CASE STUDY 1

TEXT GENERATING AI TECHNOLOGIES

Question

GPT-2 is a new AI text generation model that has been developed by OpenAI, a non-profit research organization. OpenAI decided not to release GPT-2 completely due to concerns about potential malicious use, i.e. generating deceptive, biased, or abusive language at scale.

Is OpenAI's GPT-2 release strategy ethically justified?

References for obtaining further information:

- OpenAi Blogpost, https://openai.com/blog/better-language-models/ posted on February 14th, 2019; Authors Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage & Ilya Sutskever (accessed on March 15th, 2019)
- OpenAI Charter, https://openai.com/charter/ (accessed on April 19th, 2019)
- OpenAI Code: https://github.com/openai/gpt-2 (accessed on April 19th, 2019)

Answer suggestion

1) Describe the moral question/problem

When it comes to the release of GPT-2 the alternatives available are:

- No release at all
- A partial release
- A complete release

OpenAI has chosen the partial release. The question to be addressed is whether this is indeed the best release strategy from an ethical point of view. This question is tackled from a utilitarian point of view. In light of this ethical perspective the question can thus be reformulated as: Does the partial release strategy contribute more to the common good than the other two release strategies?

2) Gather the relevant facts

In the **no-release scenario**, nobody outside OpenAI would initially know about the existence of GPT-2. However, leaks might obviously occur. In addition, in the short-to-medium terms other research organizations would probably develop similarly powerful models. Not releasing GPT-2 would thus most likely not avoid the development of similar technologies by others. It would at most postpone the emergence of this technology. In addition, the no-release scenario would not be ideal to facilitate a debate about the ethics of powerful text generating AI systems. After all, there would not be any trigger for such a discussion as nobody outside OpenAI would know about the existence of such potent dual use technologies. Of course, OpenAI could still start such a discussion without releasing anything about GPT-2. However, when asked why they engaged in such a debate, it would be disingenuous not to reveal anything about GPT-2. If in such a context OpenAI were to disclose the true backdrop of their eagerness to spark a debate, this would prompt the partial-release scenario.

The **partial-release scenario** seems ideal to both postpone the emergence of the full-fledged text generating AI technologies and to trigger a lively discussion of the dual use character of the same. On the one hand, it would postpone the appearance of the technology, thus generating time for solid reflection and the development of policy frameworks, if need be. On the other hand, it would expose enough about the prospects of the technology to make people aware of the need for a substantial discussion of the dual use problems of text generating models without anybody being immediately able to mobilize the potential of the full-fledged model.

In a **complete-release scenario**, everybody would be immediately able to use the full-fledged version of GPT-2. This would trigger the fastest development of the AI technologies. It would not generate any leeway for a discussion to take place and necessary policy frameworks to be developed in advance of substantial societal impacts. Discussion would not be perceived as acute because of the full release.

3) Consider the relevant ethical principles (informed by the ethical theory of your choice)

As mentioned above, the case is analysed from a utilitarian point of view. This means that the focus is on the expected ratio of benefits and harms of the three different release scenarios respectively.

The principle used: When choosing between alternative actions it is ethically obligatory to choose that action that has will yield the best expected ratio of benefits and harms

In all three scenarios, full-fledged text generating AI systems would eventually be developed meaning that one would encounter the likely positive and negative societal impacts of sophisticated, general language models as listed by OpenAI in their blogpost.

The potential benefits are: the development of "AI writing assistants", "more capable dialogue agents", "unsupervised translation between languages", and "better speech recognition systems" (OpenAi Blogpost, 2019). The potential harms are: the generation of "misleading news articles", the impersonation of others online, the automation of "the production of abusive or faked content to post on social media", and the automation of "the production of spam/phishing content" (OpenAi Blogpost, 2019). The only significant difference between the three different scenarios would be the time and space left in advance of these anticipated effects occurring for the development of policy and regulatory frameworks to enhance the projected benefits and soften the predicted harms.

As discussed above, in the **no-release scenario** it would be more difficult for OpenAI to start a debate. If they did suddenly engage in discussion and would be asked why, they would either have to be disingenuous or – by disclosing their real motives – move to a partial release scenario. Sticking to the no release scenario would mean that everybody outside OpenAI would remain in ignorance about the technological developments that would be about to take place (because others would develop the same technologies eventually), leaving less room for debate and policy development. OpenAI would not spark any substantial debate; neither

could it stop the development of the technology by others. So, the sketched positive and negative societal impacts would eventually occur because others would develop the AI systems.

The **partial release scenario** is the ideal scenario to start a debate, which would be in line with OpenAI's charter. Third parties would thus be better informed about the prospects of powerful text generating AI systems and have some time to discuss and develop policies allowing the development of measures that might mitigate the potential harms and enhance the projected benefits. Hence, positive and negative societal impacts would occur, but there would be a change that the ratio of benefits over harms turned out to be more advantageous than the benefit/harm ratio in the no release scenario.

The **complete release scenario** would leave others less time for debate and policies to be developed before substantial impacts of the technology would likely occur. Open AI might still play a role in facilitating debate. However, the time frame would not be accommodating for the development of effective measures. This is the fastest pathway for any positive and negative societal impacts to materialize, leaving the least amount of time for debate. Hence the benefit/harm ratio is likely less advantageous than the ratio in the partial release scenario.

OpenAI's assessment of potential benefits and harms seems broadly correct. The applications that are branded as beneficial might, for example, make life easier and work more effective. The negative impacts would severely

worsen cybersecurity in general. The generation of misleading news articles would spoil cyberspace epistemologically. It would systematically reduce the epistemological trustworthiness of cyberspace. Enhanced capabilities to impersonate others online would be used to significantly scale up social engineering attacks. The automation of the production of abusive or faked content to post on social media would disrupt the political process and again spread mistrust in the veracity of cyberspace content in general. The automation of the production of spam/phishing content would generate a massive cluster of annoyances and problems on its own. All in all, these malicious applications could lead to erosion of trust, social disruption and reputational damage amongst others. It would therefore be worthwhile to try and create some time and space for the development of appropriate regulatory frameworks and policies.

4) Make your decision/assessment

In all three scenarios, full-fledged text generating models would be available either immediately (complete-release scenario) or in the short-to-medium terms (the other two scenarios). The partial-release scenario is the only one that optimizes the conditions for a

discussion which might prove beneficial when it comes to the development of guidelines and policy frameworks in order to diminish the potential negative and enhance the expected beneficial societal impacts of this emerging technology. That is why, from a utilitarian point of view, OpenAI's release strategy, i.e. partial release, is indeed the best option of the three alternatives because it is likely to enhance the expected benefits and reduce the expected harms.

References

- Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B. and Anderson, H., 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.
- OpenAi Blogpost, https://openai.com/blog/better-language-models/ posted on February 14th, 2019; Authors Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage & Ilya Sutskever (accessed on March 15th, 2019)
- OpenAI Charter, https://openai.com/charter/ (accessed on April 19th, 2019)
- OpenAI Code: https://github.com/openai/gpt-2 (accessed on April 19th, 2019)
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., Language Models are Unsupervised Multitask Learners. https://d4mucfpksywv.cloudfront.net/better-language models/language_models_are_unsupervised_multitask_learners.pdf (accessed on April 19th, 2019)

CASE STUDY 2

BLOCKING PAYMENT SITES IN RANSOMWARE ATTACKS

Question

In ransomware attacks, the attacker usually provides the victim with an extortion note that indicates how the victim can pay the ransom. Some use a tor2web domain that basically is the entry point into the TOR network but is accessible from the "normal" Internet. This payment allows Computer Emergency Response Teams (CERTs) to block such payment sites. This disrupts the money flow of the attackers as it cuts out the main purpose of ransomware attacks i.e. monetary gain. If blocking is successful and the lucrative element is removed, the motivation of criminals to engage in this activity will drop. While blocking payment websites is technically possible, there is a political component to it which relates to censorship on the Internet. There is also the issue of autonomy to consider as blocking such sites deprives the victim of the possibility to recover their encrypted files.

Is CERTs blocking payment sites ethically justified?

References for obtaining further information:

- Europol, European Cybercrime Centre. Internet Organised Crime Threat Assessment (IOCTA) 2018. Available from https://www.europol.europa.eu/internet-organised-crime-threat-assessment-2018
- Fruhlinger, J., 2018. What is ransomware? How these attacks work and how to recover from them. [Online] Available at: https://www.csoonline.com/article/3236183/what-is-ransomware-how-it-works-and-how-to-remove-it.html [Accessed 4 June 2019].

Answer suggestion

1) Describe the moral question/problem

The alternatives available to answer the question are:

- Block payments
- Don't block payments

CERTs have blocked payment sites in the past however the question to be addressed is whether this is indeed the best option from an ethical point of view. This question is tackled from a utilitarian perspective. In light of this ethical perspective the question can thus be reformulated as: Does blocking payment sites in ransomware attacks contribute more to the common good than not blocking payments?

2) Gather the relevant facts

The case at hand is analysed from a utilitarian point of view. This section concentrates on the expected impact of blocking or not blocking payment websites.

The advantage of blocking the payment website: It contributes to a safer cyberenvironment. It does not allow businesses to reward criminals for their activity as they have blocked the payment pathway. This can act as a disincentive for criminals as the lucrative element to the attack is removed. By making the endeavour less appealing to criminals, it becomes less attractive and less popular route of attack.

The disadvantage of blocking payment websites: It removes the option for the victim to retrieve their data. For a private business or public institution whose back-ups systems have been encrypted as part of the attack, this could be critical as not having access to critical data for an extended period of time (the average downtime is approximately seven days but there have been reports of downtime lasting more than six weeks) could cost the business time and resources that they may not have. Resources will need to be reallocated in efforts to recreate the encrypted data and/or rebuild workstations. There is also the financial cost that downtime will incur and the longer the downtime, the higher the cost for the enterprise. This is potentially a serious problem for enterprises with a low tolerance for an attack and who are willing to pay large sums of money to ransomware attackers in return for their data. Take a hospital as an example. The potential collateral damage could be catastrophic and could be the difference between life and death. In such a situation, it is reasonable to assume that the hospital would

want to avoid causing unnecessary disruption to critical, lifesaving services. Therefore, the cost of CERTs killing payment websites could be argued as being too high.

The advantages of not blocking payment websites: It gives the victim the opportunity to pay the ransom. If the ransom is immediately paid, let us assume that the decryption key is provided, and the data obtained is returned in its original state. In such cases, the victim can resume business as normal as access to encrypted files will be returned and downtime will be minimal. The advantage of not blocking payment websites also gives the victim the opportunity to not pay the attacker, however history suggests that businesses, especially those with low tolerances for ransomware attacks, do not choose this option and pay the attacker.

The disadvantages of not blocking payment websites: CERTs risk handing over the responsibility to vulnerable victims to decide whether to pay or not pay the ransom. This could potentially lead to all victims paying the ransom. The most obvious harm of paying the ransom is that the victim is rewarding criminals for criminal behaviour. Paying the ransom reinforces the idea that the ransomware business model works and as a result, criminals can rest assured that their continued efforts to extort money from victims works. Another potential harm associated with paying the ransom revolves around what the ransom is contributing to. It is generally unknown to the victim to whom, where or what activity the ransom money will be used for however it is fair to assume that the ransom money will be used to fund further illegal activities in the cyber or physical world. i.e. pay for the criminals' everyday expenses such as groceries, fund other illegal endeavours like purchasing more sophisticated software to launch better attacks and/or contribute towards other illegal activities such as underage prostitution, or purchase weapons or drugs.

