SOLVING MORAL DILEMMAS WITH AI TO ADDRESS THE SOCIAL IMPLICATIONS OF THE COVID-19 CRISIS

Hubert Etienne[*]
Facebook AI Research
Ecole Normale Supérieure
Faculty of Philosophy
{hae}@fb.com

ABSTRACT

The Covid-19 crisis exposed the world to multidimensional challenges, which not only include the management of the pandemic by healthcare authorities, but also the moral dilemmas faced by practitioners in medical structures, the preservation of trust in information undermined by the spread of false news, the upkeep of human interactions during lockdown periods, as well as the dangers of online hate rising with polarizing contexts. In such circumstances and when combined with an appropriate level of human judgement, machine learning applications were revealed to be crucial resources in supporting decision-making and implementation, resulting in more efficient and better-informed responses to said issues. This paper focusses on four social dimensions (bioethical, political, psychological and economic) from which the decisions taken in the context of the Covid-19 crisis derived major ethical implications. On the one hand, I argue against the possibility of addressing these issues from a purely algorithmic approach, elaborating on two types of limitations for automated systems to address ethical issues. This leads me to discuss how different ethical situations call for different performance metrics with regards to the 'contextual explicability and performance issue', as well as to enunciate a gold principle: 'legitimacy trumps accuracy'. On the other hand, I present practical examples of machine learning applications which enhance, instead of dilute, human moral agency in better addressing these issues. I also suggest a 'moral perimeter' framework to ensure the responsibility of algorithms-assisted decision-makers for critical decisions. The unique potential of AI to 'solve' moral dilemmas by intervening on their conditions of possibility prompts me to discuss a new type of moral situation: AI-generated meta-dilemmas.

1 INTRODUCTION

The challenges of Covid-19 provoked an unprecedented mobilisation in the research community, resulting in the publishing of almost 24,000 research documents on the SARS-CoV-2 coronavirus and the diseases it causes between January 1st and June 30th 2020 (da Silva et al., 2021). Researchers in artificial intelligence (AI) also took part in this general effort, notably developing predictive models to map the spread of the outbreak and contact tracking applications to support governments and organizations in the management of the pandemic. However, the people whose lives were taken by the virus are not the only victims of the Covid-19 crisis. Quarantining a third of the world's population (Buchholz, 2020) led to a significant increase in online hate, and strongly impacted individuals' mental health with regards to loneliness and depression. Online misinformation tore families apart, physical violence surged globally, and healthcare professionals faced moral dilemmas they were not prepared for. The goal of this paper is to exhibit how machine learning solutions can be, and have been, leveraged, not only to address the sanitary aspects of the outbreak, but also to tackle the social dimensions of the Covid-19 crisis, ultimately leading to greater responsibility. Firstly, I identify four key domains where major decisions were taken in the context of the Covid-19 crisis and elaborate

[*] PhD resident at Facebook AI Research, Ecole Normale Supérieure, Faculty of Philosophy, and Sorbonne University, Faculty of Science and Engineering.

on their underlying ethical implications. I then present two types of inherent limitations of artificial agents, disqualifying purely algorithmic approaches to address these issues as much as justifying the need to find a coherent collaboration framework between artificial and moral agents. Thirdly, I present practical examples of such collaboration as enhancing moral agency for each of the four dimensions' issues. Finally, I suggest a moral scheme to avoid involuntary delegation of autonomy in critical decisions, before concluding on the unique power AI offers to solve moral dilemmas, as well as the unprecedented degree of responsibility it comes with.

2 Four dimensions of the Covid-19 crisis with social implications

Covid-19 as an infectious disease calling for a sanitary response to a pandemic should be distinguished from the crisis of Covid-19, which refers to a wider phenomenon, captured by a particular context, and containing all the social (both ethical and political) implications of Covid-19, together with the consequences of its management. The Covid-19 crisis impacted individuals' lives in various ways, doubtlessly beyond the sphere of physical health, and leading to deep irreversible changes in the structure of societies. I focus here on four key dimensions of the crisis which present ethical and political challenges, for which machine learning algorithms were able to support human decision-making.

1. With regards to bioethical issues, the Covid-19 crisis called for the management of the pandemic under limited available information and with insufficient resources such as masks, tests kits, resuscitation beds or artificial ventilation systems to treat a surplus number of patients in need. It resulted in medical staff having to make moral arbitrations in emergency situation, when facing dilemmas involving the allocation of intensive care treatments such as: *how would you fairly allocate ten artificial ventilators between thirty people in vital need, all differing in age and medical condition?* This situation was all the more difficult for non-military doctors who suddenly had to adopt the logics of 'war medicine' for saturated resuscitation services,[1] without being able to rely much on their professional organisations to assist them with such decisions. In France, while the French Society of Anaesthesia Intensive Care Medicine (SFAR) produced a short document listing clinical ethics guidelines for the allocation of intensive care treatments (Azoulay et al., 2020) – including criteria related to the patient's age, previous condition or current clinical severity – it scrupulously avoided any precise recommendations in order to avoid provoking the same polemic as in Italy, following the release of the SIAARTI's clinical ethics recommendations (Vergano et al., 2020).

2. With regards to economic consequences, the lockdowns decreed by several countries, and resulting in a third of the global population being quarantined, had critical consequences on micro- and macro economies. A number of businesses collapsed, others went through unprecedented social plans, resulting in great precarity for job seekers landing on an employment market devoid of opportunity. Trading some people's freedoms and economic situations to better secure others' health is never a small ethical affair, as there is no moral principle which always prioritises safety over freedom. The challenge further grows in complexity when involving a worsening of national debt in order to provide specific categories of people with financial support, not only regarding the fair allocation of such aids among the population, but also from an intergenerational ethical perspective, touching upon Derek Parfit's non-identity problem (Parfit, 2017). Additionally, these lockdowns also served to increase inequalities in various ways, including in working conditions for remote employees as well as for pupils already struggling at school, who then dropped out completely. All these foreseeable consequences should be taken into account in the ethical assessment of a lockdown decree.

