AI & SOCIETY (2020) 35:663–673
https://doi.org/10.1007/s00146-019-00924-9

OPEN FORUM

The role of experts in the public perception of risk of artificial intelligence

Hugo Neri[1] · **Fabio Cozman**[1]

Received: 22 May 2019 / Accepted: 31 October 2019 / Published online: 15 November 2019

Abstract
The goal of this paper is to describe the mechanism of the public perception of risk of artificial intelligence. For that we apply the social amplification of risk framework to the public perception of artificial intelligence using data collected from Twitter from 2007 to 2018. We analyzed when and how there appeared a significant representation of the association between risk and artificial intelligence in the public awareness of artificial intelligence. A significant finding is that the image of the risk of AI is mostly associated with existential risks that became popular after the fourth quarter of 2014. The source of that was the public positioning of experts who happen to be the real movers of the risk perception of AI so far instead of actual disasters. We analyze here how this kind of risk was amplified, its secondary effects, what are the varieties of risk unrelated to existential risk, and what is the dynamics of the experts in addressing their concerns to the audience of lay people.

Keywords Artificial intelligence · Social impacts of artificial intelligence · Risk · Risk perception · Experts

1 Introduction

In this paper, we apply the social amplification of risk framework [SARF] (Renn et al. 1992; Kasperson et al. 1988, 2003; Pidgeon et al. 2003) to the public perception of artificial intelligence [AI] using data collected from Twitter. A major finding presented in this paper is that we identified that experts are the real movers of the risk perception of AI. Public awareness of Artificial Intelligence has been growing since the second half of the last decade. This is a new wave of popularization of the term AI. The first wave occurred at the beginnings of the discipline, during the 1960s and early 1970s. This new wave is relevant because the phrase "artificial intelligence" became popular (i.e., Twitter users tweet more about AI) after 2014, which coincides with experts' public interventions.

In this paper, we take the set of public messages available on Twitter that explicitly used the phrase "artificial intelligence" from 2007 to 2018 as an abstract representation of the public awareness of AI. A portion of this public awareness refers to the anticipation of likely negative consequences related to the variety of applications of AI as a technology. We call that as the public perception of risk of AI or merely the risk perception of AI. Even though the risk perception is just a fraction of the public awareness of AI, according to our data, its rate of growth is higher than the public awareness excluding the risk perception (Fig. 2).

Classical risk analysis takes the concept of risk as the statistical expectation of unwanted events and the magnitude of their consequences (Freudenburg 1988). Social scientists and psychologists often question this technical approach to risk because it ignores epistemological, sociological, and subjective dimensions. There are five critical approaches to risk within social sciences: (a) the cultural approach (Douglas 1985, 1986, 1990, 1992); (b) the edgework approach (Lyng 1990); (c) the governmentality approach (Foucault 1978, 1980, 1982, 1991), (d) the risk society approach (Beck 1986, 1999, 2007), and (e) the social systems approach (Luhmann 1993). The problem with these critical approaches is that they do not offer a framework of analysis for the diffusion of behavior towards risk within society. For this reason, in this paper we assume as a guideline the SARF (Kasperson et al. 1988; Pidgeon et al. 2003), which based on the concept of risk as risk perception (Slovic 1986; Slovic et al. 2000, 2004).

✉ Hugo Neri
hugo.munhoz@usp.br

Fabio Cozman
fgcozman@usp.br

[1] Center for Artificial Intelligence, University of São Paulo, São Paulo, Brazil

In this framework, risk perception is the experience of risk not as physical harm, but as "the result of a process by which individuals or groups learn to acquire or create interpretations of hazards. These interpretations provide rules of how to select, order, and often explain signals from the physical world" (Slovic 2000: 140). SARF is a chain of factors that describes the life cycle of risk perception. It is about how individuals form messages from the selection of specific characteristics of a hazardous event and communicated them to others (Renn 1991). In a communication metaphor, the reception of a message by other individual is similar to stations that receive messages encoded in a signal. The station decodes and evaluates the messages and then communicates them further either amplifying or attenuating them. The attenuation and amplification rely on how well a message matches one's previous beliefs. The amplification occurs during the transmission and the reception. Besides, a transmitter "is also a new information source … during the reception of information and in recoding" (Kasperson et al. 1988, p. 237). There is also the propensity to take actions to the risk, and that may lead to behavioral patterns as well as secondary social or economic consequences (Kasperson et al. 1988). These secondary consequences are usually called ripple effects.

However, SARF has some limitations. Since the framework was developed 30 years ago, the process of risk communication was deeply dependent on traditional mass media. As a consequence, individuals had a limited role "unless they are eyewitnesses of risk events or directly affected by a cause of a risk" (Renn 1991, p. 302). With social media like Twitter, the individuals' role has changed to a more active and influential stations. In early investigations of risk perception in online engagement, Chung (2011) shows that the differences between online engagement and media coverage suggest that the sheer volume of news media does not represent public concern or interest in an issue. Studies that apply SARF to social media data are new (e.g., Fellenor et al. 2018; Strekalova and Krieger 2017; Witrz 2018). Witrz (2018) also summarizes the key findings of the new studies applying SARF to social media data. Social media, (a) have an amplifying effect on emotions that the other mediums such as online forums and traditional newspapers did not have (Chong and Choy 2018); (b) allow a more direct view of the perspectives of a range of publics and stakeholders" (Fellenor et al. 2018, p. 14); (c) reconfigure the classification of direct and indirect information sources and social stations (Fellenor et al. 2018; Zhang et al. 2017; Zhou et al. 2017); and (d) social media also enable individuals to amplify/attenuate signals/information from official sources/stations (Strekalova and Krieger 2017). Regarding Twitter data, we share the spirit of Sloan et al. (2013) in the context of SARF, ''Twitter can be conceptualised as a 'digital agora' (Sloan et al. 2013) that provides an insight into mass user-generated opinions, sentiments and reactions to social events.''