3) Consider the relevant ethical principles (informed by the ethical theory of your choice)

As mentioned above, the case is analysed from a utilitarian point of view. This means that the focus is on the expected ratio of benefits and harms of the two different scenarios respectively.

The principle used: When choosing between alternative actions it is ethically obligatory to choose that action that will yield the best expected ratio of benefits and harms.

Short-term benefits of blocking all payment sites: This could be taken by CERTs as a moral objection to the ransomware business model to which cybercriminals benefit. In the short-term, the criminals are not rewarded for criminal behaviour. As mentioned above, this will

benefit wider society not only in the short term, but in the long term too as it undermines the basis of the ransomware business model i.e. to extort money from victims.

Long-term benefits of blocking all payment sites: This has the potential to eliminate the lucrative lure for future ransomware attackers as cash-flow from victims will be an uncertainty and become a less attractive way to exploit individuals or businesses.

Short-term harm of blocking all payment sites: A blanket decision to block all payment sites can cause harm in the short-term for the victim. This harm pivots on the victim's inability to pay for the return of their encrypted files or data. As a result, products or services might become completely unavailable (depending on the types of ransomware used) for a short period of time. This immediate pause can create a monumental crisis for the victim especially when back-up systems are not accessible. Day-to-day activities will be disrupted until important files and systems are back up and running and the financial cost of the downtime associated with rebuilding those systems can be significant. Individuals who are directly connected to the business i.e. management, employees, shareholders, customers and suppliers will also feel the short-term effects of CERTs blocking payment sites as their day to day tasks will be interrupted.

Arguably, paying the attacker encourages bad behaviour and will contribute to its growth. However, lest not forget that it is still possible for ransomware authors to write and sell more sophisticated ransomware software as an offensive strategy to CERTs blocking payment websites. For example, criminals can decide to provide their ad-

dresses directly in the TOR network or provide unique bitcoin addresses for each victim, which has been done by later versions of ransomware. CERTs are unable to block payments when this is done.

Long-term harm of blocking all payment sites: Blocking all payment websites could start a debate on censorship. It suggests that CERTs do not trust that victims will make the right choice and they subsequently must make the decision for them. Questions worth considering include, should CERTs try to block all payment websites unequivocally? Is this a type of censorship? Is it a necessary evil or completely justified security measure? Is blocking all payment websites a move in the wrong direction i.e. a move towards a nanny-state? Should the harms and benefits of blocking payment websites be considered on a case by case basis?

Short- and long-term benefits of not blocking payment websites: If it is possible for CERTs to block payments sites but choose not to, this allows the victim to autonomously decide whether to pay the ransom. It also suggests that CERTs trust civilians to make the right decision. If the victim decides that the encrypted files are unimportant or they can revert to back-ups, the victim themselves might choose to not pay the ransom without their hand being forced. The victim might also choose to not pay the ransom on the basis of moral principles i.e. not paying a criminal for theft or fund nefarious criminal activities. This scenario has the same benefits in the short and long term as the blocking scenario with one small difference; the autonomy of the victim is not compromised. If the victim decides to pay the ransom, the attacker is the one and only true beneficiary of the outcome. If the victim decides to not pay the ransom, the victim and wider society benefit in the long term.

Short and long-term harm of not blocking: The victim has the choice to pay or not pay and paying might appear to be the obvious solution to the victim in both the short and long term. Victims may not have the foresight or wherewithal to know that funding cybercriminal activity can harm wider society, in both the cyber and physical world. There is always the possibility that the victim does not care for the impact that paying the ransom will have on civil society – they simply want to return to business as normal as soon as physically possible. The victim may also not be aware of the risks associated with paying the ransom. For example, there is always a chance that either, a decryption key does not exist or the software itself is unstable and the data itself is irretrievable with or without a key. This means that even when the victim pays the ransom, there is no guarantee that their files will be returned in their original state.

4) Make your decision/assessment

The ratio of benefits to harms in the blocking scenario would be higher and turn out to be more advantageous than the benefit/harm ratio in the not blocking payment servers as the former in all cases removes cash flow for ransomware attackers whereas the latter does not guarantee it. While removing the potential for businesses to reward criminals by paying ransoms, the blocking scenario proves more beneficial when it comes to the mitigation of cybercrime as paying the ransom will only fund further attacks and other criminal activity.

While blocking ransomware payment sites is not technically possible for all attacks, this action has the potential to diminish the long-term continuity of ransomware attacks and thus improve the safety of cyberspace for those who use and operate within cyberspace. From a utilitarian point of view, blocking payment sites where possible, is the best option of the two alternatives because it is likely to enhance the expected benefits and reduce the expected harms.

References

- Europol, European Cybercrime Centre. Internet Organised Crime Threat Assessment (IOCTA) 2018. Available from https://www.europol.europa.eu/internet-organised-crime-threat-assessment-2018
- Europol, 2018. Internet Organised Crime Threat Assessment. [Online] Available at: https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment [Accessed 1 May 2019].
- Fruhlinger, J., 2018. What is ransomware? How these attacks work and how to recover from them. [Online] Available at: https://www.csoonline.com/article/3236183/what-is-ransomware-how-it-works-and-how-to-remove-it.html [Accessed 4 June 2019].
- Goodpaster, K., 1991. Business Ethics and Stakeholder Analysis. Business Ethic Quarterly, 1(1), pp. 53-73.
- IBM, P. I., 2018. 2018 Cost of a Data Breach Study: Global Overview. [Online] Available at: https://databreachcalculator.mybluemix.net/assets/2018_Global_Cost_of_a_Data_Breach_Report.pdf
- Loi, Michele, and Markus Christen. 2019. “Ethical Frameworks for Cybersecurity.” In The Ethics of Cybersecurity, edited by Markus Christen, Bert Gordijn, and Michele Loi. Springer. https://canvas-project.eu/results/canvas-book.html.
- Mateiu, M., 2018. The Ultmate Guide to Ransomware. [Online] Available at: https://www.avg.com/en/signal/what-is-ransomware [Accessed 1 May 2019].
- Morgan, S., 2016. Hackerapocalyse: A Cybercrime Revelation, s.l.: Cybersecurity Ventures.
- National Audit Office, 2017. Investigation WannaCry Cyber Attack and the NHS. [Online] Available at: https://www.nao.org.uk/report/investigation-wannacry-cyber-attack-and-the-nhs/ [Accessed 29 April 2019].
- Phillips, R., Freeman, E. & Wicks, A., 2003. What Stakeholder Theory Is Not. Business Ethics Quarterly, 13(4), pp. 479-502.
- Symantec, 2018. 2018 Internet Security Threat Report, s.l.: Symantec.
- Walker, J., 2018. Ransomware remains biggest malware threat in 2018, says Europol. [Online] Available at: https://portswigger.net/daily-swig/ransomware-remains-biggest-malware-threat-in-2018-says-europol [Accessed 1 May 2019].

CASE STUDY 3

SYSTEMS ADMINISTRATOR DISCOVERS CONFIDENTIAL INFORMATION

Question

You are system administrator in a large hospital. In the course of cyber security testing, you end up accessing a clinical note that a psychiatrist stored in a section of his computer which he mistakenly assumed to be inaccessible to anyone else. You have to open the file to verify its integrity and you read the text of the note, which says "since he has killed his wife, the sense of guilt is devouring him". The name of the patient is not written on the note, but you have to de-anonymize the patient ID to determine who is his treating physician. The law of your country makes it a crime for public officials not to report a crime to the police and your role in the hospital is in fact that of a public official. On the other hand, it is not a crime for the psychiatrist not to report. On the contrary, reporting would be a breach of his professional duty, and even cost him expulsion from the psychiatric order. (In your country, the guild of physicians endorses an absolutist version of patient-doctor confidentiality that knows no exception).

Do you, as system administrator, report the crime to the police that you found by chance during a cybersecurity testing?

Answer suggestion

1) Describe the moral question/problem

The case at hand is analysed from a deontological perspective (duty of confidentiality), from a utilitarian perspective, from a particularist perspective, from a consequentialist maximin perspective, and from a politically conservative perspective. The moral question has the shape of a dilemma. The law grants special rights and duties of confidentiality to the medical profession, for a reason. But the law does not consider the role of third parties such as yourself. Yet, it feels like that by reporting the crime, medical confidentiality is violated. On the other hand, this person may be dangerous. And anyway, it is a crime for you not to report.

2) Gather the relevant facts

You need more information so you plan to talk to the doctor about this particular case. For example, you need to know if the man has already been arrested or his crime has gone undetected. The doctor may also tell you that the patient is actually dangerous and he fears he may kill again, which would convince you to report the crime. Or on the contrary, you may end up discovering that you equivocated the meaning of the doctor's note. It is possible, for example, that the doctor only meant "kills his wife" as a metaphor, or as a thought of a delusional patient without reference to reality. If so, reporting the crime (before talking to the doctor) risks will open up an investigation that will be incredibly costly for patient, and the doctor, and wholly unnecessary. You hope that talking with the doctor will bring light on the situation.

On the other hand, while the doctor may sympathize with your desire to know, deontological medical principles prevent him to talk with you and reveal anything about his patient.

The relevant legal facts are:

1. You have a professional duty and you are legally authorized to de-identify the patient's ID in order to assess who is the treating physician and restore the integrity of the file system.

2. In your country, the law does not consider system administrators as bound by norms of confidentiality against crimes for which a legal duty to report exists (it only protects the confidentiality of doctors, lawyers, and priests). But this appears merely accidental. One could think that the law has not been updated to the new circumstances which expose patients to breach of their confidentiality, since the role of system administrators was not foreseen in these laws. If sick people feel that their confidential medical information is not protected, they will not access medical professionals. This is exactly what the law and professional norms of confidentiality aim to avoid.

3. If you talk about the case with the chief manager of the hospital, she will follow the recommendations of the legal office and require you to report the case to the police.

The personal prudential facts are that if you simply move the file to its original place without reporting the case, it is very unlikely that anyone will detect your violation of the law. It is virtually impossible for anyone to know, and be able to prove such fact.

3) Consider the relevant principles / moral theories

3.1 Confidentiality principle

Patient-doctor confidentiality: there are moral reasons behind the law which protects the confidentiality of the patient-doctor relations. Society seems to have decided that the duty of confidentiality matters more than its negative consequences, e.g. the possibility that a criminal will not be detected and could reoffend.

But is this principle relevant in this case? If one considers the moral principle behind the law, rather than its letteral meaning, it feels that the moral principle of confidentiality is relevant for you, not just for the doctor. After all, the relevant information is medically confidential also to allow patients to talk with their doctors without the fear that they will be reported to the police. So, the moral rationale of confidentiality extends naturally, from the moral point of view, to your case.

