process. However, taboos and other cultural notions also affect the kinds of questions we can even ask in science: logically, we cannot identify the risks around any given subject if the subject itself is too taboo to even think about. This is where storytelling might play a crucial part. We can more readily work around taboos if we bracket them as make-believe. As argued by [45, 3, 14], science fiction produces scenarios connected to especially emerging technologies, and therefore, it is central realm for producing and circulating risk knowledge connected to those technologies. Even though we know science fiction is not 'real', it still affects significantly our collective imagination and perceptions on what is possible. Science fiction, then, can be a fruitful way for designers to grapple with technology risks and AI as well.

Towards a holistic way of understanding risk

What other, non-narrative ways there might be to identify and analyse risk and AI? In economic anthropology, there is a distinction between two aspects of uncertainty: 1) Known risks that we are prepared for; we already have developed strategies to cope with those risks, and 2) things that are more obscure or unknown, and we do not have any established procedures or ways to manage them [8]. This is actually similar to notions held in certain engineering organisations and the military. Donald Rumsfeld made this famous in 2002 [39], when he spoke of the *known knowns*, the *known unknowns* and the *unknown unknowns* in relation to military strategy. The phrase 'unknown unknowns' was apparently first introduced by psychologists Joseph Luft (1916–2014) and Harrington Ingham (1916–1995), in their Johari window [27], a matrix consisting of space to identify and organise issues that are known to one's self and to others; known to one's self and not others; not known to self and known to others; and unknown to both self and others, in other words, unknown unknowns. In the case of AI risks, we can utilise this framework to identify, organise and discuss uncertainty, and thus begin to accumulate knowledge in a meaningful way.

In the case of unknown unknowns, rational choice does not work very well, as there is not enough information to make any kind of statistical, rational calculation. When confronted with this type of uncertainty, people tend to turn towards more culturally informed strategies, as argued by Boholm [8]. These can refer to what is conventionally understood in a society/community as "valid", "true" and "normal", regardless of rationality. In other words, *culturally based cognitive shortcuts* are activated. Unlike computers, humans never produce error messages; we rationalise, obscuring the actual reasoning behind our decisions. This is a known process among cognitive psychologists who refer to these as 'schemata' or 'scripts'—heuristics that simplify a problem, for example "a binary structure of morally loaded mutually exclusive alternatives, or as a situation without alternatives" as Boholm [8] writes. In other words, it leads to black-and-white thinking.

Risk, for humans, is a dynamic relational order of meaningful connections between things, and not something we can objectively calculate or measure [8]. This is why it is so difficult, or impossible even, to accurately ascertain risks as they pertain to new technologies. Different individuals and different disciplines that work with and/or around novel technologies have vastly differing ontologies, epistemologies and value systems in the first place. They are different scientific subcultures or, we could even say, scientific tribes that educate their members to hold different schemata, which are activated in practice.

According to Boholm [8] cultural schemata around risk produce contexts which connect:

(1) an object of risk (a source of potential harm)
(2) an object at risk (a potential target of harm) and
(3) an evaluation (implicit or explicit) of human consequences.

From this perspective we can see that risk is experienced not as an essential property of things that is simply perceived, but as an inherently dynamic relational order of meaningful, culturally assigned connections. This perspective on the 'cultural nature' of risk makes it possible to theorize the variation in the conceptualization and management of risks among different communities or organizations. Risk, then, is experienced through a cognitive framework that produces these linkages and their meaning or content. In the case of AI, these might be:

(1) What is AI? (the object of risk)
(2) What is at risk? (the object at risk)
(3) What does it mean that it is at risk? (evaluation)

A philosopher of AI, for example, might say that "AI is a computer that works like a human mind, but might be vastly more efficient; what is at risk, is the whole future of humanity; this is very alarming and the risks are very high indeed." However, an AI engineer might feel that "AI is not very effective currently, and really consists of just certain types of algorithms to make operations more effectively and independently. What is at risk is that it will never work. This means that we will be deprived of some technological progress, and while technological progress is very valuable, it is not very risky overall." The answers, then, to these questions are likely to be very different across different disciplines and individuals. In all likelihood, they will not produce the same linkages and meanings. A more comprehensive view on AI and risk necessitates interdisciplinary collaboration.