Back to AI, neutral or beneficial secondary effects of the risk perception of AI can be seen in the addressing of ethical issues in the field of AI as an attempt to mitigate undesired side effects. Such side effects range from the more plausible scenarios such as enhancement of discrimination to less plausible like a malevolent general AI. These corresponding actions became both parts of already existing organizations, and they help in the creation of new organizations. As example, one has Machine Intelligence Research Institute (2000) and the Future of Life Institute (2014) in the United States, or the Future of Humanity Institute (2005), the Centre for the Studies of Existential Risk (2012), and the Leverhulme Centre for the Future of Intelligence (2015) in the United Kingdom. This has led to companies' foundations like Elon Musk's OpenAI (2015), as well as government reports like the White House's National Science and Technology Council Committee on AI (NSTC 2016) and the House of Representatives Bill to establish the National Security Commission on Artificial Intelligence (Stefanik 2018) in the US, or the House of Commons' Science and Technology Committee on Robotics and AI (HC 2016) and the House of Lords' Committee on Artificial Intelligence (HL HL 2018) in the UK.

Because of these social and economic effects, we know that the risk perception of AI has been going through the SARF process. If we look for the hazardous events of AI as a cause for the rise of its risk perception, we find just a few situations with significant undesired effects. Examples on the media were the Knight's Capital Group's bankruptcy due to an unexpected pattern of trading of an AI program and a fatal crash of a Tesla car. However, do these examples explain the increase in the risk perception of AI? If it is the case, we could find in the tweets a significant number of associations of adverse events that happened and AI. During the period analyzed, it was not the case. Anyway, we want to make clear that the public perception of the cause of the accidents has nothing to do with the real causes of accidents, i.e., whether it was connected to AI or not. It does not imply either that at any point in the future the public cannot start to associate AI to such events.

Since we could not assign a single or a set of hazardous events to the social amplification process of the risk perception of AI, there has just been another factor that triggers this process. Our main claim in this paper is that in the AI case, some experts played the role of being the primary source of the formation of the message containing the risk perception. However, they did not lose their role of stations as risk communicators either amplifying or attenuation the messages. We want to stress that the concern about AI is not new, but our aim here is to cover the most recent wave of risk perception of AI.

In the next session, we present how we build the data set, its properties, and how we classify the data. Then we characterize both the growth of the public awareness of AI and the risk perception of AI. We explore the perception of the public of major harmful events related to AI. After that, we explore the formation of the messages that were amplified, the role of the communicators, and we conclude with a discussion at the end.

2 Data set and methods

The English-speaking public awareness is an index of the public tweets containing the phrase "Artificial Intelligence" from January 2007 until January 2018. Since Twitter's API has a limitation in the number of 32.000 retrievable tweets from the past, we adopted as a scrapping strategy the simulation of a browser. We did that using the selenium web driver library available for Python3. We did not add the acronym "AI" to our query since it would have led us to get a large number of ambiguous results. Our query also took care of fulfilling the whitespace between the words to capture hashtag occurrences. As a result, we gathered 3.682.015 public tweets posted by 785.297 unique profiles. Even though the query was in English, it did not follow that the profiles were all from English-speaking countries. We inferred the profiles' geographical localization from their time zone information in which around 70% of the profiles contained that. From this group, we estimated that 66% of them were from an English-speaking country. Within the English-speaking world, 70% of the profiles are located in the US, 14% in the UK, 9,3% in Canada, 4% in Australia, 1,5% in South Africa, and 1,1% in Ireland. It is also worth emphasizing that the bulk of public awareness comes from US-based profiles. We hope to extend in further analysis of how public awareness and the risk perception of AI changed in different countries.

Another important discovery that we hope to explore in the further analysis is related to the content of tweets. They are not only personal opinions on subject matters related to AI, but they also refer to external pages, esp. newspaper articles, books, scientific articles, and YouTube videos. In fact, 79.9% of the posts related to an external link. In this case, Twitter served not only as a public opinion repository but mainly as a repository of shared pages on the topic that are temporally indexed. Having this sort of data is very important for future measures of the influence and the half-life of media coverage and experts' communication about the risk of AI. It is likely that the number of posts on the topic is much higher as the acronym "AI" was not explored. It is also perfectly reasonable to argue that as soon the term "Artificial Intelligence" became familiar to the public, its acronym started to be widely adopted. However, we did not test that either in this paper.

Besides, we did not explore the phrase 'Artificial General Intelligence' [AGI], which specifies the research, development, or deployment of a human-level AI. That is an expert phrase conceived to clarify part of the meaning confusion we presented above. As we will see, the risks related to an AI and the risk of an AGI are somewhat different in nature. One can say that the latter implies an existential risk, and the former does not. In a nutshell, existential risks are usually defined as threats to the future of humanity, i.e., events that can lead to human extinction.

A critical way in which people interpret risk is related to the affect heuristics (Slovic et al. 2006). It follows the idea of problem-solving and information-processing models based upon bounded rationality (Simon 1956), and judgment heuristics such as availability, representativeness, and anchoring (Tversky and Kahneman 1974; Kahneman et al. 1982). The concept was introduced by Zajonc (1980) and percolated to risk perception as soon as the concept was developed (Slovic 1986). Affect heuristics claims that "representations of objects and events in people's minds are tagged to varying degrees with affect" where "people consult or refer to an 'affect pool' containing all the positive and negative tags consciously or unconsciously associated with the representations" (Slovic et al. 2006, p.1335). Empirically, word association methods have been applied to studies testing affect heuristics (see Benthin et al. 1995). In this line of thought, we classified the risk tweets from other tweets by employing a word association method. We took a set of 99 words that could represent the risk affect pool. From this procedure, we ended up with two sets of tweets related to AI, one related to risk and therefore a negative association, whereas the second is a set of reactions to AI that can range from rejection to enthusiasm. Besides, we tested additional filters to understand better other association tendencies like "machine learning," "benefits from AI," and "experts."

Regardless of the straightforwardness of the method, it worked as a constant that allowed to keep track of the changes. However, we did not attribute any weight for a variety of words. For the SARF, the unit of analysis is the tweets tied to a unique profile, it does not matter if there is repetition across the different tweets from different profiles, but it does matter whether the same profile has repeated in various tweets the same content. In this case, they were counted as duplicates and therefore deleted. A final consideration is that a large amount of data from Twitter securely allows us to observe trends and information flow. On the other hand, we are aware of the disadvantage of having a barrier regarding individuals' subjective judgments.