3.2. Utilitarianism

From a utilitarian perspective you are recommended to do what produces the highest net value of well-being (the sum of aggregate individual benefits – aggregate individual harms). In order to determine what that choice is, you try to collect more information. Thus, you need to talk to the doctor, hoping that the doctor violates his duty of confidentiality. This information will allow you to assess the best course on action from a consequentialist point of view (e.g. what is the line of conduct that minimizes harm?). In particular, it is ethically desirable that the doctor should know that you have gained access to this information, also in order to avoid repeating the same mistake. Moreover, you may have misinterpreted the note. Or, conversely, the doctor may tell you, or let you understand, that the patient is truly dangerous, but his hands are tied by his professional obligations. He may violate his duty of confidentiality and let you understand that he feels much better now that someone would report the criminal to the police.

3.3. Particularism

Talking to the doctor is also supported by the meta-ethical theory of particularism (Dancy 2000). According to particularisms, ethical action cannot be derived from moral principles, no matter how many and how complex. Moral principles are at most rules of thumb. (Consider the principle “pleasure is good”. Is the pleasure of a sadist good? Is the world with sadist pleasure a better place, compared to a world in which the same wrongs are not accompanied by the same feelings?) From the particularist point of view, we can know what is right only by carefully analyzing the specificities of each particular act in its context. Moral rightness consists in perceiving the right ‘shape’ of the situation, which is never simply a matter of applying general rules. Every principle has exceptions. Hence, from a particularist perspective, it is mandatory that you should get as much information from the doctor, so that you may be able to perceive the real ‘moral shape’ of the case in question and make a particular judgment about it.

3.4 Conservative principle of respect for the law and tradition

Conservative principles of respect for the law and tradition. Arguably, a conservative moral outlook implies that you should follow the law and report the crime. According to Thomas Sowell (2007) the politically conservative mindset – the ‘constrained vision’ as he labels it – is essentially characterized by the following principles:

1) Human individuals are limited in their cognitive and moral capacities: they have limited information, they fall prey to cognitive biases, and their motivations are selfish.

2) As a result of (1), individuals are not to be trusted to make moral judgments, in conflict to the law and tradition (but in exceptional cases, e.g. laws that violate human rights).

3) As a result of (2), conservativism prescribes respect for the laws and traditions of one's country, which are less affected by individual biases and interests as different individual biases cancel each other out.

3.5. Maximin consequentialism

The maximin principle entails that we should choose the course of action that maximizes good consequences (or equivalently minimizes harm), in the worst-case scenario for each option (Loi and Christen 2019). In this context, the principle should be applied as follows:

1) Ask yourself what is the worst possible consequence of not reporting the crime to the police

2) Ask yourself what is the worst possible consequence of reporting the crime to the police

3) Choose the act that minimizes harm

4) Make your decision/assessment

This is a problematic case in which different moral theories support different choices.

Deontology: you have a moral duty to respect the confidentiality of the medical records (even if you do not have a legal duty): you do not report the crime.

Utilitarianism and particularism: try to talk to the doctor, hoping he reveals the relevant facts which will allow you to better understand the situation and its consequences. Then you decide based on what maximizes the good (utilitarianism), or based on a perception of the salient moral features of the situation which cannot be explained by a limited set of principles (particularism).

Conservativism: follow the law and report the crime, because it is the law, which is more likely to be ethically correct, than the individual moral conscience. If you believe that the law is outdated, you hope that your case will be discussed in the media, picked up by activists, and the discussion will cause a new social consensus to emerge.

Maximin:

-Worst case scenario by not reporting: the patient is a killer and will kill other innocent persons.

-Worst case scenario by reporting: the breach of confidentiality becomes a publicly debated fact that undermines the patients' trust in the confidentiality of their medical records. Criminals who have acted against the law or have engaged in shameful behaviour no longer feel safe revealing these facts to their treating physicians.

If you believe that the first outcome is morally worst than the second, you report to the police.

References

– Dancy, Jonathan. 2000. Practical Reality. Oxford ; New York: Oxford University Press.
– Loi, Michele, and Markus Christen. 2019. “Ethical Frameworks for Cybersecurity.” In The Ethics of Cybersecurity, edited by Markus Christen, Bert Gordijn, and Michele Loi. Springer. https://canvas-project.eu/results/canvas-book.html.
– Sowell, Thomas. 2007. A Conflict of Visions: Ideological Origins of Political Struggles. New York: Basic Books.

CASE STUDY 4

CARDIAC PACEMAKERS AND OTHER IMPLANTABLE MEDICAL DEVICES

Question

Implantable medical devices (IMDs) are employed to improve the quality of patient's life. Devices such as cardiac pacemakers, insulin pumps, biosensors and cochlear implants offer therapeutic, monitoring and even life-saving benefits. More and more IMDs are wirelessly networked and can be connected to other devices to, for example, monitor operational readiness, set parameters, exchange data or install software updates.

However, for some years, there have been reports about the dangers of implantable medical devices. In addition to risks of unintentional loss of function due to hardware defects and software errors, the connectivity of IMDs allows for malicious attacks on the devices (e.g. Baranchuck et al. 2018, Coventry & Branley 2018, Mohan 2014, Ransford et al. 2014).

Is the use of implantable medical devices morally justifiable despite, in case of malfunction, they put the patient's health or even life at risk?

References for obtaining further information:

- Baranchuk, A., Refaat, M. M., Patton, K. K., Chung, M. K., Krishnan, K., Kutyifa, V., ... Lakkireddy, D. R. (2018). Cybersecurity for cardiac implantable electronic devices: What should you know? Journal of the American College of Cardiology, 71(11), 1284–1288. https://doi.org/10.1016/j.jacc.2018.01.023
- Rios, B., & Butts, J. (2018). Understanding and Exploiting Implanted Medical Devices, https://www.blackhat.com/us-18/briefings.html#understanding-and-exploiting-implanted-medical-devices, accessed 2018-08-28.

Answer suggestion

1) Describe the moral question/problem

The medical benefits of IMDs are undeniable, but at the same time the, at least potential, risks are very great, because in the event of malfunction of an IMD the health or even the life of a patient is at risk.

2) Gather the relevant facts

IMDs serve the primary aim of increasing the physical safety of patients. Wireless IMDs are de-signed to enable continuous monitoring of vital parameters and faster communication with health care professionals both routinely and in emergency cases. While this faster access aims to enable health care professionals to use medical data more quickly, efficiently and flexibly to perform successful treatment, lack of transparency about who under what circumstances can access what information does not help to ensure patient consent and control (Mohan 2014). In addition, it seems to be a problem that patients do not have direct access to information stored in IMDs, particularly in the case of so-called 'closed-loop-devices', although these data could inform them about their own body and health status (Alexander 2018, Ransford et al. 2014).

If patients think that they might have little or no control over their own health-related data, in the long run that could contribute to a loss of confidence in health technology as well as in health care professionals. Because IMDs can be attacked and personal data stolen, patients may perceive danger for themselves and their data and thus for privacy and trust. Furthermore, there is the risk that implant users will be discriminated against as a consequence of unauthorized access to sensitive data, their uncontrollable use and disclosure to third parties (Burleson and & Carrara 2014, Coventry and & Branley 2018, Ransford et al. 2014).

Another possible negative effect on patients' trust is the lack of a clear attribution of (moral) responsibility to the various stakeholders involved (e.g. manufacturers and designers, health care professionals, insurance companies, legislators and regulators), which pursue different interests and not always primarily focus on patients' well-being (Alexander 2018; Baranchuck et al. 2018, Burns et al. 2016).

If patients were to decide who exactly has access to their IMD or if the access would be at least (through technical or regulatory measures) more protected, however, other problems (in addition to the ones mentioned above) would arise: "Requiring users to authenticate to a device before altering its functionality is a boon for security, but it introduces risks in case of an emergency. A medical professional may need to reprogram or disable a device to effectively treat a patient. [...] [E]ncryption or other strong authentication mechanisms could make such emergency measures impossible if the patient is unconscious or the facility does not possess a programming device with a required shared secret." (Ransford et al. 2014, 170). The conflict between usability and security does not only occur with the use by health care professionals. In the case of an open-loop system in which patients have access to information stored in the device, their computer literacy level must be considered to make sure that patients with little technical knowledge and understanding for security would not have disadvantages. The degree of dependency and the level of risk must also be taken into account (Alexander 2018; Ransford et al. 2014).

3) Consider the relevant principles / moral theories

In what follows, the case is analysed from the point of view of principlism (Beauchamp & Childress 2009, Loi and Christen 2019, Loi et al 2019, Weber and Kleine 2019). This means that the focus is on the following four principles:

- Respect for autonomy as a negative obligation is to avoid interfering in other people's freely made decisions. Understanding respect for autonomy as a positive obligation means informing people comprehensibly and thoroughly about all aspects of a decision, for example about its consequences. Respect for autonomy also may "[...] affect rights and obligations of liberty, privacy, confidentiality, truthfulness, and informed consent [...]" (Beauchamp & Childress 2009, 104).
- The principle of non-maleficence is being derived from the classic quote "above all, do no harm" which often is ascribed to the Hippocratic Oath. At the heart of this principle is the imperative not to harm or ill-treat anyone, especially patients.
- Beneficence must be distinguished from non-maleficence. According to Beauchamp and Childress (2009, 197), "[m]orality requires not only that we treat persons autonomously and refrain from harming them, but also that we contribute to their welfare." Consequently, care must always be taken to ensure that actions that are intended to be benevolent do indeed contribute to a benefit; advantages and disadvantages, risks and opportunities as well as costs and benefits of those actions must therefore be weighed up.
- Justice as a principle is even more difficult to grasp than the other three principles, since the different existing theories of justice produce very different results. For the purposes of our considerations, justice is to be translated as guarantee of fair opportunities and prevention of unfair discrimination, for instance based on gender or ethnicity. Justice also means that scarce resources should not be wasted; in addition it must be noted that these resources often have to be provided by others, for example by the insured, so that economically use is required.

Principlism is far from being undisputed in biomedical ethics (e.g. Clouser & Gert 1990; Hine 2011; McGrath 1998; Sorell 2011). Nevertheless, it remains highly influential for the scholarly thinking about ethical issues arising (not only) in the health domain. Thus, it should be used as a starting point for the following ethical evaluation.

4) Make your decision/assessment

Looking at the respective principles, the answer to the moral question posed above is very different. In many cases of using IMDs, the autonomy of the patients is not strengthened because they do not have access to the data necessary for the function of the IMD or produced by it. However, it can be argued that this access could mean that the patients could harm themselves through misuse of the IMD and this should not be allowed by physicians in order to comply with the principle of non-maleficence.

At the same time, however, the risks of using IMDs are considerable since the potential harm to patients can be quite significant. On the one hand, if this is taken into account, the principle of non-maleficence would require that IMDs should not be used. On the other hand, there is no example of real-world targeted cyberattacks on IMDs. Damage has not yet occurred but the benefits of IMDs are already obvious. The principle of beneficence therefore calls for the use of IMDs even when there are risks involved. Here it becomes evident that the decision as to which principle should be given greater weight is ultimately based on the consideration of opportunities and risks as well as of benefits and harms. However, the assessment of opportunities and risks in particular is often based on subjective assessments – this should be made clear at all times.