RISK AND AI

Applying this theoretical understanding to AI applications is a large undertaking that can only be begun here; to chart the territory for this work, however, we identify three important technologies that should be examined from a holistic point of view on risk, namely, general AI (GAI), autonomous vehicles (AVs) and facial recognition (FR).

General AI: Non-humans in our lives

When first confronted with the notion of AI, we often think of a fully-fledged intelligence that is very autonomous, adaptive, and human like – often superhuman. This is would be a GAI, and this is the basis from which notable scholars and fiction writers have also approached the topic. Most recently, Nick Boström [10] discussed the idea of a super-human artificial intelligence. Bostrom's central thesis could be described as a warning about the so-called hockey stick phenomenon; an exponential, explosive growth in the intelligence level of an AI system that might occur in microseconds, once the prerequisite amount of slow growth has been achieved.

While several notable AI engineers, such as Andrew Ng [19, 42] have considered this position highly improbable, there is something in Boström's argument that makes it permanently compelling. After all, while this AI explosion has not happened, nobody can ever say that it absolutely will not happen. This is reminiscent of Beck's [5] understanding of risk society. As long as it remains a risk, it will always linger. Risks that have occurred are no longer risks. We can hypothesise that *if* we assume that AI can develop into a general, human-like intelligence, *then* the risk of a superhuman intelligence developing out of it will always be in issue, much like the risk of a nuclear reactor malfunction will always exist while the technology itself exists. However, GAI does not exist and cannot be evaluated empirically. As such, this is in the realm of the known unknowns (Figure 2).

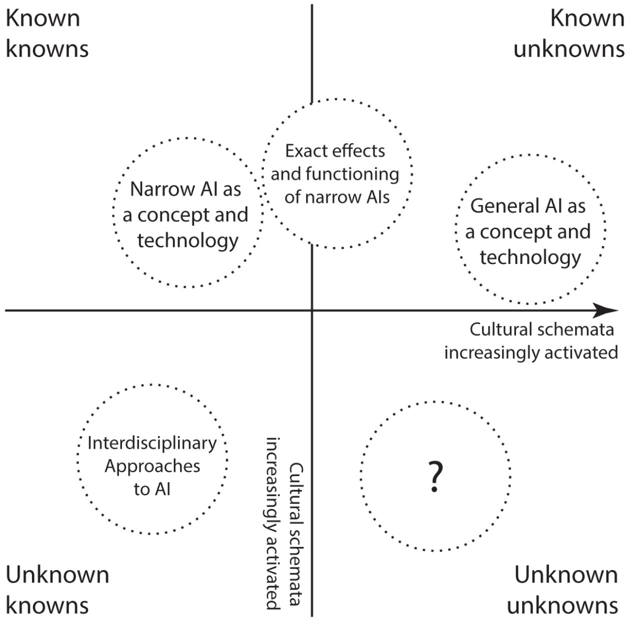
Known knowns
Known unknowns
Narrow AI as a concept and technology
Exact effects and functioning of narrow AIs
General AI as a concept and technology
Cultural schemata increasingly activated
Interdisciplinary Approaches to AI
Cultural schemata increasingly activated
?
Unknown knowns
Unknown unknowns

Figure 2. Combining a knowns/unknowns matrix with the spectrum of cultural schemata activation.

However, narrow AIs, such as the various algorithms that make discrete autonomous decisions on our behalf and learn from our behavior, are definitely in the realm of known knowns (Figure 2); they exist and we already know fairly well how they function even on an experiential level. While more empirical knowledge should be produced about the effects of these technologies on everyday life experience, they are still in our realm of experience. Therefore, we can estimate that existing narrow AIs do not activate our cultural schemata as aggressively, since we can rely on both established experiences and objective data. However, we do not necessarily know the cumulative effects of the exact functioning of all narrow AIs, such as image recognition algorithms. Understandably, then, there is much more debate concerning the effects of these algorithms in, for example, work discrimination or even doling out justice (47, 46). However, even the much more familiar and mundane narrow AIs do have vastly important effects on how travel as a huge industry will shape our lives, and indeed, the whole planet – yet societal discussion around recommendation algorithms is scant, as they are so experientially familiar to us.

