Policy and Internet

Policy & Internet, Vol. 11, No. 1, 2019

Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems

Ben Wagner

Automated decision making is becoming the norm across large parts of society, which raises interesting liability challenges when human control over technical systems becomes increasingly limited. This article defines "quasi-automation" as inclusion of humans as a basic rubber-stamping mechanism in an otherwise completely automated decision-making system. Three cases of quasi-automation are examined, where human agency in decision making is currently debatable: self-driving cars, border searches based on passenger name records, and content moderation on social media. While there are specific regulatory mechanisms for purely automated decision making, these regulatory mechanisms do not apply if human beings are (rubber-stamping) automated decisions. More broadly, most regulatory mechanisms follow a pattern of binary liability in attempting to regulate human or machine agency, rather than looking to regulate both. This results in regulatory gray areas where the regulatory mechanisms do not apply, harming human rights by preventing meaningful liability for socio-technical decision making. The article concludes by proposing criteria to ensure meaningful agency when humans are included in automated decision-making systems, and relates this to the ongoing debate on enabling human rights in Internet infrastructure.

KEY WORDS: automation, function allocation, human rights, Internet architecture, technology policy, artificial intelligence, algorithms

自动化决策正成为大部分社会范围内的常态。当人类对技术系统的控制变得越来越有限时，这就带来了有趣的责任挑战。本文将“准自动化”定义为人类作为一种基本的橡皮图章式机制纳入一个完全自动化的决策系统。笔者研究了三种准自动化案例：自动驾驶汽车，基于乘客姓名记录的边境搜索和社交媒体的内容调节。在这些案例中，个人能动性在决策中的作用是有争议的。虽然有专门的调控机制进行完全自动化的决策，但如果人为例行公事做决定(橡皮图章)，这些调控机制就不适用了。更广泛地说，大多数调控机制都遵循二元责任模式，试图对人或机器进行监管，而不是寻求对两者共同调控。这便产生了调控机制不适用的调控灰色地带，通过防止人类对社会技术决策承担有意义的责任而损害人权。文章最后提出了在人类被纳入自动决策系统时确保意义能动性的标准，并将其与如火如荼的网络基础设施促进人权辩论联系起来。

关键词： 自动化，功能分配，人权，网络架构，技术政策，人工智能，算法

doi: 10.1002/poi3.198

La toma de decisiones automatizada se está convirtiendo en la norma en grandes partes de la sociedad, lo que plantea desafíos de responsabilidad interesantes cuando el control humano sobre los sistemas técnicos se vuelve cada vez más limitado. Este artículo define la "casi automatización" como la inclusión de los seres humanos como un mecanismo básico de "rubber-stamping" en un sistema de toma de decisiones que de lo contrario sería completamente automatizado. Se examinan tres casos de casi automatización, donde la agencia humana en la toma de decisiones es actualmente discutible: automóviles sin conductor, requisas fronterizas basadas en registros de nombres de pasajeros y moderación de contenido en las redes sociales. Si bien existen mecanismos regulatorios específicos para la toma de decisiones puramente automatizada, estos mecanismos reguladores no se aplican si los seres humanos están siendo entidades de "rubber-stamping" mientras existen las decisiones automatizadas. En términos más generales, la mayoría de los mecanismos regulatorios siguen un patrón de responsabilidad binaria al intentar regular la agencia humana o de máquinas, en lugar de buscar regular ambas. Esto da como resultado áreas regulatorias grises donde los mecanismos reguladores no aplican, lo que perjudica los derechos humanos al evitar responsabilidades significativas para la toma de decisiones socio-técnicas. El artículo concluye al proponer criterios para garantizar una agencia significativa cuando se incluye a los seres humanos en los sistemas de toma de decisiones automatizados, y se relaciona con el debate en curso sobre la habilitación de los derechos humanos en la infraestructura de Internet.

