Truth in the 'Killer Robots' Angle?

Matthew Rahtz (University of Zürich/ETH Zürich; mrahtz@ethz.ch)
DOI: 10.1145/3137574.3137584

I had no idea, getting interested in AI two years ago, that being involved in the field would involve such a persistent sense of unease. I started out unequivocally excited, perhaps a little naive; but over the years, the concerned voices of economists, philosophers, and the mass media have gradually seeped into me, leaving me with an ill-defined feeling of hesitation about what we're heading towards.

Trying to understand the situation a bit better over the past months has been somewhat overwhelming. AI touches so many different areas that it's hard to hold everything in mind at once. While there are points of concern in many of these areas, there are three areas in particular which have stood out to me: unemployment, the long-term risks of AI, and lethal autonomous weapons systems (otherwise known as 'killer robots').

Reading into these areas in more depth has left me surprised. AI risk research is something that's interested me for a while, but the pressing issue right now doesn't seem to be the immediate need for research - rather, the connotations that are becoming associated with that research. With the second of these areas, autonomous weapons, I began looking into it only for the sake of completeness, but found myself completely unaware of the gravity of the situation. The biggest surprise, however, has been a change of opinion about the danger of unemployment. I realise I am no longer nearly so concerned about the imminent threat of automation.

Given all that we've been hearing on the topic of AI-related unemployment recently, this last statement clearly requires some explanation. So that the scene is set properly for the former issues, let us in fact begin our discussion with this last point.

Unemployment

I was quite prepared to spend this essay arguing that unemployment resulting from use of AI was by far the most pressing issue right now. The more I've read, though, the more I've got the impression that the change is going to be slower than it might appear.

At the start of my reading, the threat of unemployment seemed clear. Automation in various forms has been encroaching on the job market for centuries. With the recent step change in our machine learning capabilities, it seemed a very reasonable concern that we may not be far away from a step change in automation, and hence unemployment, too.

Take autonomous vehicle technology. Companies like Google and Tesla seem to be getting pretty good at it. For most of us, self-driving cars are going to be an unambiguously good thing; promises of safer roads and more leisure time abound. The question is: what happens when that technology makes, say, self-driving trucks possible? There are 1.6 million long-haul truck drivers in the US (McArdle, 2015). It's the most popular job in 29 states (Solon, 2016). It's one of the last jobs left offering middle-class pay without a college degree (Kitroeff, 2016). When autonomous vehicle technology reaches the trucking industry and makes all these drivers redundant – what then?

Despite the apparently obvious problem, I was surprised to find that some people don't seem to be all that worried. The common arguments I've heard for this position, though, haven't convinced me.

"Jobs will be lost, yes; but the AI revolution will create new jobs at the same time," some say. But I see no guarantee that the jobs created will match the skills of the people being made redundant. This is already a problem: the global talent gap. The issue is not that there aren't enough jobs. The issue is that the unfilled jobs require skills that the unemployed population don't have.

Others say, "We've seen step changes in employment before; what about the industrial revolution? We survived that alright." But this doesn't comfort me, because the world is such a different place now. The internet, for example, enables advancements in technology to spread much quicker than ever before; and this is especially significant when the relevant technology is software. In general, it seems like a bad idea to try and make predictions based on only superficially-similar historic precedents.

The difficulty of making predictions in a highly unpredictable world is clearly one of the major factors limiting the quality of these discussions. Whatever ideas one may have, it's hard to avoid concluding with anything but the inevitable cop-out of, "But then again, technology changes so rapidly, who knows what may happen?"

This got me to thinking: what are the *technology-invariant* factors here? What dynamics will stay relevant regardless of what new kinds of technology we come up with? Of these factors, there's been one in particular that's struck me as significant: the Pareto principle. And there's no better example of this principle than in the development of self-driving cars.

My impression with autonomous vehicle technology is that we're a lot further away from complete human replacement than one might think at first glance. Sure, we've seen the exciting demos of self-driving technology. But while these demos are impressive, it's worth bearing in mind that even back in 2015, a single talented hacker could get similar demo-level functionality working in about a month (Vance, 2015). The difficulty is apparently not in getting something basically working; the difficulty is in getting it to work reliably, in a wide range of conditions. As Tesla point out in their response to said hacker's efforts: "This is the true problem of autonomy: getting a machine learning system to be 99% correct is relatively easy, but getting it to be 99.9999% correct, which is where it ultimately needs to be, is vastly more difficult." (Tesla, 2015)

This dynamic is, essentially, what the Pareto principle states: that the first 80% of the results tends to be achieved with the first 20% of the efforts (though the exact proportions don't matter). And it's this dynamic that makes me think that change is going to be much slower than we might expect.