If the principle of justice is also taken into account, the moral assessment of IMDs becomes more complex and perhaps even contradictory. The hardening of IMDs against cyberattacks is technically complex and therefore ultimately expensive. Thus hardened IMDs would probably be so expensive that they would only be used in small numbers because many patients would not be able to pay for them, or because health insurance companies would refuse to pay for them. First, this would violate the principle of justice because patients with limited economic resources would no longer receive the treatment they actually need. Second, the principle of beneficence would be undermined, as a possible medical benefit would not be realized. Finally, the principle of non-maleficence would also be infringed, since without the use of IMDs, many patients would be much worse off.

All this can be used as an argument that it is morally justified to use unsafe IMDs. However, this would again in turn mean violating principles. In any case, the principle of non-maleficence would be violated because patients are knowingly exposed to a risk. Furthermore, it could be argued that the principle of beneficence would also be negatively affected, as patients would not receive the technically best possible treatment.

But if these considerations are taken into account to justify the use of hardened IMDs, the objections already mentioned would become effective again. In this case, it should also be recognised that considerable economic resources would have to be used to avert damage that has never occurred before. This would be questionable from an economic point of view and would actually pose a problem of justice, since the resources used in this way could presumably be used much more beneficially elsewhere.

References

- Alexander, N. (2018). My Pacemaker is tracking me from inside my body. The Atlantic, January 27, 2017, https://www.theatlantic.com/technology/archive/2018/01/my-pacemaker-is-tracking-me-from-inside-my-body/551681/, accessed 2018-08-19
- Baranchuk, A., Refaat, M. M., Patton, K. K., Chung, M. K., Krishnan, K., Kutyifa, V., ... Lakkireddy, D. R. (2018). Cybersecurity for cardiac implantable electronic devices: What should you know? Journal of the American College of Cardiology, 71(11), 1284–1288. https://doi.org/10.1016/j.jacc.2018.01.023
- Beauchamp, T. L., & Childress, J. F. (2009). Principles of biomedical ethics (6th ed.). New York: Oxford University Press.
- Burleson, W. P., & Carrara, S. (2014). Introduction. In W. P. Burleson, & S. Carrara (eds.), Security and privacy for implantable devices (pp. 1–11). New York: Springer.
- Burns, A. J., Johnson, M. E., & Honeyman, P. (2016). A brief chronology of medical device security. Communica-tions of the ACM, 59(10), 66–72. https://doi.org/10.1145/2890488
- Cerminara, K. L., & Uzdavines, M. (2017). Introduction to regulating innovation in healthcare: protecting the public or stifling progress?" Nova Law Review, 31(3), 305–312.
- Clouser, K. D., & Gert, B. (1990). A critique of principlism. Journal of Medicine and Philosophy, 15(2), 219–236. https://doi.org/10.1093/jmp/15.2.219
- Coventry, L., & Branley, D. (2018). Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. Maturitas, 113, 48–52. https://doi.org/10.1016/j.maturitas.2018.04.008
- FDA (2017). Firmware update to address cybersecurity vulnerabilities identified in Abbott's (formerly St. Jude Med-ical's) implantable cardiac pacemakers: FDA safety communication, https://www.fda.gov/MedicalDevices/Safety/AlertsandNotices/ucm573669.htm, accessed 2018-08-28
- Fu, K., & Blum, J. (2013). Controlling for cybersecurity risks of medical device software. Communications of the ACM, 56(10), 35–37. https://doi.org/10.1145/2508701

– Hine, K. (2011). What is the outcome of applying principlism? Theoretical Medicine and Bioethics, 32(6), 375–388. https://doi.org/10.1007/s11017-011-9185-x
– Loi, Michele, and Markus Christen. 2019. "Ethical Frameworks for Cybersecurity." In The Ethics of Cybersecurity, edited by Markus Christen, Bert Gordijn, and Michele Loi. Springer. https://canvas-project.eu/results/canvas-book.html.
– Loi, Michele, Markus Christen, Nadine Kleine, and Karsten Weber. 2019. "Cybersecurity in Health – Disentangling Value Tensions." Journal of Information, Communication and Ethics in Society, May. https://doi.org/10.1108/JICES-12-2018-0095.
– McGrath, P. (1998). Autonomy, Discourse, and Power: A Postmodern Reflection on Principlism and Bioethics. The Journal of Medicine and Philosophy, 23(5), 516–532. https://doi.org/10.1076/jmep.23.5.516.2568
– Mohan, A. (2014). Cyber decurity for personal medical devices Internet of Things. In 2014 IEEE International Conference on Distributed Computing in Sensor Systems (pp. 372–374). Marina Del Rey, CA, USA: IEEE. https://doi.org/10.1109/DCOSS.2014.49
– Pycroft, L., Boccard, S. G., Owen, S. L. F., Stein, J. F., Fitzgerald, J. J., Green, A. L., & Aziz, T. Z. (2016). Brainjacking: Implant security issues in invasive neuromodulation. World Neurosurgery, 92, 454–462. https://doi.org/10.1016/j.wneu.2016.05.010
– Radcliffe, J. (2011). Hacking medical devices for fun and insulin: Breaking the human SCADA system. White Paper. Black Hat Conference 2011, USA, https://media.blackhat.com/bh-us-
– 11/Radcliffe/BH_US_11_Radcliffe_Hacking_Medical_Devices_WP.pdf, accessed 2018-08-27
– Ransford, B., Clark, S. S., Kune, D.F., Fu, K., & Burleson, W. P. (2014). Design Challenges for Secure Implantable Medical Devices. In W. P. Burleson, & S. Carrara (eds.), Security and privacy for implantable devices (pp. 157–173). New York: Springer.
– Rios, B., & Butts, J. (2018). Understanding and Exploiting Implanted Medical Devices, https://www.blackhat.com/us-18/briefings.html#understanding-and-exploiting-implanted-medical-devices, accessed 2018-08-28.
– Sorell, T. (2011). The limits of principlism and recourse to theory: The example of telecare. Ethical Theory and Moral Practice, 14(4), 369–382. https://doi.org/10.1007/s10677-011-9292-9
– Vijayan, J. (2014). DHS investigates dozens of medical device cybersecurity flaws. Informationweek, October 23, 2014, http://www.informationweek.com/healthcare/security-and-privacy/dhs-investigates-dozens-ofmedical-device-cybersecurity-flaws-/d/d-id/1316882, accessed 2018-08-27
– Weber, Karsten, and Nadine Kleine. 2019. "Cybersecurity in Health Care." In The Ethics of Cybersecurity, edited by Markus Christen, Bert Gordijn, and Michele Loi. Springer. https://canvas-project.eu/results/canvas-book.html.
– Woods, Michael (2017). Cardiac defibrillators need to have a bulletproof vest: The national security risk posed by the lack of cybersecurity in implantable medical devices. Nova Law Review, 41(3): 419–447.

CASE STUDY 5

HACKING BACK

Question

In 2013, some hackers breached the control system of a dam near New York through a cellular modem and infiltrated the U.S. power grid system, gaining enough remote access to control the operations networks of the power system. The hackers targeted Calpine Corporation, a power producer with 82 plants operating in 18 states and Canada. Potentially they would have been able to shut down generating stations and cause blackouts, but their infiltration was discovered before they started damaging the power grid. The digital clues that were gathered pointed to Iranian hackers (Thompson 2016). In the same period, hackers linked to the Iranian government attacked American bank websites. These attacks were Iran's retaliation for Stuxnet.

In national security, is a country morally allowed to hack back another country that had previously hacked the critical infrastructure of the former country?

References for obtaining further information:

– Thompson M (24 Mar 2016) Iranian Cyber Attack on New York Dam Shows Future of War. Time. https://time.com/4270728/iran-cyber-attack-dam-fbi/.

Answer suggestion

1) Describe the moral question/problem

Enhancing passive cybersecurity measure can be extremely costly and, due to the complexity of protecting key infrastructures in an increasingly connected world, even the best feasible cybersecurity measures may leave the country exposed to threats (Viganò et al 2019). Thus, it may appear that the best defence strategy for a state which aims to protect its critical infrastructures is an active one: to build credible hacking back capabilities and profess a credible commitment to hack back, with destructive consequences, if one is hacked.

One the other hand, hacking back undermines the possibility of mutual trust in cyberspace, leading to a more dangerous world for most states and most individuals (Loi and Viganò 2019, Viganò et al 2019, Inversini 2019, Meyer 2019).

2) Gather the relevant facts

It can be said that the motivation to attack and respond to cyberattacks against national entities usually derives from the existing state of international anarchy (the absence of a world sovereign) and conflict among several states. In a cybersecurity arms race, all states end up worst off, either because they all end up more insecure, or because they all end up spending more money for their security.

Governments may exploit vulnerabilities in the computer and information systems of foreign countries, in order to perform intelligence activities or to damage the defences of other countries or individual opponents. Potential victims may try to avert these attacks, by increasing defences. But building successful defences becomes increasingly difficult. For example, it becomes increasingly difficult to protect critical infrastructure that relies on information and communication technology to function. One alternative to prevention and increased cyberdefences is hacking back.

Stuxnet was the virus targeting the Siemens software that operated the uranium enrichment facility in Iran, in which the attacked objects were the turbines themselves, not just the information in the system. In this case, the means of the attack, unlike the case involving drones, were merely informational (a piece of software), but the goal was to physically damage the turbines. The virus is widely believed to have been developed and deployed by the US in collaboration with Israel.

Some scholars have argued that the use of AI-enabled cyber weapons by states, for purposes of retaliation and deterrence, will lead to a cyber arms race from which all involved parties have to lose in terms of their national security (Taddeo and Floridi 2018, Meyer 2019). In such a situation, no country can afford to stay inactive due to the fear that other countries will gain an advantage that can be used against them.

The review of international legal and diplomatic initiatives concerning cyberwar reveals a failure of governmental actors to agree on more general principles of cyberspace behaviour, in spite of over twenty years of discussion (Meyer 2019). Thus, both hacking and hacking back makes relationships of trust among countries impossible. With mistrust as a baseline, each state rightly assumes that such an offensive strategy is favourable in contrast to purely defensive ones in order to gain the upper hand in foreign espionage or cyber-sabotage activities. Moreover,

a national government that is shown to be vulnerable to attacks can also appear less trustworthy towards other countries.

We can distinguish 4 fundamental categories of cyber-attacks to infrastructures (see table below). We may distinguish the means, which may be only software based or software and physical (e.g. hacking a self-driving car to kill civilians on the street). Stuxnet and the Iranian response to it are both instances in which the means of the attack are purely made of software. In terms of the effect of the attack, we can distinguish again mere functional disruption from physical damage. Some suggests that physical damage can be defined as damage such that "restoration of functionality requires replacement of physical components" (Schmitt 2013: 108, discussed in Viganò et al 2019). This can be contrasted with an attack having merely functional disruption as its damage, e.g. preventing the functioning of the internet for several hours. In terms of this distinction, Stuxnet and the Iranian response to it are attacks whose consequences, real or intended, were physical. Thus, both Stuxnet and the Iranian hacking back fall in the cell described in B1.