PALABRAS CLAVES: automatización, asignación de funciones, derechos humanos, arquitectura del Internet, políticas tecnológicas, inteligencia artificial, algoritmos

Introduction

There has been an increasing awareness in the last two decades that the Internet is a socio-technical system (Brey, 2005), comprising both human and technical aspects (Kitchin & Dodge, 2011). Growing discussion about the role of humans in Internet infrastructure is reflected in the literature on Internet governance (Denardis, 2011; Johnson & Post, 1996; Mueller, 2010), content moderation (Roberts, 2014, 2016; Wagner, 2016), and—more broadly—the invisible human labor that keeps much of the Internet going (Ekbia & Nardi, 2014; Irani, 2015; Kushner, 2013; Mathew, 2014). This tension in the relationship between people and technology is particularly evident in regard to Internet platforms, whose claim to a legal and technical neutrality in how they manage the content on their platforms is central to their ability to conduct business (Crane, 2011; Kelsey, 2008; Pasquale, 2016). However, the claim that they are neutral actors requires them to hide the invisible human labor involved in curating their platforms (Riesewieck, 2017), and instead claim that intelligent technical systems are able to solve many content moderation problems (Wolverton, 2018).[1]

This is also reflected in debates on the human rights architecture of the Internet,[2] with many authors making proposals to strengthen the implementation of human rights through technical architecture (Gill, Redeker, & Gasser, 2015; Suzor, 2018). While John Perry Barlow's 1996 *Declaration of the Independence of Cyberspace* could be seen as a first document of this kind, numerous similar documents have followed. While the 2005 *Tunis Agenda for the*

Information Society is notably weak on this topic, more recent documents such as a 2012 proposal to the European Parliament for "Human Rights Based Communications Infrastructure" (Wagner, 2012), the IETF "Guidelines for Human Rights Protocol Considerations"[3] (Cath & Floridi, 2017) and the Council of Europe's "Study on the human rights dimensions of automated data processing techniques and possible regulatory implications" (Wagner, 2018a) have discussed how to integrate human rights into Internet infrastructure extensively.

While some of these human rights oriented documents focus primarily on defining technical techniques (IETF, HRCBI, HRBCI) or governance principles (WSIS, CoE, HRBCI) by which the governance of the Internet should take place, only very few deal with the specific challenge of human agency. The Council of Europe study is the only one of the reports mentioned above which specifically tries to embed human agency into a human rights Internet infrastructure by arguing that "important elements (such as discretion) of decision-making processes cannot be automated" (Wagner, 2018a, p. 42). A similar statement by the Article 29 Working Party in their interpretation of the General Data Protection Regulation (GDPR), argues that it implies a "prohibition on fully automated individual decision-making, including profiling that has a legal or similarly significant effect" (Article 29 Data Protection Working Party, 2017, p. 9).

This follows a trend related to the algorithms and automation literature which both questions and emphasizes the importance of human agency within otherwise vastly automated and mostly Internet-connected technical systems (Gillespie, 2014; Kitchin, 2017; Pasquale, 2015). This article examines the relationship between human agency and human rights Internet architecture in order to attempt to understand whether there is a necessary role for human agency in human rights Internet architecture and, if so, what that role should be.

Imagining Human Rights in Socio-Technical Internet Architecture

When studying the role of human agency in Internet architecture, one of the key questions that comes to the fore is the challenge of function allocation. While it is clear that many functions conducted within the Internet's architecture—such as moving packets or displaying websites—can be conducted only by machines, many others, such as content moderation or ensuring the physical delivery of items ordered, could be completed either by a machine or a human. Thus, while only infrequently discussed as such, function allocation is a highly relevant and politically contested question for many aspects of Internet architecture. For example, while some politicians claim that Facebook should automatically find and detect terrorist content and remove it, others—such as the Council of Europe Expert Group on Internet Intermediaries (MSI-NET)—argue that only humans should make such content moderation decisions.[4]

The standard method to determine whether machines or humans should best be used in certain parts of a system has been to use the MABA-MABA lists first

published by Fitts (1951). These lists are typically used to decide whether a machine or a human can better fulfill a task (Figure 1).

However, since the development of this list, there has been considerable debate on whether this list—which Sheridan (2000, p. 203) describes as the "the function allocation counterpart of Moses' 10 commandments"—is relevant from a human rights perspective, with Billings (1991) proposing *human-centered automation* as a way of "keeping the human front and center as automation technology in one form or another continuous to engulf us" (Sheridan, 2000, p. 209). One of the key aspects of this human-centric approach to function allocation is to ensure that a "human operator [is] in the decision and control loop" (Sheridan, 2000, p. 209).