3. With regards to psychological aspects, the correlation between an extended quarantine period and a drastic degradation of people's mental health with repercussions on social interactions was largely observed and documented. A longitudinal survey conducted in the U.K.

[1] https://www.liberation.fr/checknews/2020/03/24/face-a-la-saturation-des-services-de-reanimation-comment-decident-les-medecins$_1782738/$

reveals that the proportion of people that reported experiencing at least one severe underlying mental health problem has risen from 7% to 18% for men and from 11% to 27% for women since the beginning of the pandemic, resulting in 34% of women now sometimes feeling lonely and 11% often feeling lonely (respectively 23% and 6% for men) (Etheridge & Spantig, 2020). Loneliness would most severely hit young adults (aged 18-24), especially when living alone (more than doubling the odds of being lonely), and other risks factors include lower income, lower education and recent loss of employment (Groarke et al., 2020). The impact of quarantine on depression was also observed with an increase from 8.5% up to 27.8% of the interrogated population of U.S. young adults having depression symptoms between the beginning of the pandemic and mid-April 2020 (Ettman et al., 2020). The study also suggests personal financial situation to be a strong factor, finding people with less than $5,000 in savings as 50% more likely to have symptoms of depression than others, controlled by all other demographics. These psychological implications of the lockdown then transfer to social interactions, especially considering that the boredom experienced by many quarantined people is a factor known to be positively correlated to both offline bullying (Vasileia et al., 2017) and online bullying (Antoniadou et al., 2016). All these factors added to the pre-existing high level of polarization, especially in the U.S. where the politicisation of the response to the pandemic even made wearing a mask a politically charged action, converged towards a latent animosity which often transferred to physical violence. In France particularly, the lockdown period has been synonymous with manifestations of violence inside hospitals, prisoners' riots, clashes between policemen and local population in sensitive neighbourhood, and a rise in domestic violence. The United Nations Development Programme (2020) observed a worldwide increase in domestic violence and child abuses during the lockdowns, including a rise of 30% in France, 25% in Argentina, 30% in Cyprus, and 33% in Singapore.

4. With regards to political implications, the shifts in policies and recommendations from national governments and international organizations, combined with publicly-relayed disagreements between medical experts left populations greatly confused and particularly vulnerable to false news – with some of these proving particularly dangerous for individuals' health. Analysing c.11,500 pieces of content reported by users on Facebook and Instagram in June 2020 with Onur Çelebi (2021), we observed that the two main topics of perceived misinformation are related on the one hand to the Black Lives Matters events, and on the other hand to the management of Covid-19. Considering the latter, they identify three main sub-topics of distrust and misinformation related to the credibility of research for Covid-19 vaccines, the usefulness or dangers of wearing a mask, and the political management of the crisis as a whole, often perceived as overstated, manipulative, and ultimately feeding conspiracy theories.

Let us note that none of the issues above is distinctly new, as if it were the product of the Coronavirus itself. They are all pre-existing questions, for which we have been trying to find solutions more or less impatiently. What is new, however, is that the arrival of the Coronavirus forced us to address them all together, in a very short period time, and at an unprecedented scale. Bioethical dilemmas existed but have never come up as frequently in day-to-day life, nor have the incidences of online hate and physical violence. Online misinformation did not proceed from the Coronavirus, but the latter propelled the former to a whole new level of dangerousness, while such a global-scale lockdown and its economic consequences have never been experienced before. This is why we can rightly talk about the 'crisis of the Covid-19', for which the virus was only the trigger, catalysing these issues to a greater level of criticality, thereby forcing the deployment of new solutions to address them.

3 TWO INNER LIMITS OF PURELY ALGORITHMIC APPROACHES IN ADDRESSING SOCIAL ISSUES

One specificity of the issues considered here as part of the four dimensions of the Covid-19 crisis is that they cannot be 'solved' by algorithmic systems alone. Such an impossibility is to be justified by two types of limitations inherent to algorithms: one technical, relating to the 'contextual explicability and performance issue', and the other ontological, deriving from a conception of legitimacy.

Let us start with the technical limit. Every algorithmic system comes with an accuracy rate which never is perfect, models ultimately remaining simplifications of the world. For instance, it is particularly difficult for a misinformation classifier to detect humour or irony in a piece of content – an issue often referred to as a lack of ‘common sense’ – as well as to generalise a suggested claim from a post to all other instances of this claim among the diversity of contents – known as the ‘overfitting’ issue. However, the importance of the system’s performance together with the relevant degree of explicability required varies with applications, and even acquires a moral value in some cases. The issues we are considering here are of this kind, calling for higher standards when assessing systems’ performance, due to the potential consequences their inaccuracy could have on people’s lives. Let us refer to this as the ‘contextual explicability and performance issue’ and examine the following scenarios:

(a) A model predicts the probability of an industrial machine breaking down with 60% accuracy;

(b) A model predicts with 90% accuracy the effects of a new drug resulting in full recovery for 80% of patients and no secondary effects for others;

(c) A model calculates the dosage of an anaesthetic product to inject before surgeries with 99% accuracy, but resulting in the death of the patient in the 1% of cases when it is poorly dosed.

Scenario (a) and (b) do not raise particular ethical issue in themselves, as no victim results from their lack of accuracy. The third scenario, however, is highly problematic from an ethical point of view, as the imperfection of the model leads to patients’ death.