Even Google's efforts seem to be in line with this principle. Their self-driving buggy without even a steering wheel is pretty cool, but there's a catch: it can only safely go 25 mph (Farivar, 2015). Also, it doesn't work in snow (McArdle, 2015). Turning back to self-driving trucks, if it's taken *Google* this long to get to achieve autonomy in even these limited conditions, I'd guess it's going to be a very long time before complete autonomy can be achieved for a multi-tonne truck travelling at high speed down a freeway in rain, sleet, fog and, indeed, often snow.

It's unsurprising, therefore, that current efforts at truck automation, such as those from Daimler (Davies, 2015) and recent startup Otto (Lee, 2016), are instead targeting semi-autonomous solutions. In good conditions, the truck will drive itself. In bad conditions, a human driver in the cab can take over. In good conditions, you still get the benefits of machine control, like the ability to drive throughout the night. But you're spared the difficulty of pushing all the way to the "99.9999%" that's required for complete automation.

I suspect this is a pattern we'll see throughout many industries. Sure, new technologies are going to pop up. And we'll see those technologies progress to the level of useful semi-automation pretty quickly. But it's going to take much longer for the technology to mature to the point where complete automation is possible.

This is not to say that complete automation won't happen eventually. But because of the Pareto principle, I think the change is going to be gradual. We're going to get advance warning of what's happening. It seems unlikely that it's going to be a step change.

I also don't mean to suggest that eventual wide-scale automation isn't something worth thinking about and preparing for. Indeed, I'm glad to see the issue receiving as much attention as it is. I only mean to say that it may be a better use of our energies right now to focus on other areas which are more pressing and, perhaps, more neglected.

This brings us to the first of what I believe

our current concerns really are: lethal autonomous weapons systems.

Lethal Autonomous Weapons Systems

One of the tropes brought up often in the media over the past few years has been the image of 'killer robots'. For a long time, the hyperbole that invariably accompanied such media lead me not to take the issue seriously. Even in retrospect, I'm not surprised at my ignorance. Those kinds of discussions never touched on what seems to be the *real* issue: the consequences of an autonomous weapons arms race.

First, let's clarify what we're talking about. Lethal autonomous weapons systems (LAWS), as they're known more dryly, refer to weapons which can make the decision to kill entirely on their own, without explicit go-ahead from a human. For example, the drones currently in use don't fall into this category, because the decision to kill must be made by a remote human operator. What we're talking about is, say, a drone that can find and kill a target while being completely disconnected from a pilot.

LAWS have already existed for a while – think land mines. More recently, though, advances in AI are starting to enable more sophisticated forms of LAWS. For example, automated sentry guns have already been developed and deployed along the border between North and South Korea (Rabiroff, 2010). This trend looks set to continue: with the advantages of LAWS (e.g. immunity to communications jamming; potentially more precise and accurate targeting; fewer soldiers' lives on the line), many nations are now investing heavily in their further development (Goose & Wareham, 2017). The question we now face is: should we allow this trend to continue?

There are arguments both ways. On the one hand, avoiding danger to soldiers' lives, LAWS lower the threshold of entry to conflict. And because of the difficulty in distinguishing combatants from non-combatants, LAWS could lead to an increase in the number of civilian casualties (Goose & Wareham, 2017). On the other hand, if we *can* program them with humanitarian law, perhaps LAWS could be more ethical than their stressed-out human counterparts (Arkin, 2015).

The most forceful argument I've come across, however, concerns the likely consequences of a LAWS arms race. The idea is: once one nation starts deploying sophisticated LAWS, other countries will feel the need to step up their own efforts to develop and deploy LAWS of their own, leading to a positive feedback loop (Future of Life Institute, 2015b). That race is going to lead to even more sophisticated forms of LAWS being developed, and at ever lower prices. With proliferation happening all around the world, at some point it seems inevitable that some units will fall (or be sold) into the wrong hands. Consider "the availability on the black market of mass quantities of low-cost, anti-personnel microrobots that can be deployed by one person to anonymously kill thousands or millions of people who meet the user's targeting criteria" (Russell, Tegmark, & Walsh, 2015), and you get the picture.