Will there be a singular non-human agent that, for our "own good" controls us? This question is no doubt in the realm of the known unknowns: we know that this might be the case but we cannot calculate or estimate any kind of a probability of this happening. Here, then, the role of various cultural schemata will increase. We think of the intelligent machine as inherently other and, as such, as potentially hostile and dangerous [30, 11, 2]. There are also a number of culturally important and widespread narratives of technological hubris ending in disaster, such as Icarus and the tower of Babel, and the real-life story of the Titanic. The realm of the known unknowns is not a comfortable terrain for engineering, since the value systems of engineering cultures put emphasis on empirical evidence, concrete events, and measurable outcomes that produce data from those events. Other fields, however, are perhaps more comfortable with dealing with subjective-ness, such as design and the social sciences; both attempt to understand subjective and intersubjective positions, the former to understand the *why* and designers to understand *how*; more specifically, how to imagine and implement artefacts.

In design fields, ideas that fall into the realm of unknown knowns (and inch toward the unknown unknown) are traditionally tested as visualised utopias. The first utopia was of course presented by Thomas More in *Utopia* [31], a narrative which presented a holistic vision for society. This ushered in utopias as a genre. Some of the most influential utopias in the 20th century, however, came from the world of urban planning, including Le Corbusier, Frank Lloyd Wright and Ebenezer Howard [18]; these were mainly visual and functional utopias without a narrative element. While this approach is not without its problems [15], it has remained as a staple among designers' tools in some form or another. Other speculative methods can be seen as being related to it: Concepts are a sort of utopia-light. Contemporary concepts come in the shape of audio-video, physical models and prototypes, and more rarely as virtual reality scenarios. Concepts and prototypes are definitely crucial for technology industries for both internal development and customer relations. Similarly, in design research and development, numerous methods from scenarios to personas have been used as aids to help designers empathise and troubleshoot [43]. Design fiction can be a useful method of generating materials and understanding on phenomena that are not experienced in everyday life. [45]

These traditional methods that have their origins in fiction and design can be joined with social scientific theories and empirical methods to give them real-world relevance, depth of understanding, and extend their ethical dimensions. Crucially, being aware of social scientific theories, such as the theories of risk, enable designers to distance themselves from their own deeply embedded, knee-jerk reactions that ensue from cultural schemata being activated. This is an invaluable tool for both social scientists and designers, as our imaginations are limited by what we deem valid, true and normal. Furthermore, they enable us to connect discrete studies into a larger whole. It is crucially important that we marry imagination with empirical reality in the study of AIs as they are about to be tightly coupled with our everyday lives. [41, 14, 44]

Overall, then, since techno-social systems are rapidly gaining complexity, it will no longer be adequate to gauge their effects post-facto through fundamental research—research in all fields will have to push itself further and further into the known unknowns, and, perhaps, even into the unknown unknowns. The border between fact and fiction may be much more porous than we have previously thought when it comes to AIs, and especially GAI. This calls for an ambitious development of novel methods.

Autonomous vehicles: more than a trolley problem

Autonomous vehicles, or AVs, present a famous case of a "known unknown" which has excited the imaginations of scholars. The article by MIT scholars [29] titled, "Why Self-Driving Cars Must Be Programmed to Kill" tackled the problems of AV killings (which we are beginning to experience in real life) argued for a utilitarian proposition: since AVs will eventually find themselves in a situation where it will have to choose one life over another, we should programme kill algorithms that would, somehow, find the most acceptable course of action for the machine in a morally difficult situation, and choose to sacrifice one human over, presumably, another human.

JafariNaimi has argued against this proposition [48]. Importantly, she argues for the contextualization of moral problems. In JafariNaimi's view, contextualization renders many philosophical problems non-existent. In other words, life is messy [14]; two real-life options can never be fully identical, since one always has slightly different characteristics or circumstances than the other. We agree with this basic proposition. Here, we see the threshold between pure engineering and design; the adaptation of a technology into everyday use is not an engineering problem. It is a design problem, and design can never be de-contextualised. This is why design will always be a profession that requires subjective judgment of individual contexts and use cases.

This represents a major difference in worldviews. By accepting that every context is different, we accept fundamentally that the world is an uncertain place that is not conducive to full-on risk negation. By accepting uncertainty, we do not have to make unacceptable choices. It is false to imply that we humans would be so in control of our world that machines that we make could dole out life-or-death justice without human intervention. In fact, by engineering a decision like this, we could no longer call any AV fatality an accident; rather they would be either death sentences (if legal) or murders (if illegal). It is an interesting question whether these "moral algorithms" would even be unconstitutional in most countries, for example, in the U.S. where the law requires a citizen to be tried by a jury of peers [12].