The crucial importance of the model’s performance then becomes slightly more complex when breaking down the types of errors it makes, especially false positives and false negatives deriving from misclassifications and resulting in an arbitration between two underlying performance metrics: the recall and precision rates. Thereby, a model detecting patients infected by HIV with excellent accuracy but a lower recall (meaning that it rarely classifies a non-infected patient as infected, but often classifies infected people as non-infected) is problematic as false negative people will not only not be treated, but they will also have an erroneous feeling of safety and may involuntarily contaminate others. In this case, we may want to concede a bit of precision to increase the recall. This would allow better detection of infected people at the cost of classifying sane people as infected, resulting in false positive people to go through slightly more medical testing than they should. In contrast, a model calculating the probability of a child repeating a class with great accuracy but poor precision (meaning that it better classifies students who would benefit from repeating a class than those who would not) is problematic as false positives in this case would be encouraged to unnecessarily repeat a class, consequently losing an entire year. In this case, it could be argued that prioritizing precision over recall is morally relevant, considering the induced sacrifice to be greater for false positive people who end up wasting a year than for false negative people, who still have the possibility to repeat the following class. Especially as the misclassification for false positives may be caused by an unaccounted signal, such as a depression, which likely would still not be resolved and potentially worsened by the repetition of a class.

The above cases illustrate the variation in the degree of confidence one may have in a model. However, a complete moral assessment of the model’s performance would also require considering the confidence the model itself grants to its outputs. In computer vision, a pattern recognition algorithm attributes a probability to all the possible classes of known objects, and it is common practice to compare algorithms’ accuracy not only on the basis of the top 1 prediction (does the class of objects with the highest probability match the category of the object to be recognised?) but also on the top 5 predictions (if not first, is it at least amongst the finalists?). This is when the selected threshold of confidence becomes morally relevant (should we still take into account the top 1 prediction when it has a 0.6 probability?), together with the range of top predictions considered.

Setting up a relatively low threshold and only considering the top 1 prediction, in spite of a very small spreads between the probabilities for the top predictions, might not have significant moral implications when the algorithm is requested to identify fruits and ends up classifying red apples as beets. It will, however, if it results in identifying a black person as a ‘gorilla’ as seen in the infamous bias in Google's auto-tagging feature [2]. Even more so if the biased system is implemented in an autonomous vehicle, where ethical settings will doubtless always prioritize human lives versus animals when landing on the market (Awad et al., 2018). As far as that goes, inviting the top-5 profiles out of a facial recognition solution used to identify a bank robber from CCTV recordings for the purpose of having them identified by a witness is not too worrisome. On the contrary, taking only the top 1 result for granted and using this as evidence against a culprit would be much more questionable.[3] In this case, just like in the detection of anomalies in medical imagery, it should not be forgotten that the top recommendations are always dependant of the training dataset, and that a top 1 result will always come out, even when the person or type of tumour we aim to detect is not included or very underrepresented in this dataset.

An optimist may say, however, that these technical limits are only temporary and that we could develop predictive algorithms reaching perfect accuracy on top-1 results at some point. In response to this, I will argue that, even in this ideal situation, the perfect accuracy of these algorithms would not counteract their incapacity to address the ethical and social issues considered here, as there would still be ontological limits inherent to the very nature of these systems. In fact, whether it relates to the moral arbitrations underlying the allocation of intensive care treatments, to the trade-offs between preserving some people's physical health at the cost of others' precarity and mental health, or to the capacity of ensuring trust in information, an intervention over these issues requires a specific legitimacy, which algorithms fundamentally lack.

At the collective level, a sophisticated model can surely predict the economic consequences of a six-week lockdown for target industries and households. But it certainly does not have the political authority to decree, and ensure by the use of legitimate police force, a compulsory quarantine over a whole country, with its underlying consequences on citizens' freedom, economic situation, mental health and social interactions. This is because some issues call for more than just an appropriate and efficient response: they call for the authority of a legitimate decision-maker engaging their responsibility through a free deliberate choice. A purely algorithmic management of the Covid-19 crisis could then ultimately reach greater results in minimizing the number of corporate bankruptcies under constraint of maximizing the number of lives saved, but it would also undermine the foundations of our democratic systems. From the classic contractarian philosophers to the modern constitutional theorists, it has remained unquestioned that in politics, the rightness of a decision is not to be exhausted by its sole outcomes, but also and firstly assessed on the basis of its procedural value. This value is compliance to law, and as summarised by Thomas Hobbes: 'auctoritas, non veritas facit legem' (1990). This is precisely what characterises the rule of law and justifies the preservation of political authority – i.e. the free recognition of the state's power by its citizens (Gadamer, 1976) – as the first and greatest end of modern democratic regimes. Losing sight of the strict hierarchy between the two levels of objectives – the first one being the upkeep and improvement of political authority, the second being the minimization of the Covid-19 crisis' negative consequences – results in failing to recognise the direction of the reciprocity between these two levels: the legitimate authority empowers the political decision, which in its turn maintains the authority as long as it does not betray it. In doing so, we engage in the vicious circle of the 'logic of the detour of production' characterizing 'counterproductive' societies for Ivan Illich (1975), as social systems taking their means for ends.