The proposal, therefore, is to ban LAWS before this arms race can get started.

Indeed, this is the direction that the gears of international government - the UN Convention on Certain Conventional Weapons (CCW) – *are* moving in. The problem is that they may not be moving fast enough. Only at the end of 2016, after three years of discussion, has the CCW agreed to establish a Group of Governmental Experts under whom the creation of new international law can be discussed (Wareham, 2017a). Whether this group will move quickly enough to prevent the start of the race is still uncertain (Wareham, 2017a). It's not even clear whether this discussion will really lead to a complete ban, or only regulation limiting LAWS' use (Goose & Wareham, 2017). Given that deployment of sophisticated LAWS may be only years away (Future of Life Institute, 2015b), we're at a decisive moment.

Reading about LAWS, I've been forced to admit that the 'killer robot' angle really does have a grain of truth in it. There is, however, a second angle of the scare I've gradually become convinced it's worth taking seriously: the long-term risks of AI. But it's not the risks themselves that I think are the most pressing issue right now. The bigger issue at the moment is the culture that's becoming associated

with such concerns – and the limiting effect that culture might have on the field's growth. This brings us to our second pressing issue: opinion of AI risk research.

Opinion of AI Risk Research

Though the dangers of 'killer robots' have been talked about for decades, research into the long-term risks of AI only seems to have started being taken seriously with the publication in 2014 of Nick Bostrom's book 'Superintelligence' (Bostrom, 2016). Bostrom argues that the real threat will come not from robots, but from artificial *general* intelligence (AGI): AI which is superhumanly capable across a wide range of different tasks, rather than just the narrow domains that current AI can deal with. Consider AI with superhuman cognitive abilities (without the rest of our ancestral baggage, like emotions) that can be put to work on arbitrary problems, and you get the idea.

Such technology is, of course, not around the corner. Reflecting, though, that over the course of my father's life, we went from complete ignorance of DNA to being able to precisely engineer super-muscular dogs (Regalado, 2015), and from complete lack of digital technology to small devices we can fit in our pockets with radio access to the sum of all human knowledge, it seems within the realms of possibility that AGI may happen within our lifetimes. AGI may be a way away, but not so far as to be completely intangible to us.

Despite being such a long way away, Superintelligence concludes, AI risk research is nonetheless something we need to start working on *now*. Why? Because the advent of AGI is likely to be one of the most momentous events in the history of mankind. After AGI, we're likely to be forced onto one of two paths. There's the 'bad' path, where, for example, AGI allows one organisation or state to assume control; or where an errant AGI set up with an faulty objective gradually consumes all the world's resources in order to achieve its goal. But there's also the 'good' path, where AGI allows us to solve a whole slew of problems that have thus far proved beyond the reach of our small, metabolically-limited brains. Given the all-or-nothing nature of the outcome, Superintelligence argues, and given that it may be difficult to alter our trajectory once popularity of potentially unsafe algorithms has passed some critical level, it's worth us getting started as early as possible on making sure that we get the *good* path.

The most pressing issue right now, however, isn't the immediate need for research. Yes, the proportion of the AI community that's dedicated to risk research is less than what it ideally would be. But given the long-term nature of the problem, what's important is not the starting level, but the rate of growth the field will experience over the coming decades. And this is where I get worried.

The publication of Superintelligence around the middle of 2014 succeeded in bringing awareness of the issue to a broader audience. Some of that audience were in a position to rebroadcast to a broader audience still: over the subsequent months, we saw public statements of concern from the likes of Elon Musk in October 2014 (Gibbs, 2014), Stephen Hawking in December 2014 (Cellan-Jones, 2014), and Bill Gates in January 2015 (Mack, 2015). Though beneficial in publicising the issue, taken out of the context of the broader discussion, these statements seem to have had some undesirable consequences.