If we accept moral algorithms, we accept responsibility for modelling the entire world and as such, we accept a responsibility for everything – we should contain every disaster and foresee every issue, and this burden of protection is placed upon technology to realise. However, we are not able to bear this responsibility to its conclusion. We cannot fully control external reality; we can only design or engineer it to a limited degree, and always from a place of existing within it. This is an important part of what contextualization means. However, we would disagree with JafariNaimi's proposition that we can simply choose to "not do" AVs if we cannot make them fully safe [48]. While techno-determinism should be criticised heavily, the non-adoption of technologies does not, in an empirical sense, seem to quite work this way. Technology adoption is very tightly coupled with power relations. The weapons industry is the best example of technology adoption that is not beneficial for humanity; yet we cannot seem to cease their production. Indeed, AV features have already been integrated into consumer products in the form of cruise control, lane assist and auto-navigation.

Facial recognition: emplaced privacy and safety

What, then, could help us reign in some of the more harmful aspects of these technologies? At this point in time, the European Union is attempting new regulatory measures on facial recognition technology. More specifically, the EU is looking to ban FR for the next five years in order to have time to assess risks associated with the technology [4]. There have already been concerns over the unethical use of FR in public areas, for example, in the UK [22]. China's approach to FR is the polar opposite of the EU. In China, FR is utilised by the government to observe pedestrians and even pharmacy transactions. This highlights the importance of cultural differences in the adoption of technology – and the notion of risk. In Europe, the risk is seen as citizens losing their right to privacy, and this seems to be the position of European governments, at least the European Union. In China, the government considers the risk of illicit or even merely antisocial activities as a larger risk than the risk of these technologies being used against private citizens.

Still, FR has already been accepted into everyday use in airports. This represents an interesting case: While Europeans do not accept the use of FR in their streets, they are somewhat more willing to do so at airports. Why is this? We argue that context can explain this phenomenon, as all human experiences are emplaced [35]. The airport is an odd place, a "non-place" even [1]. While they are nominally located inside different countries, airports form a network of places that are culturally gray areas. They have their own rules which vary little from country to country. They cannot easily be categorised. Additionally, they are seen as being high risk due to terrorism and international crime. It is arguably easier to adopt ethically conspicuous technologies in these places, where we deem the risks to be high. Thus, we are willing to go through a number of uncomfortable and intrusive rituals to get on an airplane, and these, we trust, keep us safe from harm. That is their subjective meaning, despite being developed by the aviation industry as objective measures. Risk, then, is inherently emplaced, as our judgments of what is risky is context driven.

CONCLUSIONS

In this paper, we examined the concept of risk in detail, and applied resultant understandings to discuss general AI, AV's and FR. Next, we will conclude our discussion by briefly distilling some key takeaways from our analysis. We will also present three ideas to help designers grapple with the notion of risk.

Key takeaways on technology design and risk

At its core, risk remains an inherently human concept. That something is at risk is a statement that must be based on values. Thus, by examining critically what we need to be "at risk" at different times, we can make interpretations about what is valuable—and what is not. For example, while the EU is considering a ban on FR technologies, no such drastic halt has been proposed in the design and implementation of autonomous driving properties in cars. What does this say about European values concerning mobility, privacy and individual freedom?

Studying the concept of risk may make it also easier to see why we are reluctant to ban or discard technologies even when we see them as dangerous in and of themselves. We all have different notions of what is at risk; one person might highlight the risk to humanity, whereas another would emphasise national security. Within the latter style of thinking, the risk of being subjugated by our own technology may seem as being lesser than the risk of being subjugated by other humans.

Risk also has profound implications for human experience and sense-making and designing for minimal risk can be problematic. AIs, such as AVs, and technology in general, aim to make our lives safer and safer by mitigating human and environmental risks. However, if we relocate decisions around risk to machine intelligences, there is a risk here that human lives might increasingly lack in meaning and feelings of autonomy – and these are absolutely central to human wellbeing.