Note that this latter point is not an essentialist argument, and that a procedural infringement would not just have symbolic implications and a loss of theoretical sense. An illustration of this was given by the French government's communication strategy around masks. It first consisted in lying to the people about the stocks and the uselessness of wearing one in March, in a context where a collective hysteria prompted people to plunder pharmacies and hospitals to rob masks, for the purpose of prioritizing the access of this limited resource to medical staff [4]. Once the country had amassed sufficient stocks, the government enacted a decree in July making mask-wearing compulsory in public spaces under financial penalty. Although such a strategy can be reasonably defended between rational economic agents, the equilibrium thereby reached is clearly not optimal from a political perspective,

[2] https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas

[3] https://edition.cnn.com/2021/04/29/tech/nijeer-parks-facial-recognition-police-arrest/index.html

[4] https://www.liberation.fr/france/2020/04/27/masques-comment-le-gouvernement-a-menti-pour-dissimuler-le-fiasco$_1786585/$

as the loss of legitimacy resulted in civil disobedience, violent riots and a profound distrust of political authority, whose consequences will certainly be greater and last longer than the short term advantages secured in terms of mask management. A more common example occurs when a court is obliged to reject a valuable piece of evidence whose sourcing is subject to a procedural defect. Indeed, accepting such evidence would automatically corrupt the case, dissolve the authority of the judge, abolish the foundations of the juridical institution, and certainly provoke a scandal among law professionals.

At the individual level, a moral decision requires a moral agent, that is an entity provided with a certain conception of ‘good’ and a free will, thus capable of making moral decisions they can justify and be held responsible for. The moral individual level is even less open to algorithmic dynamics than the collective political one because, leaving aside the legitimacy question and only focusing on the subordinated objectives, the decision is of a kind that does not even benefit from a greater amount or processing of data. In the midst of the pandemic, Judith Thomson’s (1976) thought experiment, especially the transplant cases, became a dramatic reality as medical staff faced moral dilemmas involving the allocation of artificial ventilators and resuscitation beds, calling for heart-breaking arbitrations which placed their moral responsibility at stake. The following scenarios represent situations that some of them could or did face:

Would you refuse an available resuscitation bed to a patient with a 45% chance of surviving when the system predicts the imminent arrival of three new patients with a 70% chance?

Would you rather allocate a bed to someone with a 95% chance of survival but who will occupy it for six weeks versus an average of three for others?

Would you free the bed of someone who has been occupying it for three weeks with no sign of improvement, and a low probability to fully recover, in order to offer it to a patient with a 90% chance to recover fully within three days?

Would you accept giving a bed to someone with only a 20% chance of survival, but offering to buy ten new beds to be delivered the following week?

Notice these dilemmas already include precise information provided by statistical models which can indeed inform the decision-maker, but not support the rationality proper to the moral decision itself. Moral decisions intervene when there is no solution to be found: a moral dilemma requires a responsible decision because, by design, it cannot be solved. Philosophers greatly disagree on the ultimate exit of such dilemmas, as well as on the particular moral doctrine it should be based upon. They do, however, widely agree that what primarily matters for moral decisions is the nature of the decision-maker, which can only be a moral agent. The decision-maker’s moral agency is what elevates a choice to the rank of moral decision, by engaging their responsibility to bind themselves with this choice and the consequences that will derive from it. The indissociability of a moral decision with an agent’s responsibility is precisely what justifies the conclusion that algorithms cannot make moral decisions.

As a conclusion here, both at the individual and the collective level, it is not because a machine has greater chances to maximise the objectives pursued by a moral entity when taking a decision engaging their responsibility, that its decision is to be considered as preferable. Doing so would not only demolish the meaning of the decision and cancel the legitimacy of the moral entity, but also have great chances to lead to suboptimal outcomes. A gold rule to avoid this pitfall when questioning the relevant degree of automation in the decision-making process applied to political and ethical issues could be that ’legitimacy trumps accuracy’, as an adaptation of Ronald Dworkin’s more general principle, according to which ‘rights trump utility’ (Dworkin, 2013).

4 Four types of examples where human-algorithm cooperation empowers human judgement instead of challenging it

Although a purely algorithmic approach cannot address these issues alone, it would certainly be a mistake to categorically refuse the support of technology in tackling them. There is indeed no moral value in the cultivation of one's ignorance, nor in the rejection of opportunities for self-improvement and the development of more efficient practices. The whole question then dwells in the appropriate space they should be given to support human moral agency instead of negating it. To illustrate this, I shall now present practical use cases of machine learning-based solutions, which were leveraged during the Covid-19 crisis in such a way, to address some of the issues identified in our four social dimensions.

1. With regards to bioethical aspects, the main type of moral dilemmas faced by medical staff in the context of the Covid-19 crisis derives from the necessity to distribute a limited number of vital resources to a surplus number of patients in need. While algorithms cannot take such decisions, they can help predict the evolution of the outbreak to identify when and where these dilemmas will most probably occur. This enables the transfer of resources at the service, hospital or country level, either to prevent dilemmas from happening, or to best prepare teams to address these ethical challenges. In fact, you do not need to choose which of two patients should benefit from a ventilator when you are able to predict this situation and had the time to transfer another ventilator from a hospital with fewer patients.

   Many research teams engaged in the effort of modelling the spread of the outbreak at different scales. An example of these is the collaboration between Facebook AI and New York University's Courant Institute of Mathematical Sciences, which focussed on creating localised forecasting models for the State of New Jersey, applying multivariate Hawkes processes to publicly available data to create daily Covid-19 predictions. These forecasts, once combined with hospitals' data on the location of available resources, could help them improve their resource planning to prepare for an expected increase in patients, which not only include the management of material resources (ventilators, masks, beds, etc.) but also human resources (staff schedule). Another joint effort between Facebook AI and New York University's Department of Radiology provides an example of AI use to build hospital-specific forecasting for resource planning. Researchers leverage different learning methods to predict from de-identified clinical data such as X-rays and computer tomography scans the number of patients whose condition is likely to improve or worsen in a given time period, as well as how many of them are likely to be admitted, transferred or discharged from intensive care units.[5]

2. With regards to economic aspects, the main type of issues relates to the trade-offs underlying national lockdowns. Here again, forecasting models predicting the evolution of the outbreak can be used to re-open businesses locally once indicators fall below determined thresholds, which may help temper the impact of a national lockdown over local economies. Another path is explored by Emer2gent, an alliance initiated by Rolls-Royce 'to accelerate and smooth economy recovery from the COVID-19 crisis'.[6] Using machine learning methods to cross-analyse economic, business, travel and retails datasets, these allies aim to provide insights and applications to support the global Covid-19 response – for instance by identifying lead indicators of economic recovery cycles to inform policy decisions, and limit recessionary impacts by supporting operating confidence in investment and activities.