This key recommendation of Sheridan (2000) has translated into many recommendations made about human rights infrastructure. While MABA-MABA lists do not explicitly use the term human rights, many of the terms they use suggest that their human-centric approach is essentially attempting to ensure that human judgment and discretion also flows into decision making. As noted by Billings (1991, p. 7): "automation is simply one of many resources available to the human operator, who retains the responsibility for management and direction of the automation and the overall system." A similar argument is made by Fitts (1951, p. 7), who suggests that the "nebulous ability we call judgment also appears to be unique in the human."

This assumption made by Billings (1991) and others is also reflected in the literature on automated weapons systems, where it is argued that it is preferable to keep a "human in the loop" for several reasons (Crootof, 2016; Roff & Moyes, 2016). For one, it is assumed that improving the quality of decisions made by automated systems by including humans in the process will limit the risks of their implementation. Thus, some organizations focused on promoting human rights in Internet infrastructure have argued that it is important to maintain the option for human intervention. For example, a recent Council of Europe study on automated decision-making systems states:

The Fitts MABA–MABA List

Men are better at
Detecting small amounts of visual, auditory or chemical energy
Perceiving patterns of light or sound
Improvising and using flexible procedures
Storing information for long periods of time, and recalling appropriate parts
Reasoning inductively
Exercising judgement
Machines are better at
Responding quickly to control signals
Applying great force smoothly and precisely
Storing information briefly, erasing it completely
Reasoning deductively

Figure 1. The Fitts List (Sheridan 2000, 204).

> While algorithmic decision-making is increasingly adept at replacing human decision-making, important elements (such as discretion) of decision-making processes cannot be automated and often become lost when human decision-making processes are automated.[5]

Based on the basic logic of function allocation, it could thus be argued that human decision making is an important part of what could be considered to be a human rights–based Internet architecture. What is interesting is that while some reports do not touch on the question of human agency at all, those that do so explicitly require it. This suggests that those individuals who have engaged with the concept of human agency in relation to human rights infrastructure believe some level of it to be a necessary condition for safeguarding human rights. This article examines the extent to which keeping humans in the decision loop—that is, ensuring meaningful human agency in decision-making processes—is necessary to safeguard human rights.

The claim that keeping a human in the decision loop is necessary to safeguard human rights is consistent with what both Sheridan (2000) and Billings (1991) argue for as a human-centric approach to automation. It is also consistent with the arguments developed by the Council of Europe Expert Group on Internet Intermediaries (MSI-NET), which suggest that the allocation of many key decision-making functions to humans is important to safeguard human rights.[6] Given that the current literature on function allocation and existing statements of important expert groups suggest that some level of human discretion should be safeguarded in regard to function allocation in developing a human rights Internet architecture, this claim seems like a useful starting point to analyze the role of human agency within such an architecture.

To understand what this claim means, we will consider several examples of cases in the context of Internet architecture where humans are currently kept in the loop. This will help us understand what basic assumptions are made about human decision making within Internet architecture (broadly understood), in relation to machine decision making, and help us understand what other reasons might exist for including humans in the loop beyond merely improving the quality of governance. It will be suggested that many assumptions about "human agency and ethical humans vs. impartial machines" do not serve well in designing human rights into Internet architecture.

Keeping a "Human in the Loop"? Three Cases

For a wide variety of reasons, many organizations choose to keep a human in the loop when they operate automated technical systems. This approach is one of the most typical responses to mostly automated systems to ensure that the results provided by a computer algorithm are not the sole reason for decision making, but that the decision involves human decision making as well. There are many different fields where this type of interaction between human and machine is increasingly common, with the machine suggesting a solution to a computer

operator that the computer operator can respond to before a final decision is made. The following cases are presented as a diverse set of examples of this phenomenon, to demonstrate some of the challenges associated with including humans in the loop.

Self-Driving Cars

Perhaps the most obvious example of such behavior involves self-driving cars and their drivers. Smart and partially self-driving cars such as the Tesla Sedan are permanently connected to the Internet, and as such can be considered to be large, fast, and dangerous Internet of Things (IoT) devices. This is particularly the case when they are being driven by amateur rather than professional drivers, who have not received special training in the handover procedures which exist between self-driving autopilot systems and human drivers.