One of the consequences has come about through the coincident media hype about recent advances in machine learning. This seems to have given some the impression that the concern about the risks of AGI is based on an assumption of imminence. A report in January to the US Department of Defense by the JASON advisory group, for example, states that "the claimed 'existential threats' posed by AI seem at best uninformed. . . in the midst of an AI revolution, there are no present signs of any corresponding revolution in AGI" (JASON, 2017). But this is not the basis for the concern at all. In fact, it's a measure of just how seriously those involved believe the future dangers to be that *even though* AGI is likely to be a very long way away, it's still worth preparing for now. It would be unfortunate if misunderstanding on this point were to lead to misallocation of resources.

There is, however, a second, more serious consequence. These statements, combined with the above-mentioned Terminator articles, seem to have created a media at-

mosphere where questions about the dangers of AI have become appealingly provocative (Bostrom, 2016). This provocation has, in turn, seem to have stirred up an (understandable) feeling of defensiveness among some in the AI community. Mustafa Suleyman, one of the co-founders of Google DeepMind, for example, was quoted at a conference 2015 telling the audience that "Any talk of a superintelligent machine vacuuming up all the knowledge in the world and then going about making its own decisions are absurd. There are engineers in this room who know how difficult it is to get any input into these systems." (Arthur, 2015)

Indirectly, these statements and the media reaction to them seem to have created a perception of AI risk research as being something of a silly thing to work on. And it's this perception of silliness that makes me concerned about the field's growth.

One problem is that it's going to make it harder to attract more people to the area. Given that the field is already talent-limited (Whittlestone, 2017), it would be a mistake to stunt its growth even further.

The bigger issue, though, is that this perception could lead to the broader AI community becoming actively hostile towards those involved in risk research. If such hostility arises, then no matter how many people are working on risk research, they may be prevented from having any impact. They may, for example, be unable to persuade those pursuing real-world implementation to investigate safer alternatives to existing algorithms (such as the inclusion of reward uncertainty into reward learning algorithms (Alexander, 2017)). Without a sense of everyone being on the same side, the venture seems doomed from the start.

It's hard to say just how much investment in risk research is going to be needed to ensure a 'safe' future AGI-wise. It may be enough to have only a small portion of the AI community as a whole dedicated to the issue. But judging from our current trajectory, there's no guarantee that we'll get that balance right by just letting things happen. That balance looks to be something we'll need to work on *deliberately*.

This brings us to our final section: what can we actually starting doing?

What can we do?

In summary:

- It seems unlikely that unemployment through automation will occur as a step change. Assuming that real-world application of AI continues to follow the Pareto principle, we're more likely to see that change happening gradually. Furthermore, we're more likely to see human-machine hybrid jobs than complete replacement of humans. Given these two factors, and given that the issue of unemployment is already receiving a lot of attention, our further efforts might be better spent on other issues both more pressing and more neglected.

Within this category, I see two particularly important issues:

- Deployment of LAWS based on sophisticated AI could lead to an arms race. An arms race will lead to technology proliferation. Proliferation will make it easier for groups with malicious intent to get their hands on the technology and use it to, for example, oppress a populace.
- AI risk research is in danger of becoming seen as a silly topic. This is concerning partly because the connotation will make it hard to attract extra minds to an already talent-limited field. The bigger concern, however, is that without collaboration between the risk community and the rest of the AI community, the impact of risk research may be limited.

So what can we actually do about these issues?

Despite being perhaps the most urgent problem, LAWS may be the simplest to actually deal with. A lot of progress has been made towards a full ban. All that remains is to make sure that the real issues are not drowned out by irrelevant hyperbole; to maintain sufficient attention on the situation to ensure that the remaining steps towards a full ban take place.

One way that organisations can assist in this is by lending their weight to the push for a ban. Specifically, they can endorse the Campaign to Stop Killer Robots – a group of NGOs that has been instrumental in influencing the UN. Public statements of support, such as those

from Clearpath Robotics in 2014 (Hennessey, 2014), give the campaign clout with which to maintain their influence (Wareham, 2017b).

Organisations may also be able to help by holding events to get the issue known about more widely, encouraging more people to get involved. The positions of the individual members of the Group of Government Experts will be partially a function of, for example, the number of people writing to national representatives, and media coverage resulting from events publicising the issue. Organisations such as the ACM may promote such action directly through member communications; those at academic institutions might organise local talks to get more people aware of what the risks really are.

The question of how to get AI risk research to be taken seriously is a more difficult one. The various tropes have become so firmly established in our collective consciousness that it's going to be hard to directly affect public perception. However, I think there is at least hope for change within the limited scope of the academic community.