   The potential benefits of such initiatives, conjugating machine learning techniques with economic data to inform policymakers and support market confidence, should however come with a strong warning. Indeed, this type of approach always dances around the abyss of Illichian counterproductivity, as the temptation is great to lose sight of the initial objectives and start considering growth indicators, not as a means towards these ends anymore, but as ends themselves. This is the path that another initiative, the AI Economist, seems to be taking. Built for the purpose of advising governments on their tax policy in the post-Covid-19 period, the project leverages reinforcement learning methods to develop a system that 'can predict how people would actually respond to a tax, like whether it incentivizes them to work more or work less'.[7]

[5] https://ai.facebook.com/blog/using-ai-to-help-health-experts-address-the-covid-19-pandemic/

[6] https://emer2gent.org/the-alliance/

[7] https://www.salesforce.com/news/stories/introducing-the-ai-economist-why-salesforce-researchers-are-applying-machine-learning-to-economics/

Such an approach to tax theory entirely reverses the classic conception of citizenship. Not only does it reduce citizens, from subjects of law empowering the state via tax collection, to a constraint in the model, a mere source of frictions existing through their nuisance capacity (work less, fraud more, riot), and a variable that should be minimized to reach a higher equilibrium in the tax collection. It also structures the whole system around a fundamental opposition between citizens' and governments' interests, playing an adversarial zero-sum game. By doing so, the AI Economist carries the threat of bringing the Laffer curve in tax theory to a whole new level, and harm political systems as much as the Coase theorem, when used by the Law Economics school to theoretically justify and practically enable a rights market, which resulted in a complete denaturation of rights in the modern age (Etienne, 2021a). When indicators are used not only to inform policies but also to dictate them, this marks the start of 'governance by numbers' (Supiot, 2015) and the use of statistics as normative and prescriptive. When these indicators even come to dictate the behaviours of agents they are supposed to describe, there begins the 'algorithmic governmentality' (Rouvroy & Berns, 2013) and the failure of a statistics, whose endogeneity over the feedback loop has made them monologic.

3. With regards to social aspects, the main type of issue relates to the challenge of maintaining sane interpersonal interactions in difficult times. While it is complicated to act directly on physical violence through computational means, it is possible to detect and delete hate speech online, as well as support victims of bullying. AI is already used by social media platforms to detect and remove hate speech. As an example of this, automatic detection methods enable Facebook to proactively identify c.90% of hate speech removed by the platform (representing 9.6 million pieces of content in the first quarter of 2020) thanks to deep content semantic understanding with multimodal learning.[8] This approach notably enables classifiers to identify suggested ideas across different languages and modes (e.g. a caption and its associated images). It also leverages both fusion models to build embeddings from multi-modal inputs, and BERT technique to build a cross-lingual models (Conneau et al., 2019).

   Although necessary, taking actions on the symptoms of hate remains insufficient, underlining the need to investigate ways to intervene in the causes of deleterious interactions. However, research on the psychology of bullying converges to say that the majority of people involved in bullying are subject to mental health issues, that they often are, or have been, bullied themselves, and identify boredom as a significant trigger to bullying behaviours (Vasileia et al., 2017; Antoniadou et al., 2016). The auto-generative aspect of bullying, as transforming victims into actors, then somewhat justifies that the detection and removal of online bullying not only act on bullying's effects, but also on its causes. This means of action is not trivial, as it was observed that up to 50% of studied perpetrators were also found to be victims of bullying (Haynie et al., 2001; Brown et al., 2005), with the involvement in bullying, as either bullies or victims, being associated with greater likelihood of weapon carrying and fighting (Nansel et al., 2003). Victims of bullying also tend to suffer more from a lack of self-esteem and depression (Morrison Gutman & Feinstein, 2008). Machine learning techniques have proven efficient in detecting clinical symptoms of depression from social media posts, both on Twitter (De Choudhury et al., 2013) and on Facebook (Eichstaedt et al., 2018), thus making it possible to identify most vulnerable people, and best support them. Apart from these psychological factors, it is also possible to take actions to reduce boredom, for instance by resorting to lossless compression algorithms to maintain online services during connection peaks. In March 2020, Vodafone reported a 50% rise in internet use in several markets due to the many people working remotely and spending time on streaming platforms, [9] resulting in Netflix, among other platforms, reducing the quality of its video streaming service to face the increase in demand. As far as can be seen, ensuring the access to video streaming and online gaming platforms may certainly have had prevented a number of bullying behaviours.