While private companies that operate self-driving cars are typically still required by law to have a human driver present in the car, humans are able to intervene only in order to prevent an accident, or when some other error occurs. However, the time required for a person to transition from an "automated state" to a "manual state" where the driver has full control remains relatively high, typically taking around 10 seconds for basic reassertion of control (Kyriakidis et al., 2017; Lu, Coster, & de Winter, 2017; Merat, Jamson, Lai, Daly, & Carsten, 2014), and around 30–40 seconds "to resume adequate control" (Merat et al., 2014, p. 281). At the same time, it is widely acknowledged that requiring human drivers to take control of the vehicle is not a good solution, as "disengagement of automation is not a particularly practical method for keeping drivers in the loop" (Merat et al., 2014, p. 274).

Thus, it seems reasonable to argue that human drivers are not necessarily fully in the loop, simply because their presence is required in a technical system. Rather, it can be argued that the presence of drivers in self-driving cars is to assure the public that their safety is being taken care of, and that a specific person will be liable in the event of an accident. Because current research suggests that individual response times are considerable, it has to be asked how useful a driver would actually be in the case of an emergency. If an effective response requires a reaction time of less than 10 seconds, this would seem quite difficult for a human driver to achieve. This is particularly the case as the errors that are produced by automated systems are not the same as those produced by humans and, consequently, are very difficult for humans to predict. Machine errors in automated driving are typically associated with the correct identification of the objects perceived by sensors, while human driving errors are typically associated with being able to provide sustained attention to a specific task over long periods of time. It is therefore perhaps unsurprising that frequent users of automated driving technology mostly blame other humans rather than their own automated systems for any accidents they experience while driving (Levin, 2016; Luckerson, 2015). However, other reports suggest that at least some of the accidents occurred while the cars were in self-driving mode (Pritchard, 2015). There is even some

suggestion that the companies involved blame human drivers, even when the algorithm is actually at fault (Levin, 2016; Luckerson, 2015).

In response to these challenges, some authors have proposed alternative solutions to the challenge of human/computer agency that would increase rather than decrease the agency of individual drivers. To give an example of such an alternative, the integration of an "ethical knob" (Contissa, Lagioia, & Sartor, 2016) would allow drivers to decide how self-interested or how community-interested (i.e., the ethical model: egotistical or communitarian) they would like an automated vehicle to be. The vehicle would then make decisions which were more altruistic or egoistic based entirely on the ethical choices of the driver. Such a solution could "giv[e] back to the passenger the moral decisions and the judgement as to which outcome is more acceptable" (Contissa et al., 2016, p. 9). However, this approach also ensures that the dominant frame of plausible agency remains focused on the driver, leaving the associated assignment of liability clearly focused on the human driver. Admittedly, the integration an ethical knob within an IoT device at least ensures that humans have a greater level of agency more closely aligned to the level of personal liability that they have in the process.

Police Searches Based on Passenger Name Records and Social Media Data

Another interesting example of the kinds of human intervention often involved in automated systems concerns police officers conducting passenger searches based on algorithms used to analyze passenger name records (PNRs) and social media data. Such automated systems are used to identify potential criminals on flights using the PNRs and other typically publicly available data of the individuals on a flight, such as social media data purchased from private companies to enrich the PNRs. The identification is not made with 100 percent certainty; rather, the algorithms generate statistical probabilities of individuals being likely to be part of a certain group of criminals, and individual police officers then decide at which threshold of probability they wish to check whether the hypothesis made by an automated system is in fact correct. Based on passengers' PNR data and other publicly available social media data, particular passengers might then be selected by police officers for further inspection at the border. This decision-making process does however serve to blur the boundary between whether police officers themselves make the decision to search an individual suspected of drug smuggling or terrorism, or whether the decision is made by the algorithm. While the officers are free to do whatever they want in theory, in the context of the socio-technical system they are embedded in, they will *de facto* end up doing certain things that the algorithm suggests.

At an initial stage, the decision-making process seems relatively clear. Border agents are given computer output, which they are asked to interpret and based on which they make an informed decision (Saunders, Hunt, & Hollywood, 2016). However, such an understanding of the socio-technical system mistakes the wider role of the police officers engaged in complex information technology systems.