One low-hanging fruit may be for more organisations to offer awards and grants for work related to AI risk, as the Future of Life Institute is already doing (Future of Life Institute, 2015a). As well as enabling research, grants may help to signal to the community that risk research is something credible to be working on.

Another source of easy gains may be to encourage universities to offer courses on the broader context of AI – an area that seems to be conspicuously lacking in the curriculum currently. In addition to informing students of the *real* arguments for risk research, such courses could address other problems that have been pointed out in the computer science curriculum, such as the need for ethics education and awareness of the dangers of dataset bias (National Science and Technology Council, 2016). Attracting computer science students towards such courses is, of course, going to be a challenge; but I see a lot of possibility for creative solutions here, such as courses based on readings in science fiction (Burton, Goldsmith, & Mattei, 2015).

Perhaps the most effective remedy to the issue would simply be getting people together to talk about it. Consider, for example, the apparent success of the Future of Life Institute's 'Beneficial AI' conference held at the beginning of the year in Asilomar. Part of the success of the conference was a step towards a greater sense of unity in the field, drawing on the range of expertise represented at the conference to form the Asilomar AI principles – a set of 29 principles agreed by the participants of the conference as important to uphold, touching on aspects from AI safety to ensuring that the benefits of AI will be shared throughout society. However, the event was also apparently successful in starting to break down the cultural barriers surrounding the issue. One participant noted that the conference was in part "a coming-out party for AI safety research. One of the best received talks was about 'breaking the taboo' on the subject, and mentioned a postdoc who had pursued his interest in it secretly lest his professor find out, only to learn later that his professor was also researching it secretly, lest everyone else find out." (Alexander, 2017) Creating more opportunities for this kind of conversation – whether in the form of conferences, an evening of talks, or simply group discussions – can only be a good thing.

Having got a better sense of the bigger picture throughout the course of writing this essay, I find myself feeling optimistic. Though it's clear there are dangers ahead of us – those covered here, along with many others – they're not just being swept under the rug. People *are* taking notice of them.

Of all that I've read about, I think it's the Asilomar conference that has given me the most hope. The fact that people from so many different parts of the field – machine learning (Yann LeCun; Yoshua Bengio), risk research (Nick Bostrom; Eliezer Yudkowsky), funding (Sam Altman; Elon Musk), and so on – were willing to come together to talk about where things are going gives me a sense that culturally, we're on the right track.

Whatever issues we may face over the coming decades, at the broader level, it's nurturing and maintaining this culture that strikes me as the most important thing going forward. It's only through this attitude that we're going to be able to continue making course corrections where necessary – and that long-term, therefore, we really will be able to reach the kind of

future that we all hope AI will take us to.

References

Alexander, S. (2017, 02). *Notes from the Asilomar Conference on Beneficial AI.* Retrieved 2017-02-25, from https://slatestarcodex.com/2017/02/06/notes-from-the-asilomar-conference-on-beneficial-ai

Arkin, R. (2015, 08). *Warfighting Robots Could Reduce Civilian Casualties, So Calling for a Ban Now Is Premature.* Retrieved 2017-02-07, from http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/autonomous-robotic-weapons-could-reduce-civilian-casualties

Arthur, C. (2015, 06). *DeepMind: 'Artificial intelligence is a tool that humans can control and direct'.* Retrieved 2017-02-07, from https://www.theguardian.com/technology/2015/jun/09/deepmind-artificial-intelligence-tool-humans-control

Bostrom, N. (2016). *Superintelligence: Paths, Dangers, Strategies*.

Burton, E., Goldsmith, J., & Mattei, N. (2015). Teaching AI Ethics Using Science Fiction. In *1st International Workshop on AI, Ethics and Society, Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence.*

Cellan-Jones, R. (2014, 12). *Stephen Hawking warns artificial intelligence could end mankind.* Retrieved 2017-02-15, from http://www.bbc.com/news/technology-30290540

Davies, A. (2015, 05). *The World's First Self-Driving Semi-Truck Hits the Road.* Retrieved 2017-02-04, from https://www.wired.com/2015/05/worlds-first-self-driving-semi-truck-hits-road/

Farivar, C. (2015, 11). *Cops pull over Google car for doing 24mph in a 35mph zone.* Retrieved 2017-02-26, from https://arstechnica.com/tech-policy/2015/11/google-self-driving-car-pulled-over-for-not-going-fast-enough