[8] https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/

[9] https://www.vodafone.com/news-and-media/vodafone-group-releases/news/vodafone-launches-five-point-plan-to-help-counter-the-impacts-of-the-covid-19-outbreak

4. With regards to the political aspects of the crisis, the main issue is maintaining social trust at a time when misinformation is rampant. Considering the c.33 million pieces of unique content reported by c.24 million users worldwide just in September 2020, it seems evident that content moderators cannot tackle the supra-human challenge of online misinformation alone. The good news is that the majority of these reports actually proceed from other objectives than accurately reporting potential false news (Etienne et al. 2021). Misinformation classifiers can leverage a number of signals such as the fact-checkability of the post (does it contain a claim that could be true or false?) or the number of disbelieving comments it is associated with, to attribute a misinformation prevalence score to each piece of content. Above a given threshold, the contents with the highest score are enqueued and made available to independent third-party fact-checkers and actions are then taken on the posts confirmed to contain a false claim by fact-checkers (which requires provision of a debunking source as a justification). This labelling effort is then leveraged by SimSearchNet, a convolutional neural network-based model detecting near-exact duplicates, so that fact-checkers are less likely to review the same debunked false news multiple times.[10]

   Although perfectible, I hold this misinformation management process to offer a great illustration of the desirable complementarity we may look for between algorithmic solutions and human judgement, as a means for humans to operate on supra-human dimensions. In this example, AI is used first to prioritise the work of fact-checkers – identifying the most relevant (or problematic) pieces of contents – thus reducing the detection task from a supra-human to a human dimension, and second to leverage their work, applying their labels at the supra-human level. From the perspective of the whole process taken as a system, we can observe an interesting mise en abyme of AI, acting just like an autoencoder, operating dimension reduction and augmentation between the latent state of individuals and a higher space: the infosphere.[11]

   As is every system, it is subject to the risk of degeneration, and Deepfake provides an evident illustration of this. While it is principally used today to produce pornographic content (Ajder et al., 2019), and while the few political applications remain mostly as warning educative examples, things would be very different were this technique to be used more and more for the purpose of misinformation. Although this would not directly impact social trust, as argued elsewhere (Etienne, 2021b), the whole dynamic behind the adversarial game consisting in building solutions (notably based on autoencoders) to counter other solutions developed from the same techniques, i.e. that the algorithms created to detect misinformation are also used now to produce it, would be an edifying example of Illichian counterproductivity.

5 AI PAVES THE WAY FOR A FUTURE OF HYPER-RESPONSIBILITY TOWARDS META-DILEMMAS

Although imperfect and always submitted to the risk of counterproductivity, I believe the examples discussed above show us the right direction in preserving responsibility and moral agency, as they illustrate cases where artificial systems do not take major decisions themselves, but complement human judgement, to empower it and leverage moral decisions. Such a framework for the distribution of tasks between humans and artificial systems is however challenged by some people supporting a moral recognition of artificial agents (AAs).

Overcoming the ontological limit for the purpose of enabling automated systems to make moral decisions necessitates a 're-ontologisation' of AAs (Floridi, 2013). Some philosophers have been addressing this question, arguing both pro (e.g., Floridi & Sanders (2004); Sullins (2011)) et contra (e.g., Moor (2011); Basl (2014)) the moral recognition of AAs, but without yet reaching a convincing position that would ensure an effective responsibility of AAs as moral agents, thus making such a

[10] https://ai.facebook.com/blog/using-ai-to-detect-covid-19-misinformation-and-exploitative-content/

[11] I refer here to Luciano Floridi's concept, defined as "the whole informational environment constituted by all informational entities (thus including information agents as well), their properties, interactions, processes, and mutual relations" (Floridi, 2013).

change socially desirable. Contrariwise, it seems that the main reason supporting the discourses for the recognition of AAs' rights is not essentially grounded, but mainly instrumental, for the purpose of justifying delegating them a number of important decisions, thereby scarifying the meaning and responsibility on the altar of convenience. Indeed, at the antipodes of animal ethics, it is quite evident that the motivation for raising AAs to the rank of moral agents is not to recognise our own moral obligations towards them – otherwise we would rather focus on moral patienthood, a much easier position to defend – but to empower AAs to take decisions that would violate moral obligations between individuals and undermine human autonomy. Once again, the danger does not come AAs as the theorists of existential risks like to believe, but from humans using AAs as a means to achieve ends that would otherwise exceed their biological capacity and moral legitimacy, would these be performed by them directly.

In practice, we do not, however, need to formally grant moral agency to AAs to enable them to make decisions that would normally require the legitimacy of a moral agent or a political authority. Automated systems are often unconsciously empowered to make such decisions without the intervention of a human being, leading to technologies being 'paternalistic by design' (Millar, 2015), but it also happens in a more direct way on a daily basis. Think about a judge having to take a decision about a case, for which an algorithm calculating recidivism scores declared that the culprit has ninety five percent chances to re-offend: is the judge's decision still as free and autonomous as is expected of a moral agent? Of course not. We can make reference to cognitive biases, such as anchoring effects (Sherif et al., 1958) limitating the freedom of his decision-process, but we can also discuss the impact of the information itself, its reliability and the extra-justifications the magistrate would have to provide, if deciding to go against the intuitive decision deriving from the algorithm's information. In fact, all the data considered in forming the hypotheses of the dilemmas previously discussed (e.g. would you rather allocate a bed to someone with a 95% chance of survival but who will occupy it during six weeks versus three for others?) should not be accepted as axioms of the dilemmas, but should be part of the moral decision.

There is a clash between the degree of informed decision-making and that of free will. Analysed through Illich's framework, the information coming from a decision-making support solution refers to the 'heteronomous mode' of knowledge production, which aims to support the autonomy of the decision, but only if the agent is capable of making informed use of this instrument. Otherwise, they enter a state of dependency on this source of knowledge, and so does their decision. Besides formally granting AAs moral agency, another pitfall then consists in granting them *de facto* moral agency, as it surely is more comfortable to disengage one's own responsibility and let machines take the hard decisions for us, hiding behind the artificial neutrality of their pure rationality. At the antipodes of this posture, an alternative to save and even enhance moral responsibility, would consist in extending an agent's moral responsibility (and even legal liability in some cases) from its sole decision (which autonomy is more and more 'artificial'), to the full extent of the instruments that have mainly participated in the decision-making process, thus forming a moral perimeter. When a public person makes bombastic claims based on false news picked up from a conspiracist website, it is well-accepted that this person is rightly mocked, as it was their responsibility to verify the credibility of the source. Likewise, there is no reason for a doctor, a judge or a political representative, when making decisions involving existential consequences for a great number of individuals, to have the privilege to rid themselves of the responsibility when utilising instruments of which they know little. This is precisely when the 'contextual explicability and performance issue' strikes, calling for a minimal performance and degree of explicability exigency, but also specific training and elementary knowledge for the individuals using these tools to make decisions involving a specific authority, so they can truly be held responsible for it.