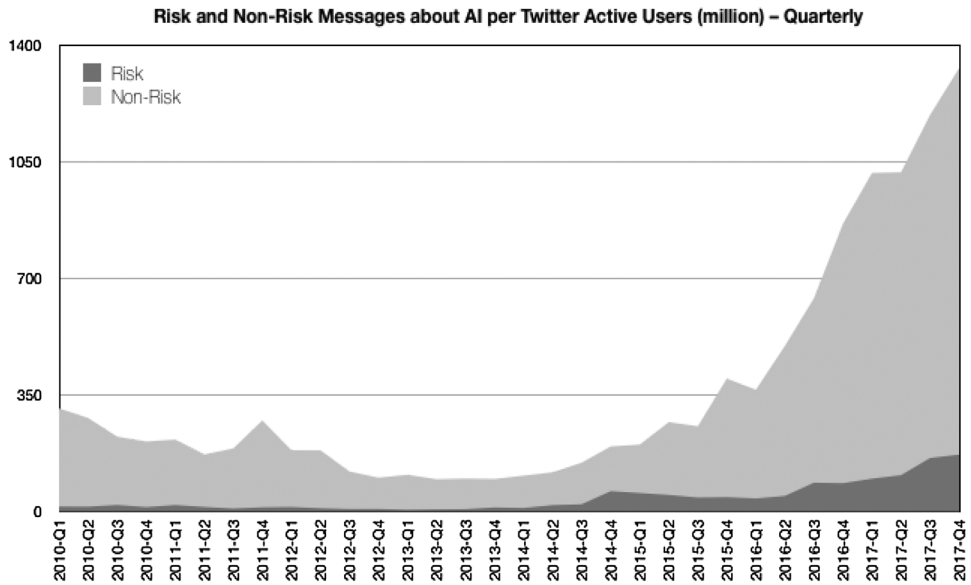

Fig. 1 The growth of risk and non-risk messages associated with artificial intelligence per Twitter-active users quarterly measured—non-stacked graph

2.1 The growth of public awareness and risk perception of AI

We start to analyze the growth of the public awareness of AI and its following public perception of risk by plotting the number of unique profile-tweets quarterly standardized by the total amount of Twitter active users. As this information is available from 2010 (Statista 2018), we considered that in our analysis. The standardization is relevant since the growing of the users of any social network service increases the likelihood that any specific topic is more talked about. The result is depicted in Fig. 1 below.

The first thing that calls attention in Fig. 1 is the decreasing popularity of the usage of the phrase "artificial intelligence" up to the end of 2013. It shows a turning point in the first quarter of 2014 when it grows progressively quicker. Although it has a false start in the fourth quarter of 2015, it booms after the first quarter of 2016. The trend we see in Fig. 1 after the first quarter of 2016 is a process of popularization of the use of the phrase "artificial intelligence." For the analytical purpose of this paper, popularization is a token of public awareness. It is also true that the risk perception grows with the popularization of the use of the phrase AI without any association with risk. If we join both sets, we see that the average proportion of the tweets about risk perception was of 9.6% for the entire period. Before the last quarter of 2013, this relation was always below the average. However, from the first quarter of 2014 on the proportion has increased, and it reached a significant peak of 24% in the last quarter of 2014. It means that during that quarter for every four tweets posted publicly on the web, one associated the idea of risk to AI.

If we compare the rate of growth of both the risk perception set and the public awareness set without risk (Fig. 2), we can see that their rate of growth is almost the same until the third quarter of 2013. The difference between the rates starts after the second quarter of 2014, which coincides with the beginning of the turning point of the popularization of Artificial Intelligence. Still, within the turning point year of 2014, we can see the gulf between both rates of growth. Another characteristic that we should stress out is the abrupt variations of the risk perception rate compared to the steady pace of growth of the general perception. It suggests two things; the first is the process of popularization as a continuous growth until it becomes part of the daily vocabulary and public awareness, whereas the risk perception has abrupt changes. As the risk perception set is a smaller one, it is vulnerable to changes even though changes started after only 2014. It may be the case that this suddenly changes are associated with news related that relates Artificial Intelligence and risk. According to SARF, we would expect to see here news or people talking about events that entailed undesired consequences involving AI. Moreover, considering the increasing rate of growth of the risk perception of AI, we would be able to see a growing number of events as well. And that is not what happened.

2.2 Actual harmful events and communication

The first death arguably related to an AI system happened with the accident of the self-driving car Tesla Model S in May 2016, "in the middle of a sunny afternoon, on a divided highway … an early adopter … died. The car failed to see a white truck that was crossing his path" (Stilgoe 2018, p. 2). At the same day, the public profile AI Briefing posted "Tesla driver dies in first fatal crash while using autopilot mode http://bit.ly/2987r43 #AI #robotics | artificial intelligence" (2016-05-30 15:50) and a few hours

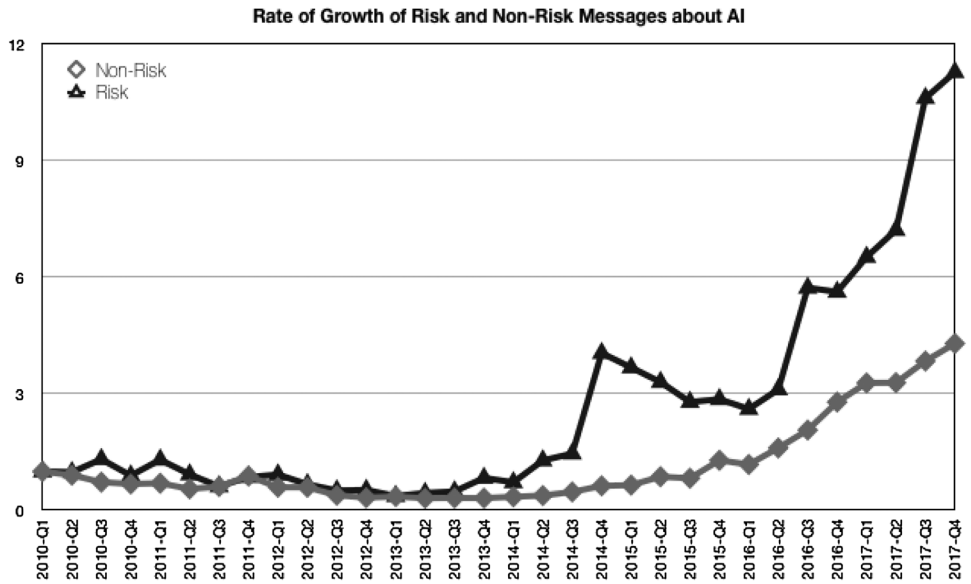

Fig. 2 Rate of growth of risk messages associated with AI compared to the riskless association to AI per Twitter active users quarterly measured