Table 5 Types of attacks on critical infrastructure

Damage →	1. Physical or physical-functional	2. Merely functional
Means of attack ↓		
A. Physical or cyber-physical	A1	A2
B. Merely cyber	B1	B2

3) Consider the relevant ethical principles (informed by the ethical theory of your choice)

The problem of hacking back will be examined from the point of view of three different moral principles:

1) Law of retaliation (tit for tat): do unto others as they have done to you.

2) Rule Utilitarianism: choose the moral rule that maximizes the net well-being produced by the consequences of all the actions following from that rule.

3) The Golden Rule: Do unto others as you would have them do unto you.

4) Make your decision/assessment

Law of retaliation (tit for tat): it may be argued that, irrespective of the consequences, a country that is attacked has the right to respond with a similar attack.

Even if the law of retaliation is compatible with the emergence of trust in small groups (Ostrom 2000), there are several difficulties with the law of retaliation, considered as a moral principle. We can point out to "internal" difficulties, e.g. internal contradictions or showing that it is hard to implement the norm, and "external" objections,

that is, objections from the standpoint of alternative moral frameworks.

Internal criticisms of the law of retaliation

First, consider a country that is justly attacked, e.g. a country who is engaging a genocide receives a massive cyber-attack against its physical infrastructure sufficient to stop the genocide without casualties. It seems absurd that this country has a moral right to retaliate against its attackers with a similar act.

Second, the behaviour of a country will be recognized as following the law of retaliation only if countries agree on how they classify cyberattacks (e.g. under which conditions a cyberattack count as acts of war). This is not trivial, as the recent history shows (see Meyer 2019, Viganò et al 2019, Schmitt 2013, Roscini 2017). In the absence of shared criteria, retaliatory acts based on law of retaliation will not be distinguishable from arbitrary ones.

Third, the traditional principles of just war, considering both jus ad bellum and jus in bello (Lazar 2017) do not justify all retaliatory acts. From the point of view of jus ad bellum,

retaliatory acts are just only if the following two criteria are satisfied (beside just cause and proportionality, that we assume here to be satisfied for the sake of the argument):

a) Just peace: the outcomes of the war must be tolerable if war is to be considered just.

b) Last resort: other more peaceful means (e.g. diplomatic ones) to achieve self-defence are not sufficiently likely to succeed.

There can be cases in which retaliatory cyber-attacks for a just cause (e.g. self-defence) are proportionate but not just, because they do not however lead to tolerable outcomes, and there are less harmful ways to achieve the military and strategic objectives the cyberwar is trying to achieve. Moreover, from the point of view of traditional jus in bello, a retaliatory act of cyberwar that kills civilian is never just, even in retaliation to a similar act.

It is interesting to observe that Iran tried to respond to Stuxnet (a merely software attack causing physical damage) with an act having the same characteristics as the original US one. This is maximally coherent with tit for tat. From the point of view of just war theory, and assuming for the sake of the argument that both the Stuxnet attack and the Iranian responses were acts of war (which is still highly controversial), Iran should have considered other options, e.g. whether it could have achieved comparable self-defence objectives, through merely functional attacks, without placing the lives of US non-combatants at risk and without raising a possible escalation of physical violence.

Rule utilitarianism:

Rule utilitarianism supports hacking back if hacking back maximizes net aggregate well-being of all the countries affected when all countries react to cyberattacks by hacking back. Assessing the net aggregate outcomes on human well-being of hacking back, adopted as a universal norm among states, is not easy. We will argue that it would lead to a race to the bottom with respect to trust, security, and/or higher costs for society to deliver adequate cyber protections to its citizens.

Against the view defended here, it may also be argued that hacking back produces good consequences when it achieves the level of mutually assured destruction, or some approximation. The doctrine of mutually assured destruction claims that international peace and stability will be reached when a full-scale use of offensive capabili-

ties by the offender and defender would cause the complete destruction of both the offender and the defender. If each offender is capable to annihilate the opponent and each opponent can launch an equally deadly second strike before being annihilated (before it gets hit, or after being hit with its surviving forces), no attacker can advance his self-interest by attacking. By analogy it may be argued that once a very high level of cyber damage capacity is reached by all (e.g. the explosion of nuclear power plants combined with a generalized blackout and the interruption of all health-care services), all parties will refrain from serious attacks.

This objection does not consider that, even if rational actors may become peaceful, this will not deter small, more anarchic actors, for example terrorist groups and individuals, for whom it is easier to escape the consequences of their acts. There is a high probability that more harmful cyberweapons will be produced in the arms race and fall in the hands of such groups. Even if international stability is achieved in this way, this comes at the price of more advanced cybersecurity protections. This peace diverts a high amount of societal resources into developing competitive cybersecurity weapons and cyber-defence capacities. In the absence of those threats, these resources could have been used to enhance human well-being.

The Golden Rule implies that Iran should not have retaliated with a similar attack, but sought justice through an international court, or the condemnation of the US by the international community. Yet, it is unclear that Iran and other countries in similar positions can realistically act on the basis of the Golden rule, without undermining their national security. The Golden Rule may be a norm applicable, realistically, by those countries with predominantly collaborative international relations.

For example, although the interests of countries such as France and Italy, or UK and Germany, are not identical, they are also not fully conflictual either. Inversini 2019 argues that countries have ethical and pragmatic reasons to move from a state of negative peace to a state of positive peace. A state of negative peace is a state in which peace is simply the status quo. Peace persist only because, and as long as, it is in the interest of each player to avoid waging war to others. The moment an agent sees an opportunity for gain by violating peace, negative peace will be perturbed. Inversini 2019 argues that the internet today is a state of negative peace. The scenario of stability through deterrence is also a state of negative peace, and a very bad one, due to the existence of very deadly cyberthreats. By contrast, positive peace consists in a state in which each player is motivated to build and preserve the conditions for peace of all, and this motivation prevails on the motivation to achieve strategic gains. It is a state in which all state can act by the Golden Rule without compromising their security, but instead while enhancing it.

In the context of cyberpeace, the Golden Rule ("do unto others as you would have them do unto you") amounts a duty to promote mutual trust and trustworthiness: "promote the conditions for international trust and preserve the international cyber-infrastructures of other trustworthy countries, in such a way that the trustworthy countries you trust have also reasons to trust you". This is a feasible strategy for countries that are already highly interdependent (e.g. in the economic sphere) and trust each other in sharing intelligence and infrastructure. In network of trust (Loi and Viganò 2019), an attack against the cyber-infrastructure of one country, becomes an attack against all countries. An improvement in the cyber-infrastructure of one becomes an improvement in the cybersecurity of all. If a country receives a cyberattack from a formerly trusted country, this attacker can be punished more severely by excluding it from the network of trust and its benefits, than by hacking it back. This is why hacking and hacking back is unlikely in a network of trust.

Thus, within a network of trust, this interpretation of the Golden Rule is also supported by rule-utilitarianism, because it promotes a kind of peace with a higher level of well-being than peace through mutually assured destruction. Rule-utilitarianism, in other words, promotes seeking the conditions of internet peace in which applying the Golden Rule is generally compatible with the national interest of all states involved.

References

- Lazar, Seth. 2017. "War." In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Spring 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2017/entries/war/.
- Loi, Michele and Viganò 2019, Eleonora "Achieving Trust in EU Cybersecurity", CANVAS Policy Brief No. 1. https://canvas-project.eu/assets/results/canvas_briefing-package-1.pdf
- Lucas, George. 2019. "Cybersecurity and Cyber Warfare: The Ethical Paradox of 'Universal Diffidence.'" In The Ethics of Cybersecurity, edited by Markus Christen, Bert Gordijn, and Michele Loi. Springer. https://canvas-project.eu/results/canvas-book.html.
- Meyer, Paul. 2019. "Norms of Responsible State Behaviour in Cyberspace." In The Ethics of Cybersecurity, edited by Markus Christen, Bert Gordijn, and Michele Loi. Springer. https://canvas-project.eu/results/canvas-book.html.
- Miller, Seumas. 2019. "Freedom of Political Communication, Propaganda and the Role of Epistemic Institutions in Cyberspace." In The Ethics of Cybersecurity, edited by Markus Christen, Bert Gordijn, and Michele Loi. Springer. https://canvas-project.eu/results/canvas-book.html.
- Ostrom, Elinor. 2000. "Collective Action and the Evolution of Social Norms." The Journal of Economic Perspectives 14 (3): 137–58.
- Roscini, Marco. 2017. "Military Objectives in Cyber Warfare." In Ethics and Policies for Cyber Operations: A NATO Cooperative Cyber Defence Centre of Excellence Initiative, edited by Mariarosaria Taddeo and Ludovica Glorioso, 99–114. Philosophical Studies Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-45300-2_7.
- Schmitt, Michael N. 2013. Tallinn Manual on the International Law Applicable to Cyber Warfare. Cambridge University Press.
- Taddeo, Mariarosaria, and Luciano Floridi. 2018. "Regulate Artificial Intelligence to Avert Cyber Arms Race." Nature 556 (7701): 296. https://doi.org/10.1038/d41586-018-04602-6.
- Viganò, Eleonora, Michele Loi, and Emad Yaghmaei. 2019. "Cybersecurity of Critical Infrastructure." In The Ethics of Cybersecurity, edited by Markus Christen, Gordijn Bert, and Michele Loi, https://canvas-project.eu/results/canvas-book.html. Springer.

CASE STUDY 6

STUXNET

Question

Stuxnet is the virus exploiting four undisclosed vulnerabilities (i.e. zero-days exploits) in Microsoft Windows that allegedly the governments of U.S. and Israel used to sabotage Iran's ability to enrich uranium in 2009-2010. This bit of malware is usually considered the first cyber weapon in the world that was used to damage a major infrastructure (Baylon 2017).

Was the Stuxnet attack morally legitimate?

References for obtaining further information:

– Farwell, James P., and Rafal Rohozinski. 2011. "Stuxnet and the Future of Cyber War". Survival. Global Politics and Strategy 53 (1): 23–40. https://doi.org/10.1080/00396338.2011.555586. Available in: https://www2.cs.duke.edu/courses/common/compsci092/papers/cyberwar/stuxnet2.pdf

Answer suggestion Case Analysis

1) Describe the moral question/problem

The problem at stake is assessing whether the Stuxnet attack was morally permissible to prevent Iran from developing nuclear weapons which would have been very dangerous for international peace and life on Earth.

2) Gather the relevant facts

Stuxnet is thought to have damaged 1000 centrifuges at Iranian nuclear facility in Natanz (Baylon 2017) and infected over 60,000 computers, more than half of them in Iran (Farwell and Rohozinski 2011), which was the target of the operation; but it also infected computers in the United States, the United Kingdom, Australia, Finland and Germany as collateral damage. Currently, the power of this worm is limited by antidotes and its built-in expiration date is 24 June 2012 (Farwell and Rohozinski 2011).

When the Stuxnet attack was discovered in June 2010, Iran counterattacked by means of similar cyber weapons: hackers linked to the Iranian government attacked American bank websites and tried to access the U.S. electric grid (Thompson 2016).