Finally, the main aspect of the ethical issues I have discussed so far relates to the decision itself and the appropriate distribution of tasks between humans and algorithms, but I have not discussed what comes before the momentum of the moral decision. There is a crucial difference between a moral dilemma deriving from a thought experiment and one proceeding from an applied case study, and that is the mode of temporality. While an applied case study demands the moral agent make a decision in praesenti, often in a time lapse that does not even allow the necessary deliberative process to engage its full moral responsibility – e.g. in the non-autonomous vehicles dilemmas (Etienne, 2020) – the thought experiment refers to a dilemma that still exists on the mode of a potential. This key difference offers two advantages to the moral agent: on the one hand, they have enough time to reflect and discuss the best alternative to make a genuine deliberate moral decision were the situation ever to occur, and on the other hand they have the possibility to take action on the conditions of occurrence of the dilemma themselves.

Even were we to find an unquestionable solution for this dilemma, this would not guarantee that people would necessarily endorse it. Moral psychologists have observed that people tend to stick to their moral intuitions even when their supporting arguments have been proven wrong (Haidt, 2001). A great inconsistency was also experimentally verified in many people between the choice they make in the context of a thought-experiment, and the one they actually implement when facing the real situation (Francis et al., 2016; Bostyn et al., 2018). In my opinion, a moral dilemma is synonym of aporia and encountering one is an evidence of failure. It certainly has an instrumental value in challenging our moral intuitions, such as the famous case of the fat man and the bridge (Thomson, 1976), but by definition, and unlike a problem, does not allow for the possibility of a solution as it is. This is why it calls for a decision – of which the Latin root *decidere*, literally cut the victims' throat, reminds us that it implies a sacrifice – and this is certainly why, regardless of the decision made, we are never fully satisfied. The problem of the dilemma, or the dilemmic aspect of the issue, resides in its own formulation and the point of view it forces us to adopt on a question that is purposely designed to not be resolved. Therefore, whereas we have no other choice when facing an actualised moral dilemma than making a heart-breaking moral decision for which, fortunately, we may not be held entirely responsible, the only solution to solve a moral dilemma is to make every effort to prevent its conditions of occurrence.

The whole project of AI being put to use towards a better mastery of the future and probabilistic prediction of events to occur will give us a unique power to develop an *ex ante* control over moral dilemmas by neutralising their conditions of possibility. This is already happening, with forecasting models allowing us to operate a general supervision of limited resources in order to minimize the likeliness of a dilemma occurring at the local level, when the demand for a vital resource exceeds its supply in a hospital. This power is to come with an unprecedented level of responsibility deployed on a whole new temporality, superposing dilemmas and meta-dilemmas: when all foreseeable dilemmas cannot be avoided together, which ones should we allow to occur? This means that we will not only be responsible for the consequences of past decisions (what happened and what did not happen), but also for counterfactual consequences (what could and could not have happened) as we will be able to better model these, together with the whole field of possibility in the future (what can still happen and what cannot). Environmental ethics already provides illustrations of events that have not happened yet, and for which we can already be held responsible for. This is considering that we somewhat provoked them, but also and mostly because we did not prevent them from happening, despite being able to foresee them and do something about it while it was still possible. More than being responsible for these events, we may be considered responsible for their inevitability and the impossibility of future generations to address them.

6 Conclusion

For the purpose of showing how AI can not only help us address the sanitary aspects of the pandemic, but also the social dimensions of the Covid-19 crisis as a whole, I focused on four key dimensions which give rise to ethical issues. I presented two major limitations of automated decision systems which invalidate solving these issues with algorithmic solutions alone. This led to the gold principle which holds that 'legitimacy trumps accuracy', and justifies the need for sensible collaboration between artificial and moral agents. Examples of such collaboration enhancing moral agency were provided for each of the four dimensions and tempered by the inherent risk for every system to degenerate, which was approached here under the prism of Illich's counterproductivity.

I finally argued that although moral agency should not be granted to artificial agents, many situations reveal a partial delegation of moral agents' autonomy involving a reduction of responsibility. In contrast, delimiting a moral perimeter could enhance responsibility and best tackle the contextual explicability and performance issue. AI can not only help us better address moral dilemmas, establishing a circular relationship between human-dimensioned and a supra-human-dimensioned spaces, but also solve them through and *ex ante* management, enabling us to intervene on their conditions of possibility. Acknowledging the role played by algorithms in our decision space, together with the unprecedented predictive power they offer, unlocks a new breed of moral situations: AI-generated meta-dilemmas. This propel us into a space of hyper-responsibility towards the conditions of possibility of future actions, over an even further future.

REFERENCES

Gender-based violence and covid-19. Technical report, United Nations Development Programme, 2020.

Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. The state of deepfakes: Landscape, threats, and impact. Technical report, Deeptrace, 2019.

Nafsika Antoniadou, Constantinos M Kokkinos, and Angelos Markos. Possible common correlates between bullying and cyber-bullying among adolescents. *Psicología Educativa*, 22(1):27–38, 2016.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

Elie Azoulay, Sadek Beloucif, Guidet B Vivien, Dominique Pateron, and Matthieu Le Dorze. Décision d'admission des patients en unités de réanimation et unités de soins critiques dans un contexte d'épidémie à covid-19, 2020.