later "Tesla drivers post viral, self-driving 'stunts' using autopilot technology http://bit.ly/296zG4P #AI #robotics | artificial intelligence" (2016-05-30 19:36). The number of messages talking about the crash that related it to AI was around 60 tweets throughout that year. At the end of the year, on the 28th December, a video was released showing the benefits of Tesla's AI autopilot which manages to avoid a collision, 1 day after the video was republished with an article about the technical achievement made by Tesla. It also had no repercussion as well, less than a dozen messages.

A second example which is nowadays more popular because Bostrom's book Superintelligence (Bostrom 2014) depicts two cases of severe financial damage caused by a high-frequency trading algorithm that may be called today an AI, the Flash Crash (CFTC and SEC, 2010) in May 2010, and the bankruptcy of the Knight Capital Group in August 2012. In the first case, the E-Mini S&P 500 futures was halted by an automatic circuit breaker when a trillion dollars had been wiped off the market. After the market closure, the decision made by regulators was to cancel all trades that had been executed at prices 60% or more away from their pre-crisis levels (CFTC and SEC 2010). In Knight Capital's case, its high-frequency trading algorithm caused a significant stock market disruption in the prices of 148 companies listed at the New York Stock Exchange. This error cost $440 million to Knight Capital and was described as a "technology breakdown" (Farrell 2012). In both cases, bugs and some managerial incompetence caused huge monetary side effects. However, neither of them was broadly associated with AI at that time, and consequently, they did not integrate the public perception of risk of AI. In other words, there was no significant number of tweets that talked about those cases and related them to AI.

We should state again that there is no reason to think that in the future these accidents may be strongly associated with AI in the public perception. Both kinds of cases have material factors such as property damage (millions of dollars) and deaths (one death registered) that would be enough to trigger ripple effects in a sense that the risk perception of AI could be amplified. However, it has not happened. Thus, we claim that these actual events were not responsible for triggering the messages of risks that were amplified and generated secondary consequences related to AI as expected by SARF. It is also worthy to state that there was not any other event that caused actual harm depicted in the tweets database associated with AI. During the period, this paper concerns, according to our data, the only association was the self-driving car fatality that was not robustly associated with the public perception of AI. It may be due to the non-extensive coverage of media or a lack of public response of the cases, so they have either attenuated right away or not propagated at all. It shows the critical role of the individuals as stations receiving, decoding, and communicating the messages of risk. As we saw, there was a message that began to be formed related to the self-driving car crash fatality and AI, but it was not amplified.

Within SARF, those individuals who are specialized to communicate a risk message are called risk communicators. Roughly speaking, any station or any individual who attenuates or amplifies a message of risk is a risk communicator. It happens on a local scale. However, some public people or institutions can communicate with broader audiences. In this case, the starting message of risk relating AI to the fatal crash was attenuated and eventually ignored because it was associated to something more prosaic like an autopilot, which would be merely considered a smart device familiar to many people hence anything new. Someone like Elon Musk,

the founder and head of Tesla at that moment, which also funds or is a member of boards of the institutes and centers for the study of the risk of AI, is himself a risk communicator. Besides, for obvious reasons, he managed to either attenuated or ignored the message. However, what sort of messages did he amplify?

2.3 Messages formation

At the end of 2014, Stephen Hawking's interview for the BBC about the potential threat of developing human-level AI became viral. His statement was clear, "the development of full artificial intelligence could spell the end of the human race." (Cellan-Jones 2014). At the fourth quarter of 2014, his interview represented 14.6% of all messages related to AI and 46.5% of all risk perception messages about AI. What is remarkable about this topic is that in 2014, the movie about his life "The Theory of Everything" was released at the end of the year. Of course, the massive repercussion of his interview may have happened because an availability bias (Kahneman et al. 1982), since in Hawking's biographical movie he was portrayed both as the highest genius after Einstein and a resilient person, a living example for everyone. Hawking went through a process of popularization. The availability bias had indeed played a role in the robust amplification process of Hawking's alert because he had expressed his fears about AI in May 2014 (Kolodny 2014), when the repercussion was not so high. It represented 29.5% of all risk associated messages of AI and 4.9% of all messages regarding AI. In this case, Hollywood indirectly affected the prestige related to a specific person who turned out to be a risk communicator, a highly trusted source of information.

The message Hawking transmitted can be categorized as an existential risk message.[1] Existential risk draws from the classical formulation of risk, i.e., the combination of probability of the events and the magnitude of their consequences (Freudenburg 1988), the following extreme scenario, even if there is a slight chance for such a doom event happen, it is worthy of caring about it (Bostrom and Ćirković 2008). Existential risk researchers are not only producing scholarly studies on the topic for scholars, but they are also talking to the broader public. Also, they are not strictly researching within scientific departments but within institutes and centers-like institutions that maintain strong ties to universities and are supported financially by entrepreneurs concerned with the future of humanity. We mentioned before, some of these institutes located in the US and in the UK whose roles are educational.

Along with the scholarly work on this kind risk of AI, fiction plays an essential role in the risk perception of AI too. "2001: A Space Odyssey" by Arthur C. Clarke and Stanley Kubrick is a classic example. But fiction is also driven by experts. Back at the 1960's Clarke and Kubrick hired as advisors for the movie the mathematician I. J. Good, who came up with this idea of intelligence explosion and superintelligence (Good 1965), and Marvin Minsky one of the fathers of AI and optimist that soon we would achieve human-level intelligence. Fiction writers have also influenced the generation of transhumanists, who are scholars, entrepreneurs, and enthusiasts arguing for a technological singularity, an idea coined by Vernor Vinge (1993). If "the public is influenced by emotion and affect in a way that is both simple and sophisticated. So are scientists. Worldviews, ideologies, and values influence the public. So are scientists, particularly when they are working at the limits of their expertise" (Slovic et al. 2000). It is rare to be even in a scholar discussion on cultural, economic, and social impacts of AI and not stumble across a movie reference like Terminator, Matrix, I-Robot, Transcendence, Ex Machina, Her, Blade Runner 2049. Here we can see the combination of two critical factors that operate under the same logic. In both cases, risk events are based on counterfactual scenarios that are a sort of future risk factor (Kasperson 1992).