Risk, then, is a design decision as much as any other. This should be made explicit in design processes, user understanding, and user participation. What is the risk to the person? What do they value? What do I, as a designer or researcher, value? Is there a conflict? Am I putting things valued by others at risk in ways that I do not understand (because I do not understand what they value)? Furthermore, designers must necessarily work within the framework of a "risk society" – a society that is very much vulnerable to global risks, such as pandemics, while holding institutions and individuals responsible for avoiding or mitigating disasters. Within a risk society, AI can be either seen as a risk itself, or, possibly also the solution to avoiding a risk. The COVID-19 pandemic represents a tragically perfect example of a risk society issue, where the problem could not be confined to any country and no clear, singular solution to the problem existed. Instead, countries had to consider their shared values in deciding which measures to take; for example, to what extent they will sacrifice citizen's privacy to take advantage of technology to perform surveillance to control the spread of the disease. In this moment, AIs are also far more easily seen as a desirable technology to help us control this uncertain situation.

We suggest three ideas to deal with risk in design of AIs

1) *Firstly, on the deepest level, we should accept uncertainty as an essential part of human life*. Since we cannot fully understand the world, we cannot fully control it either. Accepting full responsibility for natural disasters and human errors is not feasible. By accepting this, we do not have to make unacceptable design choices that de-contextualise life and death decisions. We can also refer to the risk evaluation framework of Boholm, which begs us to ask the three basic questions:

(1) What is AI? (the object of risk)
(2) What is at risk? (the object at risk)
(3) What does it mean that it is at risk? (evaluation)

These questions cannot be answered once-and-for-all; rather, they are questions that designers should reflect on during every design project, preferably as explicitly as possible.

2) *Secondly, we emphasise contextualization as a key strategy in studying the ethics of AIs.* Context enables us to imagine and study more pragmatic, real-world based ethical problems and solutions to the various challenges brought about by AIs. Context takes innumerable forms, but based on our short discussion on FR, we can tentatively argue for the central importance of *place-based* design in this work.

3) *Third, we should consciously harness speculation and see it as a part of AI designers and researchers' toolbox.* Since moving toward the unknown activates our cultural schemata, we suggest that we may have some choice here: Rather than merely succumb to the tyranny of these shortcuts, we can be more deliberate and attempt to harness our imagination and cultural sensibilities in a constructive and critical manner. Speculative methods that activate the imagination are likely most suited to grapple with this in AI research as well. Due to the increasing pace of technological progress, we argue that more attention will have to be given to these sorts of methods.

How AIs should and should not be implemented is a design question where the stakes are enormous. While bans and non-adoption are useful measures to provide society with the time to build a better understanding of these risks, it would not seem plausible that the development of AIs will be completely halted; in fact, nobody seems to be even seriously suggesting it. If we pursue AIs further, what remains then, is contextualization and values-driven work. Understanding how AIs translate into differing contexts, identifying the knowns and unknowns, and recognising our values will have to guide this development. This is work that must be undertaken by an interdisciplinary design research community.

It is also important to note here, that this discussion of risk cannot and should not be used to foster an attitude of complete nonchalance when it comes to risk-taking. To accept risk as a part of life does not mean that anything goes. In fact, it means the opposite, since all decisions that concern a risk implicitly communicate and contain our values.

Unnecessary risk-taking, then, is always an affront to our deepest-held ideas of what is important. As we move toward the unknowns, and uncertainty, these technologies touch on our deepest fears—loss of autonomy, loss of being able to make sense of the world, and loss of life. These fears may well be justified, and we must scrutinise them. However, these do not issue from the machine alone. They also come from our own values, as these are inevitably baked into the artefacts we make. If we decide to design and make the unacceptable, then we will experience the unacceptable.

Finally, while AI is a particularly relevant technology in relation to risk, we recognize that our discussion of risk here may also be applied to the design of any interactive system, or, indeed, any design project. As we have repeatedly stated, risk is a central element of all design. Thus, it should never be taken for granted, merely as something vaguely negative and uncritically scrutinised. We welcome the design research community into a critical examination of risk and how it should be addressed through and in design.

ACKNOWLEDGEMENTS

Dr. Luusua gratefully acknowledges the grant from the Academy of Finland (grant number 316136).

REFERENCES

[1] Marc Augé. 1992. *Non-Places: An Introduction to Anthropology of Supermodernity.* Verso.

[2] Yoseph Bar-Cohen and David Hanson. 2009. *The coming robot revolution: Expectations and fears about emerging intelligent, humanlike machines.* Springer Science & Business Media.