Stuxnet is a kind of attack that is functional (or cyber) and physical, as it gathered information and produced malfunctioning in the centrifuges but it also physically damaged them (Viganò, Loi, and Yaghmaei 2019).

3) Consider the relevant ethical principles (informed by the ethical theory of your choice)

We will analyze the ethical legitimacy of the Stuxnet attack from the perspective of utilitarianism. According to this family of moral theories, the morally right action is the action that produces the greatest good for the greatest number of persons (Bentham 1776). We thus will adopt an ethical perspective that is consequentialist: it assesses an action on the basis of the consequences that the action brings about.

On the one hand, the aim of the Stuxnet program was to prevent Iran from developing nuclear weapons which would have been very dangerous for international peace and in the long run represented a threat for the survival and well-being of humanity and the Earth.

In addition, compared to traditional military actions such as air strikes, the Stuxnet program did not kill or hurt anybody; it also did not involve the life of any soldier. Furthermore, although there are no data on the cost of the Stuxnet program, it was less expensive than traditional military actions and it was almost certainly less costly than the cost of a single fighter-bomber (Farwell and Rohozinski 2011).

On the other hand, first of all, Stuxnet was not completely effective in slowing down Iran's nuclear program. The number of centrifuges used to enrich uranium at the main Natanz plant declined of 23% from mid-2009 to mid-2010 but the production of low-enriched uranium has remained fairly constant (Markoff and Sanger 2010). Although the full extent of damage is not easy to assess, it was estimated that Stuxnet delayed the Iranian nuclear

program of two years (Baylon 2017), which is not a huge amount of time.

Secondly, there was suspicion from the U.S., Israel, and several other members of the international community but no actual proof that Iran was acquiring nuclear weapons; thus, the negative effects that Stuxnet prevented (the nuclear attack) could have not occurred in case Stuxnet had not been developed.

Thirdly, the Stuxnet attack caused collateral damage that needs to be taken into account: the damage to computers in many countries that were not involved in the Iran's nuclear program. The Siemens software is run in nuclear plants in several countries and according to the CEO of Kaspersky Lab, Eugene Kaspersky, a Russian nuclear plant was involved (Vincent 2013). It is likely that other nuclear facilities were infected but they did not disclose the attack in order to limit financial losses.

Fourthly, Stuxnet increased the risk of a cyber warfare: Iran's nuclear weapons production was slowed down but this was compensated by the acceleration of the cyber weapons program of Iran and of other countries, once Stuxnet was discovered (Baylon 2017). The threat of possible future cyber-attacks motivated the states to enhance their cyber capabilities. This aim could be achieved by directly enhancing a state's defense, but also by enhancing surveillance and retaliatory capabilities, which bring about respectively a limitation of citizens' privacy and a threat for long-term cyber peace (Viganò et al. 2019). Considering the increasingly effectiveness of cyber-attacks to critical infrastructures, that which Stuxnet allegedly prevented (a nuclear attack) could be triggered by another cyber-attack in a cyber arms race.

Finally, once the worm was discovered, information on its code was published, and the unintended consequence was that cybercriminals copied some of the techniques of Stuxnet to commit online theft (Baylon 2017). Similarly, the publishing of the Stuxnet code may inspire terrorist groups to use similar weapons (Baylon 2017) either by purchasing vulnerabilities in the black market or by hiring hackers.

4) Make your decision/assessment

From a utilitarian perspective, the aim of the Stuxnet attack was good but it is not sure if Iran was acquiring nuclear weapons, thus the benefit brought about by that cyber-attack should be discounted by the probability that Iran was developing nuclear weapons and intended to use them. Furthermore, even though Stuxnet, as a cyber attack, did not hurt or kill anybody, it brought about several negative consequences: damages to other nuclear plants and computer systems that were not involved in the Iranian nuclear program, retaliation with similar cyber means by Iran, a cyber arms race in which countries try to exploit software vulnerabilities, and the repurposing of some of Stuxnet's characteristics in malware used for online theft. Stuxnet also increased the risk that other negative effects will occur in the future: the interruption of cyber peace due to the cyber arms race and the adoption of malware similar to Stuxnet by terrorist groups.

Therefore, utilitarian theories would have not recommended the Stuxnet attack, as its positive effect, which was huge but discounted by the probability that Iran was developing and going to use nuclear weapons, was compensated by its many negative effects and the increase in the risk that further negative effects will take place in the future.

References

- Baylon, Caroline. 2017. “Lessons from Stuxnet and the Realm of Cyber and Nuclear Security: Implications for Ethics in Cyber Warfare.” In Ethics and Policies for Cyber Operations, 213–29. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-45300-2_12.
- Bentham, Jeremy. 1977 [1776]. “A Comment on the Commentaries and A Fragment on Government”. In The Collected Works of Jeremy Bentham, edited by J. H. Burns and L. A. Hart. London: Oxford University Press.
- Farwell, James P., and Rafal Rohozinski. 2011. “Stuxnet and the Future of Cyber War”. Survival. Global Politics and Strategy 53 (1): 23–40. https://doi.org/10.1080/00396338.2011.555586.
- Kushner, David. 2013. “The Real Story of Stuxnet.” IEEE Spectrum. 2013. https://spectrum.ieee.org/telecom/security/the-real-story-of-stuxnet.
- Markoff, John, and David E. Sanger. 2010. “In a Computer Worm, a Possible Biblical Clue”. The New York Times. 2010. https://www.nytimes.com/2010/09/30/world/middleeast/30worm.html?mtrref=undefined&gwh=6DDD1CF7878E12CEC51B30BAD0D9487C&gwt=pay.
- Thompson, Mark. 2016. “Iranian Cyber Attack on New York Dam Reveals Future of War”. Time. 2016. https://time.com/4270728/iran-cyber-attack-dam-fbi/.
- Viganò, Eleonora, Michele Loi, and Emad Yaghmaei. 2019. “Cybersecurity of Critical Infrastructure”. In The Ethics of Cybersecurity, edited by Markus Christen, Gordijn Bert, and Michele Loi. Springer.
- Vincent, James. 2013. “Russian nuclear power plant infected by stuxnet malware says cyber-security expert”. Indepedendent. https://www.independent.co.uk/life-style/gadgets-and-tech/news/russian-nuclear-power-plant-infected-by-stuxnet-malware-says-cyber-security-expert-8935529.html.

CASE STUDY 7

SCHWARTZ' HACKING

Question

Ethical hackers are computer security professionals that are hired by companies and governments to protect the security of their computer systems. Ethical hackers employ the same tools and techniques as hackers to break into computer systems, but with a different aim from that of hackers. They break into computer systems to evaluate the target system's security, report back its vulnerabilities, and find remedies to vulnerabilities (Palmer 2001).

Is it morally permissible, as security consultant, to hack without authorizationthe computer system of one's own company in order to improve its security?

References for obtaining further information:

– Palmer, C. C. 2001. "Ethical Hacking." IBM Systems Journal 40 (3): 769–80. https://doi.org/10.1147/sj.403.0769. Also available at https://thehacktoday.com/wp-content/uploads/2016/01/palmer.pdf

Answer suggestion

1) Describe the moral question/problem

Ethical hackers employ the same tools and techniques as hackers to break into computer systems. In the question examined, though, the security consultant did not get authorization to hack his company's computer system, which is a condition that is required in ethical hackers' ethical codes.

2) Gather the relevant facts

Randal Schwartz used to work as a security consultant for Intel Corporation, at the Supercomputer Systems Division (SSD). In 1992, he was hired as a system administrator for Hawthorn Farms, which is an Intel division, but he kept an account on the SSD network. He inadvertently discovered that a fellow employee had a weak password. Without informing Intel's staff administrators, he decided to run Crack – a cracking program on the password database file, which is also used by corporation spies and hackers –, because he suspected that the security team at Intel was not running a password checking program (Lewis 1995). He discovered several additional weak passwords and he also moved a "passwd" file from the SSD cluster to a faster machine on the Intel network, in order to run Crack against it and check its vulnerabilities. The Intel security team detected Schwartz' activity. As he had already been reprimanded for two previous unauthorized incursions into computers at Intel, Intel officials decided to terminate and prosecute Mr. Schwartz. In the two previous occasions, he had installed a program bypassing Intel's firewall in order to access his Intel e-mail account when he was away from the site. He did this in order to facilitate his job and he did not perform malicious actions.

In 1995 Mr. Schwartz was convicted of three felony counts of computer hacking under the Oregon computer crime statute 164.377, which states that altering a computer system without authorization and accessing to a system with the intention of committing theft are felonies (Lewis 1995). He was sentenced to probation for five years, and a 480-hour community service order. He was also ordered to pay Intel a fine of $68,000, as well as being obliged to stump up $170,000 in legal fees. After an appeal, the restitution order was dropped in 2001 but the court declined to quash the conviction, which was sent back to the lower court to be re-examined. In 2007 The Oregon court ordered an expungement of his conviction (Espiner 2007).

3) Consider the relevant ethical principles (informed by the ethical theory of your choice)

Mr. Schwartz case will be analysed from two moral perspectives: one is the principle of non-maleficence, the other is a deontological ethics based on the four values that are the most prominent in professional ethics of ethical hacking.

Firstly, the main ethical consideration involved in Mr. Schwartz' case is the assessment of harm inflicted to Intel by Mr. Schwartz. The prohibition to cause harm, also known as the principle of non-maleficence, is acknowledged in many ethical traditions. It is one of the four principles that constitute Principlism, which is one of the most widespread approaches to bioethics (Beauchamp & Childress 1979). The four principles (respect for autonomy, beneficence, non-maleficence, and justice) are considered part of our common morality. Non-maleficence basically

requires not to intentionally cause harm to somebody in acts of commission or omission (Beauchamp 2019). Mr. Schwartz did not cause any harm to his company: he did not access confidential information, neither did he use such information to his own advantage. Intel also accused Mr. Schwartz of theft, as it affirmed that his transmission of the „passwd" file from one computer to another was a form of theft. But Mr. Schwartz had not taken the passwords out of the system, thus, he did not damage the company.