Within our dataset, 88% of risk perception of AI was associated to existential risks, in the sense that AI would mean the end of the human race or AI would become a malevolent superintelligence, the same message transmitted by Stephen Hawking. This framing of the risk of AI is not new, which is the hypothesis of the super intelligence and intelligence explosion (Good 1965). This argument is present in the expressions of every expert or risk communicator who conceives the side effect caused by the development of a human-like AI. At the moment Hawking communicated his message on the public venue, other experts have also been transmitting such messages. To test the influence of other experts who acted as communicators, we created a new subset of tweets which contained both AI and a list of names of experts who appeared on the media in the last 10 years. We compared both rate of growth, i.e., the risk perception (as in Fig. 2) and the experts, and we depict the outcome below in Fig. 3. There we can see again the critical moment of the last quarter of 2014 as a moment when experts formulated the message of the risk perception and transmitted in the news. Hawking's example shows how important the risk communicators and experts are for the risk perception of AI, at least for its amplification. According to the SARF, if the main message regarding the risk perception of AI was formed and amplified by the experts, we can see that such message was able to keep its transmission throughout the quarters as a wave. The new expositions some of these

[1] As we saw in the introduction, existential risks are events that threaten "to cause the extinction of Earth-originating intelligent life" (Bostrom 2002, p. 381).

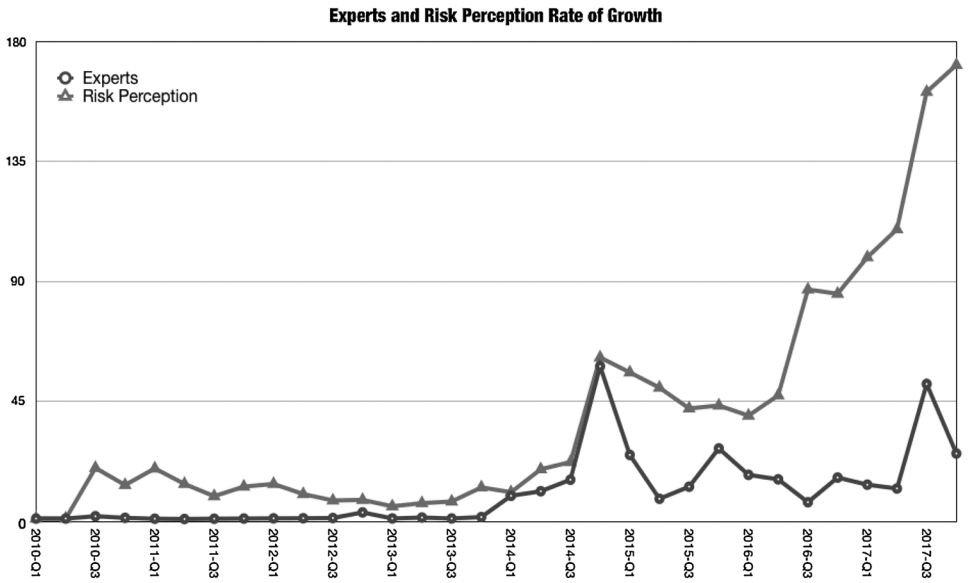

Fig. 3 The rate of growth of the experts and the risk perception of AI

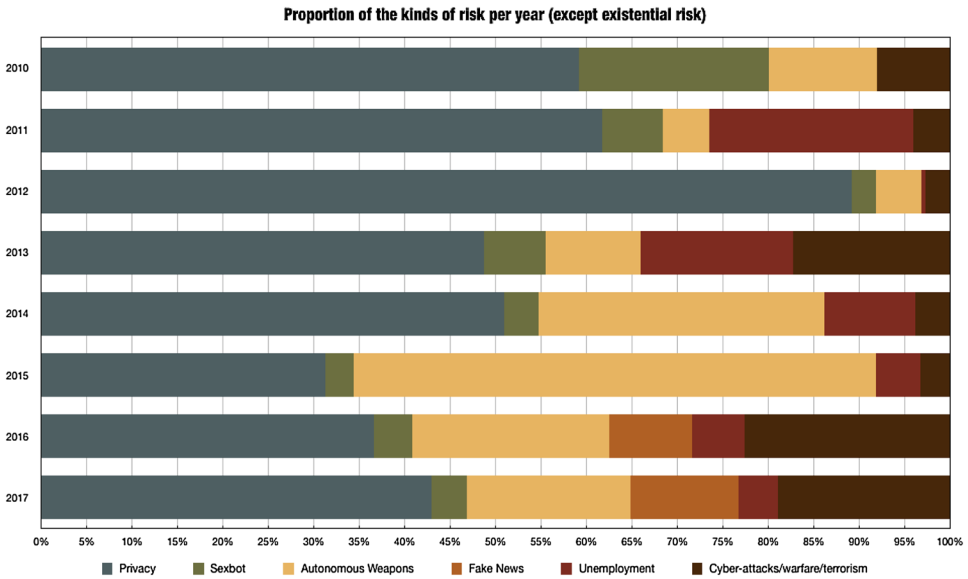

Fig. 4 Proportion of the kinds of risk of AI per year (except existential risk)

experts made on the public may have helped to keep pace; however, it seems to be a relatively independent movement.[2]

As a matter comparison with the risk perception of other technologies, Li et al. (2016) analyses the public perception of nuclear energy taking as the reference event the Fukushima Daichii nuclear accident following Tohoku earthquake and tsunami in 2011. One of their findings has some similarities with what we found in Fig. 3. Firstly, the volume of nuclear-related tweets varied during the time framework of this study (from December 2010 to May 2012), i.e., a short period before the accident and 1 year after that. Despite the volume of tweets dramatically increased within the week following the accident, it rapidly declined thereafter. Nonetheless, "the volume of nuclear-related tweets still remained at a relatively high level compared to that before Japan's earthquake, which indicated an escalation of public concerns over nuclear safety caused by this event" (Li et al. 2016, p.14). The content of the tweet had also changed. Before the accident most of the tweets "merely expressed straightforward information rather than interpretations of the causes or implications of the disaster, whereas the later tweets became more interpretive [Binder 2012]" (Li et al. 2016, p. 4). In this context, Twitter can function as a useful tool to assess genuine and spontaneous opinion generated by the Fukushima accident (Li et al. 2016, p. 14).