[3] Jeffrey Bardzell and Shaowen Bardzell. 2014. "A great and troubling beauty": cognitive speculation and ubiquitous computing. *Personal Ubiquitous Comput.* 18, 4 (April 2014), 779-794.

[4] BBC. 2020. Facial recognition: EU considers ban of up to five years. Retreived Jan 30, 2020 from https://www.bbc.com/news/technology-51148501/

[5] Ulrich Beck. 1992. *Risk Society: Towards a New Modernity*. SAGE.

[6] Mikkel Bille and Tim Flohr Sørensen. 2007. An anthropology of luminosity: The agency of light. *Journal of material culture*, 12(3), 263-284.

[7] Margaret A. Boden. 2016. AI: Its nature and future. Oxford University Press.

[8] Åsa Boholm. 2003. The cultural nature of risk: Can there be an anthropology of uncertainty? *Ethnos* 68, 2, 159-178.

[9] Nick Boström, & Eliezer Yudkowsky. 2014. The ethics of artificial intelligence. The Cambridge handbook of artificial intelligence, 316-334. Cambridge University Press.

[10] Nick Boström. 2014. Superintelligence: Paths, dangers, strategies. Oxford University Press.

[11] Stephen Cave and Dihal Kanta. 2019. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence* 1, 2, 74-78.

[12] United States Government. Constitution of the United States of America: Analysis, and Interpretation – Centennial Edition – Interim" (PDF). S. Doc. 112-9. Washington, D.C.: U.S. Government Printing Office. Retrieved Jan 30, 2020 from https://www.govinfo.gov/content/pkg/GPO-CONAN-REV-2014/pdf/GPO-CONAN-REV-2014-9-4.pdf/

[13] Mary Douglas and Aaron Wildavsky. 1982. *Risk and culture: An essay on the selection of technological and environmental dangers.* University of California Press.

[14] Paul Dourish and Genevieve Bell. "Resistance is futile": reading science fiction alongside ubiquitous computing. *Personal Ubiquitous Comput.* 18, 4, 769-778.

[15] Steven Dorrestijn and Peter-Paul Verbeek. 2013. Technology, wellbeing, and freedom: The legacy of utopian design. *International journal of design*, 7, 3.

[16] Hurbert Dreyfus. 2001. Disembodied Telepresence and the Remoteness of the Real, in *On the Internet*. Routledge. pp. 50-72,.

[17] Anthony Dunne and Fiona Raby. 2013. *Speculative everything: design, fiction, and social dreaming.* MIT press.

[18] Robert Fishman. 1982. *Urban Utopias in the Twentieth Century: Ebenezer Howard, Frank Lloyd Wright, and Le Corbusier.* MIT Press.

[19] Caleb Garling and Andrew Ng. 2020. Why 'Deep Learning' Is a Mandate for Humans, Not Just Machines. Retrieved Jan 30, 2020 from https://www.wired.com/brandlab/2015/05/andrew-ng-deep-learning-mandate-humans-not-just-machines/

[20] David Harvey. 2003. The right to the city.*International journal of urban and regional research* 27, 4, 939-941.

[21] Robert G. Hollands. 2008. Will the real smart city please stand up? Intelligent, progressive or entrepreneurial? *City* 12, 3, 303–320.

[22] Leo Kelion. 2019. Met Police gave images for King's Cross facial recognition scans. BBC News. Retrieved Jan 30, 2019 from https://www.bbc.com/news/technology-49586582/

[23] Scott R. Klemmer, Björn Hartmann and Leila Takayama. 2006. How bodies matter: five themes for interaction design. In *Proceedings of the 6th conference on Designing Interactive Systems* (DIS'06), 140-149.

[24] Richard Kurzweil. 2005. *The singularity is near: When humans transcend biology*. Penguin.

[25] A. Luusua, Johanna Ylipulli, Hannu Kukka and Timo Ojala. 2017. Experiencing the Hybrid City: The role of

digital technology in public urban places. In: Hannigan J & Richards G (eds) *The SAGE Handbook of New Urban Studies*. SAGE.

[26] A. Luusua, Henrika Pihlajaniemi and Johanna Ylipulli. 2016. Northern Urban Lights: Emplaced Experiences of Urban Lighting as Digital Augmentation. In *Architecture and Interaction*, 275-297. Springer.