The question that needs to be asked in this context is what the legal ramifications are for an individual police officer in a specific legal jurisdiction of receiving a recommendation by the PNR system. Is the police officer required to interpret the response of the automated system as a "tip" and investigate it? Or are they able to ignore the results of the system if they believe the individual to not pose a threat or to have been falsely identified?

In general, the amount of time which border guards have to ascertain whether a traveler is a threat or not is relatively short. For example, the European Union (EU) Border agency Frontex suggests that an "EU border guard has on average just 12 seconds to decide whether the traveler in front of them is legitimate or not" (Fergusson, 2014, p. 15). It seems highly unlikely that border guards would have significantly longer to assess the results of an automated system. Furthermore, based on conversations with individuals knowledgeable about the matter, it seems highly unlikely that a police officer would not follow any individual leads provided by the automated system (Karppi, 2018). This is in part because police forces use individual searches to calibrate the system and ensure that they are targeting the right groups or individuals. Thus, each search is not just a search for the purpose of finding criminal activity, it also contributes to testing a police hypothesis about likely criminals, for which both positive and negative responses are important to validate the hypothesis (Kirkpatrick, 2017).

In consequence, if an individual police officer is not actively able to ignore a specific individual recommendation of an algorithm to search a person at the border, this also means that the decisions made by the algorithm are *de facto* automated. While the "human in the loop" is of course necessary to conduct a search, the police officer involved also ends up becoming liable for all decisions made by that algorithm, because—at least formally—they were made by a human (Degeling & Berendt, 2017; Ensign, Friedler, Neville, Scheidegger, & Venkatasubramanian, 2017). Thus, the decision to search an individual is essentially made in a quasi-automated fashion, which is essentially automated and includes a human in the loop who does not have active agency at an individual level. If the automated system makes a mistake in correctly or incorrectly identifying a criminal, its results are likely to be optimized on this basis. However, there is currently no framework in which software developers could be held liable for the errors made by this system. By contrast, police officers are held liable personally and directly for any mistakes made.[7]

What could be considered a counter-balance in this context is that at least some of the police officers involved are able to adapt the algorithm based on its effectiveness.[8] By working out when it is effective and when it is not, police are more easily able to refine the algorithm, increase its effectiveness, and reduce the number of false positives it creates. At the same time, this form of agency is only after the fact and would not influence the decision making of actual police officers in individual cases. Here they are bound to conduct each search that the algorithm suggests, if only to decide as a result that the algorithm is indeed working or not.

Outsourced Facebook Content Moderation

Another key system that employs humans in the loop is Facebook's content moderation system. Facebook employs numerous filters and algorithms to decide what content individual users should see, as a result of which a large part of the Facebook platform is automated. However, some parts of the system that Facebook uses also include human input. This involves several Facebook content moderation teams around the world who are responsible for responding to complaints made by users about the content that they see on the platform. There is also another group of humans who are responsible for filtering out images which are considered to be against Facebook's community standards. For this purpose, Facebook employs a series of call centers to go through problematic images uploaded to the platform which have been flagged by users. There has been much debate about the lack of legal basis for Facebook's content moderation decisions, as "breastfeeding mothers are off limits [but] crushed heads, limbs, etc., are OK as long as no insides are showing" (Anderson, 2012). Facebook makes almost all content moderation decisions based on its terms of service and not based on the law of any given country (Wagner, 2016). The automation or nonautomation of content moderation decisions on the Internet's largest social media platform raises a key question for the governance of Internet architecture. By changing the design of this large social media platform (Wagner, 2018c), it may also be possible to better promote, protect, and uphold human rights—or to harm them.

However, the role of these particular humans in the loop has previously received far less attention than the online platform itself. While described by some as the "trash collectors on the Internet" (Bouhs, 2016), they often work in very difficult conditions with low pay and are typically given only a few seconds to decide about each piece of content (Anderson, 2012; Bouhs, 2016; Roberts, 2014). So while algorithms have automatically identified pieces of content of concern, the people tasked with confirming this barely have time to make a meaningful decision. As the workers are paid very badly, these jobs do not attract skilled workers and particularly not those who have any kind of legal or technological qualifications. While they are certainly able to learn on the job, their work consists of making quick decisions at the behest of a computer algorithm—which are in turn checked for validity and consistency by another computer algorithm. In this sense, they are doing multiple rounds of coding of a large data set in order to ascertain which images can stay and which can not. This is very different from substantively deciding what content is placed on Facebook and what is not (Wagner, 2016).