[2] A further study to show how the life cycle of this critical moment of communication of this risk perception propagated would need us to do a network analysis reconstructing the spread.

Back to AI, even though the existential risk was the leading risk related to AI, there others less speculative risks that were also had experts as their risk communicators (Fig. 4). Since 2010, the most frequent risks except the existential risk are Privacy and Surveillance (41%), Lethal Autonomous Weapons (21%), Cyber-attacks/warfare/terrorism (14%), Unemployment [3](13%), Fake News (8%) and Sexbots[4] (4%). This proportion changes yearly, and these kinds of pragmatic risks are becoming more diverse. Although privacy and surveillance are still the most associated risk of AI perceived within the dataset, their importance has been attenuated. New risks started to be introduced like fake news example. Of course, as the deployment of AI does not offer that level of risk that the development of the hypothetical strong AI, these kinds of risk perception suffer from the hot debates available at the point. For example, fake news is a phenomenon introduced in 2016, mainly due to Donald Trump's election as President of the United States. And Lethal Autonomous Weapons were a significant topic in 2015, when Stuart Russell, one of the most critical experts in the field of AI, joined to the public advocacy for the ban on development of such weapons (Russell 2015a, b).

2.4 Communicators, experts, intellectuals

The trust in the risk communicators is high because some of them drew their prestige from their scientific career within his or her field of research. Whenever they come to the public, they act as public intellectuals. According to Baert and Morgan (2017), intellectuals is the denomination of the group of scientists (whether computer scientists, physicists, mathematicians, or sociologists) and philosophers who have both the expertise and authority to speak publicly about matters categorized as scientific or philosophical. They do so "while possibly drawing on their expertise in a specific area, address a broader public and engage with what are considered to be significant social and political issues of the day which go well beyond their narrow field of professional focus" (Baert and Morgan, 2017, p. 2). In this sense, they can be considered as experts according to the risk perception terminology we have employed here.

Intellectuals use the positioning, which, through the attribution of some characteristics to other intellectuals, positions them in a value spectrum, usually but not only political. Russell has been positioning himself publicly to warn of possible undesirable effects of AI since 2013 and therefore accusing the irresponsibility of experimenting with AI without any safeguard. Conversely, skeptical computer scientists that address the subject to the public frequently, like Oren Etzioni, Edwan Felten, Jaron Lanier, and Roger Schank, position those experts who are highly concerned with the mainly existential risk that AI can bring as a "not to take too seriously" group. Critics of this position, such as computer scientist Jaron Lanier, consider this to be a religious narrative that is being built upon AI; would be a "Frankenstein myth," in which the creature turns against the creator. According to him, other risks of AI are more plausible, such as the fact that AI is a farce, or that the result of the search and recommendation systems could lead to "mass incompetence" or "generalizations and senseless answers and suggestions." That is, the behavioral risk of the human being to guide their decisions exclusively by algorithms, such as a car route (which may pass dangerous places), the recommendation of movies, books, news and even sexual partners (Lanier 2014).

The distinction between the public intellectual and the non-public intellectual is sometimes tenuous. However, based on Russell's performances communicating the dangers of AI and its applications, such as Lethal Autonomous Weapons in world forums to discuss socio-political issues, are a current example of this type of public intellectual today. Moreover, intellectuals increasingly become aware of their performances. This statement can be verified based on the increasing frequency of these people's appearance in interviews, lectures, and articles for the general public, as well as references to these intellectuals and their positioning on issues (e.g., Wolcholver 2015).

Besides the actions, the intellectuals are the legitimate individuals to lead the way to the formation of counterfactual scenarios where the feared event and its consequences may happen. Even though the people who conceive the scenarios do not believe that one of them will happen with him or with someone close to him in the near future, the imagination may exert practical effects on someone's decision and planning. As Salvadori et al. (2004) conclude about their risk study, the choices and the judgment of risk became a matter of trust "the less we know about an activity" (p. 1290). A plausible hypothesis about the dread variable is that it comes along with the future risk in our case, since the imagined future risk scenario may take the shape of the most fearful thing they relate to AI, for example losing the job for an AI program.

[3] For the discussion see, Mokyr (2014), Brynjolfsson and McAfee (2014), Frey and Osborne (2013), and Glaeser (2014).

[4] See Levy (2007). Sex bots are taken as risk for some group of people for different reasons. On the one hand, there are feminists' groups that argues that sex bots will just enhance the gender discrimination and stereotypes. On the other hand, there are conservatives' religious groups that are by definition against the hedonism.

3 Conclusion

From our analysis, some experts playing the role of public intellectuals started up the recent idea that AI could be a real threat and endanger all humans. In this sense, they framed and communicated the message that work as a critical event that impacted the public perception. This message of risk was based on counterfactual scenarios instead of actual events, such as any particular self-driving car crash. The counterfactual scenarios were at the basis of the messages of existential risks related to AI that were transmitted and amplified. As we found out, 88% of all risk related tweets were related to existential risks. In this regard, the process of propagation of the messages related to the varieties of risks of AI is associated with the primary senders.