[27] Joseph Luft and Harry Ingham 1955. The Johari window, a graphic model of interpersonal awareness. In *Proceedings of the Western Training Laboratory in Group Development*. University of California.

[28] John McCarthy and Patrick J. Hayes. 1969. Some philosophical problems from the standpoint of artificial intelligence. *Readings in artificial intelligence*, 431-450.

[29] MIT Technology Review. 2015. Why Self-driving Cars Must Be Programmed to Kill. Boston, MA. Retrieved Jan 30, 2020 from http://www.technologyreview.com/view/542626/why-self-driving-cars-must-be-programmed-to-kill/

[30] Alison Mountz. 2009. "The Other". *Key Concepts in Political Geography*, 332. SAGE.

[31] Thomas Moore. 1516/1967. Utopia, trans. John P. Dolan, in *The Essential Thomas More*, James J. Greene and John P. Dolan (eds.), New American Library.

[32] Isak Niehaus. 2002. Bodies, heat, and taboos: Conceptualizing modern personhood in the South African Lowveld. *Ethnology* 41, 3, 189-208.

[33] Richard Oldenburg. 1989. *The great good place: Cafés, coffee shops, community centers, beauty parlors, general stores, bars, hangouts, and how they get you through the day*. Paragon House Publishers.

[34] Nick Pidgeon, C. Hood, D. Jones, B. Turner and R. Gibson. 1992. Risk Perception. In *Risk: Analysis, Perception and Management.* The Royal Society.

[35] Sarah Pink. 2011. From embodiment to emplacement: re-thinking competing bodies, senses and spatialities. Sport, Education and Society, 16, 3, 343-355.

[36] Kristen Purcell, Lee Clarke and Linda Renzulli. 2000. Menus of Choice: The Social Embeddedness of Decisions. In *Risk in the Modern Age: Social Theory, Science and Environmental Decision Making*, (ed) M.J. Cohen. Macmillan Press.

[37] Ortwin Renn. 1998. Three decades of risk research: accomplishments and new challenges." *Journal of risk research* 1, 1, 49-71.

[38] Eugene A. Rosa. 1998. Metatheoretical Foundations for Post-Normal Risk. *Journal of Risk Research* 1, 1, 15-44.

[39] United States Department of Defence. 2002. DoD News Briefing – Secretary Rumsfeld and Gen. Myers, United States Department of Defense. Retrieved Jan 21, 2020 from https://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636/

[40] Jonathan Van Loon. 2013. *Risk and technological culture: Towards a sociology of virulence.* Routledge.

[41] Alex Wilkie, Martin Savransky and Marsha Rosengarten. 2017. *Speculative Research: The lure of possible futures.* Routledge.

[42] Chris Williams. 2015. AI guru Ng: Fearing a rise of killer robots is like worrying about overpopulation on Mars. Retrieved Jan 30, 2020 from https://www.theregister.co.uk/2015/03/19/andrew_ng_baidu_ai/

[43] Kim Goodwin. 2009. *Designing for the Digital Age.* Wiley.

[44] Johanna Ylipulli, A. Luusua and Timo Ojala. 2017. On Creative Metaphors in Technology Design: Case "Magic". In *Proceedings of the 8th International Conference on Communities and Technologies*. (C&T'17) 280-289). ACM Press.

[45] Johanna Ylipulli, Jenny Kangasvuo, Toni Alatalo and Timo Ojala. 2016. Chasing Digital Shadows: Exploring future hybrid cities through anthropological design fiction. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, 1-10. ACM Press.

[46] CEPEI 2020. European Commission for the Efficiency of Justice (CEPEJ). Retrieved Jan 30, 2020 from https://www.coe.int/en/web/cepej/justice-of-the-future-predictive-justice-and-artificial-intelligence/

[47] Buolamwini, Joy (Feb 7, 2019). Artificial Intelligence Has a Problem With Gender and Racial Bias. Here's How to Solve It. [Accessed Jan 30 2020] <https://time.com/5520558/artificial-intelligence-racial-gender-bias/>

[48] Nassim JafariNaimi. 2018. Our bodies in the trolley's path, or why self-driving cars must* not* be programmed to kill. *Science, Technology, & Human Values*, 43, 2, 302-323.