Future of Life Institute. (2015a). *First AI Grant Recipients.* Retrieved 2017-02-25, from https://futureoflife.org/first-ai-grant-recipients

Future of Life Institute. (2015b, 07). *Open Letter on Autonomous Weapons.* Retrieved 2017-02-07, from https://futureoflife.org/open-letter-autonomous-weapons

Gibbs, S. (2014, 10). *Elon Musk: artificial intelligence is our biggest existential threat.* Retrieved 2017-02-15, from https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat

Goose, S., & Wareham, M. (2017, 01). *The Growing International Movement Against Killer Robots.* Retrieved 2017-02-07, from http://hir.harvard.edu/growing-international-movement-killer-robots/

Hennessey, M. (2014, 08). *Clearpath Robotics Takes Stance Against 'Killer Robots'.* Retrieved 2017-02-22, from http://www.clearpathrobotics.com/press_release/clearpath-takes-stance-against-killer-robots/

JASON. (2017, 01). *Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD.* Retrieved 2017-02-28, from https://fas.org/irp/agency/dod/jason/ai-dod.pdf

Kitroeff, N. (2016, 09). *Robots could replace 1.7 million American truckers in the next decade.* Retrieved 2017-02-04, from http://www.latimes.com/projects/la-fi-automated-trucks-labor-20160924

Lee, T. B. (2016, 10). *Self-driving trucks are here, but they won't put truck drivers out of work - yet.* Retrieved 2017-01-31, from http://www.vox.com/new-money/2016/10/25/13404974/otto-self-driving-trucks

Mack, E. (2015, 01). *Bill Gates Says You Should Worry About Artificial Intelligence.* Retrieved 2017-02-15, from https://www.forbes.com/sites/ericmack/2015/01/28/bill-gates-also-worries

-artificial-intelligence-is-a-threat

McArdle, M. (2015, 05). *When Will Self-Driving Trucks Destroy America?* Retrieved 2017-01-31, from http://origin-www.bloombergview.com/articles/2015-05-27/when-will-self-driving-trucks-destroy-america-

National Science and Technology Council. (2016, 10). *Preparing for the Future of Artificial Intelligence.* Retrieved 2017-01-13, from https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Rabiroff, J. (2010, 07). *Machine gun-toting robots deployed on DMZ.* Retrieved 2017-02-23, from http://www.stripes.com/news/pacific/korea/machine-gun-toting-robots-deployed-on-dmz-1.110809

Regalado, A. (2015, 10). *First Gene-Edited Dogs Reported in China.* Retrieved 2017-02-23, from https://www.technologyreview.com/s/542616/first-gene-edited-dogs-reported-in-china/

Russell, S., Tegmark, M., & Walsh, T. (2015, 08). *Why We Really Should Ban Autonomous Weapons: A Response.* Retrieved 2017-02-08, from http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/why-we-really-should-ban-autonomous-weapons

Solon, O. (2016, 06). *Self-driving trucks: what's the future for America's 3.5 million truckers?* Retrieved 2017-01-31, from https://www.theguardian.com/technology/2016/jun/17/self-driving-trucks-impact-on-drivers-jobs-us

Tesla. (2015, 12). *Correction to article: "The First Person to Hack the iPhone Built a Self-Driving Car".* Retrieved 2017-02-20, from https://www.tesla.com/support/correction-article-first-person-hack-iphone-built-self-driving-car

Vance, A. (2015, 12). *The First Person to Hack the iPhone Built a Self-Driving Car.* Retrieved 2017-02-20, from https://www.bloomberg.com/features/2015-george-hotz-self-driving-car/

Wareham, M. (2017a, 01). *Banning Killer Robots in 2017.* Retrieved 2017-02-08, from https://www.hrw.org/news/2017/01/15/banning-killer-robots-2017

Wareham, M. (2017b, 08). Personal communication.

Whittlestone, J. (2017). *Research into risks from artificial intelligence.* Retrieved 2017-02-23, from https://80000hours.org/career-reviews/artificial-intelligence-risk-research

Matthew Rahtz is a student on the MSc in Neural Systems and Computation joint between the University of Zürich and ETH Zürich. His research interests are in the long-term risks of artificial intelligence, with a particular focus on safe reinforcement learning.