Risk perception's studies and the SARF can be enriched by further analyzing the role of experts in the formation, amplification, and attenuation of risk perception. In the way the framework was designed, it ignores the possibility of trigger events arising from sheer human conjectures. By doing that, it ignores the active role of authoritative individuals, such as the experts, in social interplay of the public position, and what such positioning can bring to the expert. It is to be expected that for any technological development, experts will not have a consensual public position about the risk of such technology. Experts display three different positions related to technology, they can be antagonists, pragmatists or neutrals, and enthusiastic experts. In the case of AI, antagonists believe there are insurmountable barriers to achieve full-fledged, human-level AI; so any risk scenario related to that is nonsensical. Pragmatists or neutrals believe that it is hard to even depict what are the real challenges to develop a full-fledged, human-level AI; even though we may achieve that at some point. For this group, the real dangers are in the short-term related to the portion of the technology that already works in the world, such as the effect of biased data sets for machine learning algorithms and unemployment (Frey and Osborne 2013). Finally, the enthusiastic experts believe that the full development is just a matter of time, and such development will bring a profound change. However, changes can be positive or negative. Because of that, enthusiastic experts can be grouped into pessimists and optimists. Existential risk scenarios are framed by the pessimists.

According to Baert and Morgarn (2017), there would be a positioning dispute between those experts whenever they come to public. And the best way to frame this dispute is with SARF. Because of the speculative nature of the harm caused by AI, pessimist experts are the kinds of experts (or intellectuals) who conceive such counterfactual dreadful and future risk scenarios. By doing that, they formulate the message that is going to be conveyed. Other pessimist experts may also play the role of amplification stations for this message. Now when pragmatic experts are forced to public positioning themselves, they play a clarification role, which can be identified with the role of attenuation stations. While rejecting extremely speculative scenarios, some of them may end up wanting to stress out the "real" dangers of the technology. When they do so, they create a new message, and a new process begins.[5] But if those pessimist experts who happen to be risk communicators as well are capable of amplifying their messages based purely in the conception of such counterfactual scenarios, in a way it triggers many indirect effects within society.

Funding This research was funded by the São Paulo Foundation (Fapesp) grants 2018/09681-4 and 2019/07665-4 and Brazilian National Council for Scientific and Technological Development (CNPq) grant 312180/2018-7.

References

Baert P, Morgan M (2017) A performative framework for the study of intellectuals. Eur J Soc Theory. https://doi.org/10.1177/1368431017690737

Beck U (1986) Risikogesellschaft: auf dem Weg in eine andere Moderne. Suhrkamp, Frankfurt am Main

Beck U (1999) World risk society. Polity, Malden

Beck U (2007) Weltrisikogesellschaft: Auf der Suche nach der verlorenen Sicher- heit. Suhrkamp, Frankfurt am Main

Benthin A, Slovic P, Moran P, Severson H, Mertz CK, Gerrard M (1995) Adolescent health-threatening and health-enhancing behaviors: a study of word association and imagery. J Adolesc Health 17:143–152

Binder A (2012) Figuring out #Fukushima: an initial look at functions and content of us twitter commentary about nuclear risk. Environ Commun J Nature Culture 6(2):268–277

Bostrom N (2002) Existential risks: analyzing human extinction scenarios and related hazards. J Evol Technol 9

Bostrom N, Ćirković MM (2008) Global catastrophic risks. Oxford University Press

[5] The propensity to take corresponding actions may lead to behavioral patterns, which generate secondary social or economic consequences that extend far beyond direct harm to humans or the environment, including significant indirect impacts such as liability, insurance costs or loss of trust in institutions (Kasperson et al. 1988). The consequences can also be good or neutral, with the foundation of new companies and institutes, as in the case of AI. Such secondary effects often trigger demands for additional institutional responses and protective actions (like the Whitehouse Policy initiative), or conversely (in the case of risk attenuation), place impediments in the path of needed protective actions. An interesting contribution for further studies would be the analysis of how media institutions (profiles and websites) frame the risk perception of AI, how and why layman amplifies specifics risk. In this sense, the research on risk perception may benefit from further developments on the role that experts play not only in the assessment of risk, but also their communication to the public, and sometimes how they shape or formulate the risk.

Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press, Oxford
Brynjolfsson E, McAfee A (2014) The second machine age. W.W.Norton & Company, New York
Cellan-Jones R (2014) Stephen Hawking wars Artificial Intelligence could end mankind. BBC Technology. https://www.bbc.co.uk/news/technology-30290540. Accessed 2 Dec 2014
CFTC and SEC (Commodity Futures Trading Commission and Securities and Exchange Commission) (2010). Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues. Washington, DC
Chong M, Choy M (2018) The social amplification of haze-related risks on the internet. Health Commun 33(1):14–21. https://doi.org/10.1080/10410236.2016.1242031
Chung IJ (2011) Social amplification of risk in the internet environment. Risk Anal 31(12):1883–1896
Douglas M (1985) Risk acceptability according to the social sciences. Routledge & Paul Kegan, London
Douglas M (1986) How Institutions Think. Syracuse University Press, Syracuse, NY
Douglas M (1990) Risk as a forensic resource. Daedalus 119(4):1–16
Douglas M (1992) Risk and blame: essays in cultural theory. Routledge, London; New York
Farrell M (2012) Knight's bizarre trades rattle markets. CNN. http://buzz.money.cnn.com/2012/08/01/trading-glitch/. Accessed 1 Aug 2012
Fellenor J, Barnett J, Potter C, Urquhart J, Mumford JD, Quine CP (2018) The social amplification of risk on Twitter: the case of ash dieback disease in the United Kingdom. J Risk Res 21(10):1163–1183. https://doi.org/10.1080/13669877.2017.1281339
Freudenburg WR (1988) Perceived risk, real risk: social science and the art of probabilistic risk assessment. Science 242:44–49
Frey C, Osborne M (2013) The future of employment: how susceptible are jobs to computerisation? Technical Report, Oxford Martin School, University of Oxford, Oxford, UK
Foucault M (1978) Governmentality. Ideol Conscious 6:5–12
Foucault M (1980) Power/knowledge: collected interviews and other essays 1971–1977. Harvester Press, Brighton
Foucault M (1982) The subject and power. Crit Inq 8:777–795
Foucault M (1991) Governmentality. In: Burchell G, Gordon C, Miller P (eds) The foucault effect: studies in governmentality. Harvester Wheatsheaf, London, pp 87–104
Glaeser E (2014) Secular joblessness. In: Teulings C, Baldwin R (eds) Secular stagnation: facts, causes, and cures. Centre for Economic Policy Research (CEPR), London, pp 69–82
Good IJ (1965) Speculations concerning the first ultraintelligent machine. Adv Comput 6:31–88. https://doi.org/10.1016/S0065-2458(08)60418-0
HC—House of Commons (2016) Robotics and artificial intelligence. https://publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/14502.htm. Accessed 8 June 2018
HL—House of Lords (2018). AI in the UK: ready, willing and able? https://www.parliament.uk/business/committees/committees-a-z/lords-select/ai-committee/news-parliament-2017/ai-report-published/. Accessed 4 May 2018
Kahneman D, Slovic P, Tversky A (1982) Judgment under uncertainty: heuristics and biases. Cambridge University Press, New York
Kasperson R (1992) The social amplification of risk: progress in developing an integrative framework of risk. In: Krimsky S, Golding D (eds) Social theories of risk. Praeger, Westport, p 153
Kasperson R, Renn O, Slovic P, Brown HS, Emel J, Goble R, Kasperson J, Ratick S (1988) The social amplification of risk: a conceptual framework. Risk Anal 8(2):177–187
Kasperson J, Kasperson R, Pidgeon N, Slovic P (2003) The social amplification of risk: assessing fifteen years of research and theory. In: Pidgeon N, Kasperson R, Slovic P (eds) The social amplification of risk. Cambridge University Press, New York, pp 13–46
Kolodny C (2014) Stephen Hawking is terrified of artificial intelligence. Huffington Post. https://www.huffingtonpost.co.uk/entry/stephen-hawking-artificial-intelligence_n_5267481. Accessed 5 May 2014
Lanier J (2014) The myth of AI: a conversation with Jaron Lanier. Edge.org. https://www.edge.org/conversation/the-myth-of-ai#26015. Accessed 14 Nov 2014
Levy D (2007) Love and sex with robots: the evolution of human-robot relationships. Harper/HarperCollins Publishers, New York
Li N et al (2016) Tweeting disaster: an analysis of online discourse about nuclear power in the wake of the Fukushima Daiichi nuclear accident. J Sci Commun 15(05):A02
Luhmann N (1993) Risk: a sociological theory. A. de Gruyter, New York
Lyng S (1990) Edgework: a social psychological analysis of voluntary risk taking. Am J Sociol 95:851–886
Mokyr J (2014) Secular stagnation? Not in your life. In: Teulings C, Baldwin R (eds) Secular stagnation: facts, causes and cures. Centre for Economic Policy Research (CEPR), London, p 83
NSTC—Executive Office of the President National Science and Technology Council (2016) National Science and Technology Council Committee on Technology. Preparing for the future of artificial intelligence. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf
Pidgeon N, Kasperson R, Slovic P (2003) The social amplification of risk. Cambridge University Press, New York
Renn O (1991) Risk communication and the social amplification of risk communicating risks to the public. Springer, Berlin/New York, pp 287–324
Renn O, Burns W, Kasperson J, Kasperson R, Slovic P (1992) The social amplification of risk: theoretical foundations and empirical applications. J Soc Issues 48(4):137–160
Research Center and the Ohio Aerospace Institute, 30–31 March 1993. http://www.rohan.sdsu.edu/faculiy/vmge/misc/singularity.html
Russell S (2015a) Ban Lethal Autonomous Weapons. The Boston Globe 08 09
Russell S (2015b) Take a stand on AI weapons. Nature 521(7553):415–416
Salvadori L, Savio S, Nicotra E, Rumiati R, Finucane M, Slovic P (2004) Expert and public perception of risk from biotechnology. Risk Anal 24:1289–1299
Simon HA (1956) Rational choice and the structure of the environment. Psychol Rev 63:129–138
Sloan L, Morgan J, Housley W, Williams M, Edwards A, Burnap P, Rana O (2013) Knowing the Tweeters: deriving sociologically relevant demographics from Twitter. Sociol Res Online 18:7. https://doi.org/10.5153/sro.3001
Slovic P (1986) Informing and educating the public about risk. Risk Anal 6(4):403–415
Slovic P (2000) The perception of risk. Earthscan, London
Slovic P, Kunreuther H, White G (2000) Decision process, rationality and adjustment to natural hazards. In: Slovic P (ed) The perception of risk. Earthscan, New York
Slovic P, Finucane M, Peters E, MacGregor D (2004) Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. Risk Anal 24(2):2004
Slovic P, Finucane M, Peters E, MacGregor D (2006) The affect heuristics. Eur J Oper Res 177(2007):1333–1352
Statista (2018) Number of monthly active Twitter users". https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Stefanik E (2018) 115th Congress 2nd Session. H.R. 5356 to establish the National Security Commission on Artificial Intelligence. https://www.congress.gov/bill/115th-congress/house-bill/5356

Stilgoe J (2008) Machine learning, social learning and the governance of self-driving cars. Soc Stud Sci 48(1):25–56. https://doi.org/10.1177/0306312717741687

Strekalova YA, Krieger JL (2017) Beyond words: Amplification of cancer risk communication on social media. Journal of Health Communication 22(10):849–857. https://doi.org/10.1080/10810730.2017.1367336

Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Science 185:1124–1131

Vinge V (1993) The coming technological singularity: how to survive in the post-human era. Presented at the VISION-21 Symposium sponsored by NASA Lewis

Witrz C et al (2018) Rethinking social amplification of risk: social media and Zika in three languages. Risk Anal 38(12)

Wolcholver N (2015) Concerns of an Artificial Intelligence Pioneer. Quanta Magazine. Publicado em 21 de Abril de 2015. https://www.quantamagazine.org/artificial-intelligence-aligned-with-human-values-qa-with-stuart-russell-20150421/

Nilsson N (2010) The quest for artificial intelligence: a history of ideas and achievements. s.l.:web version

Zajonc RB (1980) Feeling and thinking: preferences need no inferences. Am Psychol 35:151–175

Zhang L, Xu L, Zhang W (2017) Social media as amplification station: factors that influence the speed of on-line public response to health emergencies. Asian J Commun 27(3):322–338. https://doi.org/10.1080/01292986.2017.1290124

Zhou M, Wang M, Zhang J (2017) How are risks gener- ated, developed and amplified? Case study of the crowd col- lapse at Shanghai Bund on 31 December 2014. Int J Disaster Risk Reduct 24(Supplement C):209–215. https://doi.org/10.1016/j.ijdrr.2017.06.013