Secondly, the relationship between Mr. Schwartz and Intel is not only regulated by his job contract but also by professional ethics, i.e. the personal and corporate standards of behavior that are expected in a profession and comprise duties, responsibilities, and limits of a job. Yet, as ethical hacking is a relatively new profession, there are no uniformed professional ethics, neither uniformed ethical codes for such profession. For this reason, Mr. Schwartz' behavior is not easy to assess: there are no univocal duties, responsibilities, and limits of an ethical hacker. However, we can infer some of these standards from the tasks and aims of the ethical hacker. The latter is hired to evaluate the security of a client's system or network. Consequently, the ethical hacker inevitably handles sensitive and confidential information, and this poses the company in a position of vulnerability. For this reason, at least four ethical values are usually present in the ethical codes of ethical hackers (Johansen 2017; Jaquet-Chiffelle & Loi 2019): honesty, transparency, respect of privacy, and trustworthiness. Honesty requires the ethical hackers to stay within the scope of his/her client's expectations (Jaquet-Chiffelle & Loi 2019), without taking advantage of the vulnerabilities of the company's IT-system or of the information accessed. Transparency between the client and the ethical hacker is fundamental, as the client has the right to know what the ethical hacker is carrying out. Respect of the client's and its employees' privacy (Jaquet-Chiffelle & Loi 2019) is also important, as some information of the company is confidential and need to remain as such. The fulfillment of these values grounds a relationship of trust between the client and the ethical hacker. Trustworthiness is the quality descending from these values. Ethical hackers possess similar expertise to those of hackers, programming and computer networking skills, continuous education to keep up with innovation in computer science, and also patience in monitoring systems for weaknesses. The only thing that differentiates an ethical hacker from a hacker is trustworthiness (Palmer 2001). An ethical hacker should be trustworthy because the company or government hiring him/her basically gives him/her sensitive information that is secret. Mr. Schwartz did not respect transparency when he cracked the passwords of the SSD, as he did not inform Intel of his cracking activities. In the eyes of Intel's officers, he probably lost trustworthiness after his first violations of Intel's firewall, which weakened the trust relationship with Intel. It is true that Mr. Schwartz did not disclose to third parties the password he found, neither he used them for his own advantage. He respected the privacy of Intel, as sensitive and confidential information was not leaked or used for personal reasons.

With regards to honesty, i.e. the requirement to stay within the scope of the client's expectations, Mr. Schwartz said he always believed that actively trying to break into systems was a standard way of checking security (Lewis 1995). The job of penetration testing is not easily defined but there is agreement that the tasks of a system administrator include protecting the security of a system against hackers, vandals, and spies. To make sure that a system is secure includes checking its vulnerabilities. In fact, as we have seen, ethical hackers use the same tools as hackers. In the case of Mr. Schwartz, was cracking the passwords of the SSD's IT-infrastructure part of Mr. Schwartz' job? On the one hand, Mr. Schwartz was moved to another division of Intel and thus he was no more responsible for the security of SSD. Also, he did not ask permission to stress Intel's IT-security. It has been said that ethical hackers hold the keys to a company (Palmer 2001), as they can easily have access beyond the target areas of the system or network that the client granted in the work agreement. Remaining within the specified

target areas is thus a duty of the ethical hacker and Mr. Schwartz did not respect this duty, as he accessed a network of a division he did not belong to. On the other hand, the definition of target areas is not easy in the case of security assessment, since a compromised network can be the basis for attacking another network (Lewis 1995).

4) Make your decision/assessment

From the perspective of the principle of non-maleficence, Mr. Schwartz did not cause any harm to Intel neither to the society. Thus, from this perspective, he is not blameworthy. However, his conduct is not only assessed from the perspective of applied ethics but also from that of professional ethics, as Mr. Schwartz' conduct pertains to his job as an ethical hacker. From this point of view, we can see that Mr. Schwartz violated at least two of the four basic values of his professional ethics: he was not transparent, as he did not inform Intel of his cracking activities, and, accordingly, he was less trustworthy and weakened his relationship of trust with Intel. Hence, he was blameworthy of lack of transparency and limited trustworthiness. Finally, his honesty to remain within the scope of his company expectation cannot be assessed as the scope of the company's expectation is not easy to define: as a compromised network can be the basis for attacking another network, Mr. Schwartz' target areas for security inspection could have extended to the SSD.

References

- Beauchamp, Tom L., and James F. Childress. 1979. Principles of Biomedical Ethics. New York: Oxford University Press.
- Beauchamp, Tom. 2019. "The Principle of Beneficence in Applied Ethics." In Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta. Stanford University.
- Beauchamp, Tom L., and James F. Childress. 1979. Principles of Biomedical Ethics. New York: Oxford University Press.
- Espiner, Tom. 2007. "Intel 'hacker' Sentence Expunged." CNET. 2007. https://www.cnet.com/news/intel-hacker-sentence-expunged/.
- Jaquet-Chiffelle, David-Olivier, and Michele Loi. 2019. "Ethical and Unethical Hacking." In The Ethics of Cybersecurity, edited by Markus Christen, Bert Gordijn, and Michele Loi. Springer. https://canvas-project.eu/results/canvas-book.html.
- Johansen, Rowena. 2017. "Ethical Hacking Code of Ethics: Security, Risk & Issues." Panmore Institute. 2017. http://panmore.com/ethical-hacking-code-of-ethics-security-risk-issues.
- Lewis, Peter H. 1995. "Technology: On the Net; An Intel Computer Security Expert Runs Afoul of the Law. So Much for the 'Hacker Ethic'? - The New York Times." The New York Times, November 27, 1995.
- Loi, Michele, and Markus Christen. 2019. "Ethical Frameworks for Cybersecurity." In The Ethics of Cybersecurity, edited by Markus Christen, Bert Gordijn, and Michele Loi. Springer. https://canvas-project.eu/results/canvas-book.html.
- Palmer, C. C. 2001. "Ethical Hacking." IBM Systems Journal 40 (3): 769–80. https://doi.org/10.1147/sj.403.0769.

CASE STUDY 8

ANONYMOUS HACK ISIS

Question

In 2015, after the attack on Charlie Hebdo, the famous hacker group Anonymous actually declared war to al-Qaeda, ISIS and other terrorist groups (Douglas 2015, Martin 2015); Anonymous shut down some of their accounts on social networks. This ethical hacking was criticized, though, because taking down extremists' websites and other parts of their communication infrastructure would make them harder to monitor.

Was Anonymous' cyberattack on the communication infrastructure ethically justified?

References for obtaining further information:

- Douglas, E., 2015. Anonymous declares war on al Queda, Islamic State for Charlie Hebdo attack. The Washington Times, 2015-01-09. https://www.washingtontimes.com/news/2015/jan/9/charlie-hebdo-attack-prompts-anonymous-declare-war/, last visited 2019-06-24.
- Martin, A., 2015. Anonymous hackers 'declare war' on al Qaeda and Islamic State. We Live Securiry, 2015-01-12. https://www.welivesecurity.com/2015/01/12/anonymous-hackers-declare-war-al-quaeda-islamic-state/, last visited 2019-06-24.

Answer suggestion

1) Describe the moral question/problem

The question at stake regards whether Anonymous was morally justified in using (cyber-)force to crack down on terrorist groups. Furthermore, one may ask whether there were different ways Anonymous could have chosen to act. A, probably incomplete, list of possible courses of action is:

- No action at all.
- Attack ISIS' communication infrastructure without public announcement.
- Attack ISIS' communication infrastructure with public announcement.
- Supporting national security authorities in attacks against ISIS' communication infrastructure.

In fact, Anonymous chose the second option to attack ISIS' communication infrastructure with public announcement. The question to be addressed is whether from an ethical point of view this was indeed the best strategy. This question can be tackled, for instance, from a utilitarian point of view. In light of this ethical perspective the question can thus be reformulated as: Does attacking ISIS' communication infrastructure with public announcement contribute more to the common good than other strategies. However, other ethical perspectives can also be adopted: One can, for example, ask whether an attack on ISIS' communication infrastructure by a non-state actor is covered by the theory of just war (cf. Colarik & Ball 2016).

2) Gather the relevant facts

Many of the relevant facts have already been reviewed in the description of the case study. This section concentrates on the expected impact of the different release strategies.

- If the first option of no action at all is chosen, this means on the one hand that there would be no attacks on ISIS, at least not by Anonymous. ISIS could therefore continue to use its infrastructure undisturbed and thus plan and support terrorist attacks. But this would also fail to provide the publicly visible sign of rejection of terror, which can be understood as an expression of civil resistance. On the other hand, in this way no members of Anonymous put themselves in danger, no one breaks the state's monopoly on the use of force, and there is no escalation of violence, at least on the part of non-state actors of the attacked states.
- An attack on ISIS's communications infrastructure without public announcement means that ISIS may be hampered in its terrorist actions, potentially protecting human lives and resources. However, the public would not know. Therefore, Anonymous could not boast of having actively acted against terrorism. The public, on the other hand, would not know that such actions have taken place. This would not achieve an important effect of such attacks, because ISIS would not be deterred and people in the terrorized states would not know that there are measures that can hinder terrorist actions.
- A publicly announced attack on ISIS's communications infrastructure also means that ISIS may be obstructed in its terrorist actions, potentially protecting human lives and resources. But now the members of Anonymous are exposed to some risk of retaliation. However, such action sends an important signal to the people of those countries suffering from the terror of ISIS, who now know that they are not alone and may get the impression that there are organizations helping them. Such attacks may discourage ISIS from becoming overly visible on the Internet and in social media, reducing the possibilities for supporting terrorist attacks.

– In the case that Anonymous would support national security authorities in attacks against ISIS' communication infrastructure it is to be expected that the public would not know any-thing about such an attack. Therefore, most of the comments already made on the second option apply here. In addition, however, Anonymous' co-operation with government agencies could reduce Anonymous' credibility once information about it becomes public. In addition, it is questionable whether all Anonymous members would be willing to do so.

3) Consider the relevant ethical principles (informed by the ethical theory of your choice)

In what follows, the case is analysed from a utilitarian point of view. This means that the focus is on the expected ratio of benefits and harms of the different scenarios respectively. However, it must be stressed that an ethical analysis, for example from a deontological perspective or from a theory of just war, could produce significantly different results.

The principle used: When choosing between alternative actions it is ethically obligatory to choose that action that has will yield the best expected ratio of benefits and harms.

One problem with applying this principle, however, is that it is very difficult to determine costs and benefits. In addition, risks to material goods are difficult to compare with risks to people; one could even say that they are incommensurable. Last but not least, the question arises as to how far into the future costs and benefits can and should be estimated and taken into account.

The potential benefits are: (a) ISIS is discouraged from using the Internet and social media for its purposes. (b) The number of terrorist attacks is dropping. (c) The public is informed. (d) The public is reassured. Presumably there are further advantages, but in the following it will be argued on the basis of the mentioned benefits.

The potential harms are: (e) In the future, ISIS will use hidden communication channels and will be more difficult to monitor. (f) ISIS escalates its terrorist attacks. (g) ISIS attacks Anonymous and its members. (h) The public panics. Again, presumably there are further harms or costs, but in the following it will be argued on the basis of the mentioned harms.

– If the first option of no action at all is chosen, none of the above advantages will be achieved. But it is also very likely that none of the harms will take place. In fact, nothing will change but beyond those you are already endangered by terrorist attacks no other stake-holders like members of Anonymous will be at risk. One could argue that the sum total of benefits and harms is zero.
– An attack on ISIS's communications infrastructure without public announcement could help bring about benefits (a) and maybe (b) but not (c) and (d) but at the same that action could bring about harms (e) and (f). It is therefore an empirical question that one would first have to answer in order to decide whether this option would be morally prefera-ble. In addition, the causal relationship between the action and the consequences would certainly have to be established. The advantage of this option is that ISIS does not know who attacked, so retaliation would be difficult: Harm (g) will therefore not occur, nor (h). In principle, however, it is difficult to decide whether the sum total of benefits and harms is positive or negative.