John Basl. Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. *Philosophy & Technology*, 27(1):79–96, 2014.

Dries H Bostyn, Sybren Sevenhant, and Arne Roets. Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological science*, 29(7): 1084–1093, 2018.

Stephen L Brown, David A Birch, and Vijaya Kancherla. Bullying perspectives: Experiences, attitudes, and recommendations of 9-to 13-year-olds attending health education centers in the united states. *Journal of School Health*, 75(10):384–392, 2005.

Katharina Buchholz. What share of the world population is already on covid-19 lockdown?, 2020.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Jaime A Teixeira da Silva, Panagiotis Tsigaris, and Mohammadamin Erfanmanesh. Publishing volumes in major databases related to covid-19. *Scientometrics*, 126(1):831–842, 2021.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 3267–3276, 2013.

Ronald Dworkin. *Taking rights seriously*. AC Black, 2013.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018.

Ben Etheridge and Lisa Spantig. The gender gap in mental well-being during the covid-19 outbreak: Evidence from the uk. Technical report, ISER Working Paper Series, 2020.

Hubert Etienne. When ai ethics goes astray: A case study of autonomous vehicles. *Social Science Computer Review*, pp. 0894439320906508, 2020.

Hubert Etienne. *Le Cens de l'Etat*. in press (Les Belles Lettres), 2021a.

Hubert Etienne. The future of online trust (and how deepfake is advancing it). in press, 2021b.

Hubert Etienne and Onur Çelebi. A general classification for user misinformation reports reveals new manipulation strategies and reporting behaviours. under review, 2021.

Catherine K Ettman, Salma M Abdalla, Gregory H Cohen, Laura Sampson, Patrick M Vivier, and Sandro Galea. Prevalence of depression symptoms in us adults before and during the covid-19 pandemic. *JAMA network open*, 3(9):e2019686–e2019686, 2020.

Luciano Floridi. *The ethics of information*. Oxford University Press, 2013.

Luciano Floridi and Jeff W Sanders. On the morality of artificial agents. *Minds and machines*, 14 (3):349–379, 2004.

Kathryn B Francis, Charles Howard, Ian S Howard, Michaela Gummerum, Giorgio Ganis, Grace Anderson, and Sylvia Terbeck. Virtual morality: Transitioning from moral judgment to moral action? *PloS one*, 11(10):e0164374, 2016.

Hans Georg Gadamer. *Vérité et méthode: les grandes lignes d'une herméneutique philosophique*. Seuil, 1976.

Jenny M Groarke, Emma Berry, Lisa Graham-Wisener, Phoebe E McKenna-Plumley, Emily McGlinchey, and Cherie Armour. Loneliness in the uk during the covid-19 pandemic: Cross-sectional results from the covid-19 psychological wellbeing study. *PloS one*, 15(9):e0239698, 2020.

Jonathan Haidt. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):814, 2001.

Denise L Haynie, Tonja Nansel, Patricia Eitel, Aria Davis Crump, Keith Saylor, Kai Yu, and Bruce Simons-Morton. Bullies, victims, and bully/victims: Distinct groups of at-risk youth. *The Journal of Early Adolescence*, 21(1):29–49, 2001.

Thomas Hobbes. *Leviathan or The Matter, Forme and Power of a Commonwealth Ecclesiasticall and Civil in Carrive, Lucien and Paulette, Œuvres de Hobbes*. Vrin, 1990.

Ivan Illich. *Energie et Equité*. Seuil, 1975.

Jason Millar. Technology as moral proxy: Autonomy and paternalism by design. *IEEE technology and Society Magazine*, 34(2):47–55, 2015.

James H Moor. The nature, importance, and difficulty of machine ethics. *Machine ethics*, pp. 13–20, 2011.

L Morrison Gutman and L Feinstein. Children's well-being in primary school: Pupil and school effects. Technical report, Centre for Research on the Wider Benefits of Learning, 2008.

Tonja R Nansel, Mary D Overpeck, Denise L Haynie, W June Ruan, and Peter C Scheidt. Relationships between bullying and violence among us youth. *Archives of pediatrics & adolescent medicine*, 157(4):348–353, 2003.

Derek Parfit. Future people, the non-identity problem, and person-affecting principles. *Philosophy & Public Affairs*, 45(2):118–157, 2017.

Antoinette Rouvroy and Thomas Berns. Gouvernementalité algorithmique et perspectives d'émancipation. *Réseaux*, (1):163–196, 2013.

Muzafer Sherif, Daniel Taub, and Carl I Hovland. Assimilation and contrast effects of anchoring stimuli on judgments. *Journal of experimental psychology*, 55(2):150, 1958.

John P Sullins. When is a robot a moral agent. *Machine ethics*, 6(2001):151–161, 2011.

Alain Supiot. *La gouvernance par les nombres*. Fayard, 2015.

Judith Jarvis Thomson. Killing, letting die, and the trolley problem. *The Monist*, 59(2):204–217, 1976.

Vassou Vasileia, Vassiou Aikaterini, Stavropoulos Vasileios, and Chaintouti Vasiliki. The relationship between boredom, interpersonal closeness/bullying and victimization in the school environment. *Psychology and Behavioral Science*, 6(3), 2017.

Marco Vergano, Guido Bertolini, Alberto Giannini, Giuseppe R Gristina, Sergio Livigni, Giovanni Mistraletti, Luigi Riccioni, and Flavia Petrini. Clinical ethics recommendations for the allocation of intensive care treatments in exceptional, resource-limited circumstances: the italian perspective during the covid-19 epidemic, 2020.