– An attack on ISIS's communications infrastructure with public announcement could help bring about benefits (a) and maybe (b) as well as surely (c) and probably (d). Simul-taneously that action could bring about harms (e) to (h). In order to evaluate this option, it would need to be much better known how serious the risk of retaliation against Anonymous and its members is, and how this risk compares to the potential benefits of informing and reassuring the public. Basically, the welfare of a few is competing with the welfare of many.
– If Anonymous would support national security authorities in attacks against ISIS' communication infrastructure that could help bring about benefits (a) and maybe (b) but probably not (c) and (d). Furthermore, it is it be expected that this action could bring about harms (e) and (f). Therefore, a moral assessment resembles that of the second option, but potentially questions the credibility and cohesion of Anonymous.

It is to be expected that also the application of other ethical theories will result in rather ambig-uous statements, since for clarification much more information would have to be known, e.g. about the long-term consequences of the ISIS hack, about hierarchies of obligations or the prioritization of moral principles. The ISIS hack could therefore be one of those ‚hard cases' that do not allow a straight moral answer.

4) Make your decision/assessment

There is not only an answer to the question of what course of action can be justified from an ethical perspective. Probably a utilitarian answer would be that an attack on ISIS's communications infrastructure with simultaneous media coverage could best promote the common good – the option actually chosen. In that case, however, it must be assumed that the danger to the members of Anonymous is outweighed by the benefits generated, but this can be as-sessed differently. In addition, one could argue that ISIS can learn from these attacks how they work and how it can either use such methods itself or better defend itself against them in the future. It could also be argued that in the long run violating the state's monopoly on the use of force produces more harm than good. The further possible consequences lie in the future, the more unclear a utilitarian calculation becomes. Moral theories of just war or the right to self-defence would probably provide similarly ambivalent answers, for questions about the proportionality of measures taken would have to be answered. These, too, are ultimately based on cost-benefit considerations and bring with them similar problems with regard to the consideration of future contingencies.

References

– Colarik, A., Ball, R., 2016. Anonymous versus ISIS: The role of non-state actors in self-defense. Global Security and Intelligence Studies 2(1), 20–32. https://doi.org/10.18278/gsis.2.1.3.
– Jaquet-Chiffelle, David-Olivier, and Michele Loi. 2019. "Ethical and Unethical Hacking." In The Ethics of Cybersecurity, edited by Markus Christen, Bert Gordijn, and Michele Loi. Springer. https://canvas-project.eu/results/canvas-book.html.
– Gerstel, D., 2016. ISIS and innovative propaganda: Confronting extremism in the digital age. Swarthmore International Relations Journal 1(1), 1–9. https://doi.org/10.24968/2574-0113.1.5. 1
– Martins, R., 2017. Anonymous' cyberwar against ISIS and the asymmetrical nature of cyber conflicts. The Cyber Defense Review 2(3), 95–106. https://cyberdefensereview.army.mil/CDR-Content/Articles/Article-View/Article/1588748/anonymous-cyberwar-against-isis-and-the-asymmetrical-nature-of-cyber-conflicts/, last visited 2019-06-24.
– Parkin, S., 2016. Operation Troll ISIS: Inside Anonymous' war to take down Daesh. Wired, 2016-10-06. https://www.wired.co.uk/article/anonymous-war-to-undermine-daesh, last visited 2019-06-24.

CASE STUDY 9

SELLING VULNERABILITIES

Question

A zero-day vulnerability is a flaw that is found in software programs or operating systems and that does not have a patch or update to fix it. Such vulnerabilities are difficult to find and to exploit and they enable accessing computers and networks, stealing data, altering the functioning of processes. Accordingly, they are costly.

Is morally justified to sell vulnerabilities on the grey and black markets?

Answer suggestion

1) Describe the moral question/problem

The zero-day exploits can be used as a form of weapon, as they can disrupt and destroy computers and their network. For this reason, government intelligence, law enforcement agencies, and terrorist groups are very interested in zero-day exploits and willing to pay large amounts of money. As a consequence, the grey and the black markets of zero-day vulnerabilities are flourishing. Selling these products is very lucrative but at the same time very controversial from an ethical perspective.

2) Gather the relevant facts

In what follows, a realistic case of zero-day sale is presented. The case is invented but based on real data regarding what we know about selling zero-day exploits on the grey and black markets, their prices, and the ideas of the hackers running zero-day companies that were interviewed (Perlroth and Sanger 2013; Greenberg 2012, 2015).

After months of work on spotting vulnerabilities on software Beta, Alan – a skillful security researcher – finds a flaw in the system that enables to gather data from the computers using Beta. He discovers that a vulnerability similar to the one that he spotted could be sold on the grey market for 30,000 € to an information security company that acquires zero-day exploits from security researchers and sells them only to customers from NATO governments. Then he finds an advertisement from the Beta vendor saying that the latter has launched a bug bounty program, which consists in paying hackers for revealing to the company the bugs in Beta, to prevent that they sell these bugs on the black market or keep them to themselves. A vulnerability such as Alan's is rewarded 3,000 € by the company. Alan wants to see all the options that he has for profiting from its discovery and thus goes on the black market and monitors the prices of vulnerabilities of Beta. He finds that the latter would be paid around 150,000 €, but in the black market he would not know the identity of the customer, so he is aware that the flaw in Beta that he found could be used for malicious use. He reconsiders the grey market, in which his vulnerability sale would be quite remunerative, albeit not as much as in the black market, and would not risk being sold to a rogue state. Alan decides thus to sell the vulnerability to the information security company in the grey market.

3) Consider the relevant ethical principles (informed by the ethical theory of your choice)

We will analyze the three options of Alan's decision.

In business ethics, it is immoral to not reward or to reward not enough the job of an individual. Zero-day researchers contend that bug bounty programs are too low to appropriately pay for the researcher's time and effort. Hence, some of them do not consider the white market a fair

place where to exchange their discoveries (Perlroth and Sanger 2013). Alan worked several months on spotting flaws in Beta and thinks that he deserves to be paid for its work and that the bug bounty offered by the Beta company is not enough: it does not reward the time he invested in hacking the software. Although it is difficult to compute the economic value of the time and effort of zero-day researchers, from that perspective, black and grey

markets reward researchers in a fairer way than the white market.

From a consequentialist perspective, i.e. a perspective assessing the morality of actions on the basis of their consequences, trading in the black market risks injecting money into dubious circles, as hackers intersect with cybercrime circles, which in turn intersect with organized crime, drug cartels, and terrorism (Egelman, Herley, and van Oorschot 2013). Furthermore, the main participants of grey and black markets are government agencies, which use zero-day exploits for national security, reconnaissance, and cyber-attacks against other countries. This means that selling zero-day exploits in the black and grey markets contributes to the global cyber arms race. The fact that the grey market usually allows for more ethical exchanges than the black market (e.g. it is possible to know the identity of the customer and to refuse to sell products to it) does not exclude grey markets from the cyber arms race because even governments respecting human right and international agreements are engaged in this arms race and incentivize other countries to do the same. Zero-day sales in the grey and black market also contribute to "a related ethical concern: the danger that a country might launch a cyber weapon by mistake—for instance, by accidentally triggering a logic bomb it has planted in another country's systems" (Baylon 2017).

Finally, as the Beta vulnerability enables one to gather data from other people's computer, it is suitable for activities of espionage between countries, by terrorist groups, and for spying the activity of a country's citizens. Thus, it is possible that on the black market an authoritarian country will buy Alan's vulnerability and use it to spy on its political opponents with the aim of neutralizing their activity and imprisoning them. Is it also likely that the Beta vulnerability will be employed as a part of a code to build a virus stealing data from critical infrastructure, as it happened in the Stuxnet case, in which the virus stole the data of the Natanz nuclear enrichment plant in Iran in 2010 (Kushner 2013). Selling the vulnerability in the black market would reward Alan's effort but at the same time he worries about the possibility that the vulnerability would be used to imprison innocent people or destroy critical infrastructures of a country. Therefore, from a consequentialist standpoint, Alan should not sell the Beta vulnerability on the black market and also on the grey one.

Another fundamental ethical issue in selling vulnerabilities is that the business model of the companies selling vulnerabilities and exploits to private buyers is unethical because it is based on not revealing vulnerabilities to vendors, which decreases the security of products (Baylon 2017). Thus, according to this principle, Alan should reveal his discovery to the Beta vendor and stay in the white market.

4) Make your decision/assessment

In the end, was Alan right in selling the Beta vulnerability on the grey market? We considered three ethical principles regarding selling vulnerabilities. One is in favor of it (fairly rewarding one's own work) and the other two are against (hiding vulnerabilities to the vendors makes products less safe and has huge negative effects, i.e. a cyber arms race, supporting morally dubious circles and rogue states). The two considerations against selling vulnerabilities all apply to the black market. For this reason, selling vulnerabilities in the black market is ethically wrong. Selling vulnerabilities in the grey market is more ethical as it is supported by the principle of fairly rewarding work but at the same time it is potentially very risky in terms of cyber peace and safety of products. Selling the vulnerabilities on the regular market through bug bounty programs avoids the two principles against that action, but it is not supported by the principle of rewarding work fairly. While the immorality of selling vulnerabilities on the black market is not controversial, the two remaining options are more controversial: bug bounties do not reward researchers in a fair way and the grey market may cause huge harm due to the consequences of a cyber arms

race or funding hackers that are cybercriminals. The final decision depends on the relative importance that we give to bad consequences and fair retribution. Considering the severity of the consequences of a cyber war and of supporting cybercriminal actions, it would be recommended to sell the vulnerability to the Beta vendor.

References

- Baylon, Caroline. 2017. “Lessons from Stuxnet and the Realm of Cyber and Nuclear Security: Implications for Ethics in Cyber Warfare.” In Ethics and Policies for Cyber Operations, 213–29. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-45300-2_12.
- Egelman, Serge, Cormac Herley, and Paul C. van Oorschot. 2013. “Markets for Zero-Day Exploits.” In Proceedings of the 2013 Workshop on New Security Paradigms Workshop - NSPW ’13. https://doi.org/10.1145/2535813.2535818.
- Greenberg, Andy. 2012. “Shopping For Zero-Days: A Price List For Hackers’ Secret Software Exploits.” Forbes. 2012. https://www.forbes.com/sites/andygreenberg/2012/03/23/shopping-for-zero-days-an-price-list-for-hackers-secret-software-exploits/#15e23acd2660.
- ———. 2015. “New Dark-Web Market Is Selling Zero-Day Exploits to Hackers.” WIRED. 2015. https://www.wired.com/2015/04/therealdeal-zero-day-exploits/.
- Kushner, David. 2013. “The Real Story of Stuxnet.” IEEE Spectrum. 2013. https://spectrum.ieee.org/telecom/security/the-real-story-of-stuxnet.
- Perlroth, Nicole, and David E. Sanger. 2013. “Nations Buying as Hackers Sell Flaws in Computer Code - The New York Times.” 2013. https://www.nytimes.com/2013/07/14/world/europe/nations-buying-as-hackers-sell-computer-flaws.html.
- Viganò, Eleonora, Michele Loi, and Emad Yaghmaei. 2019. “Cybersecurity of Critical Infrastructure.” In The Ethics of Cybersecurity, edited by Markus Christen, Gordijn Bert, and Michele Loi. Springer.