This raises the interesting question of why Facebook includes them in the system at all. Part of the reason may be that they fulfill a function that cannot yet be automated, although both Google and Facebook have acknowledged that they are taking extensive steps to replace this kind of work with automated systems (Wagner, 2016). However, it can also be suggested that much of their work is not actually there to contribute to human decision making, but rather to suggest that

humans—both the users of Facebook and the staff at Facebook—actually have agency.

This suggestion comes from the primarily algorithmic construction of the system by which Facebook presents individual cases to human reviewers for review. This system has several components in its decision making, which include:

1. How many users have complained about a piece of content;
2. What they have complained about, or;
3. How urgent Facebook considers the complaint;
4. Whether the media is reporting about the case yet, and;
5. Whether the complaint comes from a privileged source such as a law enforcement officer or judge.

At the same time, the system of escalation is a complete black box to Facebook users and the general public. They have no way of knowing whether or not Facebook will respond to an individual complaint, how it will respond, and whether a human will be tasked with any such response. While there are currently debates about whether the new EU GDPR may provide a "right to explanation" for algorithmic decisions (Goodman & Flaxman, 2016), it is entirely unclear what this right would actually look like in practice (Wachter, Mittelstadt, & Floridi, 2016), and whether it would be fit for the purposes discussed here (Edwards & Veale, 2017). Despite this, it is important to acknowledge that the GDPR provides numerous additional rights to users which should contribute to shedding greater light on algorithmic decision making (Article 29 Data Protection Working Party, 2017). However, it should be noted that these rights to explanation only apply to automated decision making. Thus, it is conceivable that by involving a "human in the loop," companies could avoid this right to explanation. The Article 29 Working Party acknowledges this and—similarly to the proposal in this article—is concerned about human decisions as "rubber stamping" or quasi automation. To avoid this, they suggest a set of criteria similar to the ones proposed below; that is, to ensure meaningful "human intervention, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture" (Article 29 Data Protection Working Party, 2017, p. 10).

Numerous anecdotal reports in this context suggest that dozens of users need to complain about a single piece of content for any action at all to be taken by a human at Facebook. Aside from very urgent cases like self-harm or child sexual abuse material, Facebook typically waits to see how many users take issue with a specific piece of content before sending it to be checked by a Facebook staff member. This threshold also seems to be based on an automated algorithm, allowing an automated system to decide (based on the number of complaints) which piece of flagged content is even seen by an individual reviewer, before they are given a chance to make a decision on it. It should also be noted that the current working conditions that many content moderators work in—particularly

as part of outsourcing and subcontractor arrangements—are themselves likely to violate numerous human rights (Riesewieck, 2017).

Thus, it seems reasonable to argue that Facebook includes humans in the loop for a variety of different reasons. While this might suggest the presence of human agency at some levels, humans are so deeply embedded in the algorithmic systems that decide things around them that it would be hard to call them actual decision makers *per se*. Rather, the analysis seems to indicate they are actually embedded there to suggest a limited degree of human agency within a highly complex system. While human labor may be necessary in some areas, part of the reason for human involvement seems to suggest human agency when there is actually very little.

Challenges Faced by the "Human in the Loop" Model

Quasi-Automation as a Challenge

The uncomfortable reality about much of the discussion about safeguarding human rights in Internet architecture by putting a human in the loop is that the debate ignores many of the vast number of cases in which significant automation is taking place, as long as somewhere in the process a human is still "in the loop." Indeed, many supposedly automated or artificial intelligence (AI) systems can function only because of the numerous humans who serve to fix the mistakes that technical systems make, or who replace the decisions or technical systems completely (Ekbia & Nardi, 2014; Kushner, 2013).

Yet, such rules are deceptively simple and do not do justice to the wide scale of diverse human-technology interactions that already exist. However, existing legal rules that, for example, forbid or allow certain forms of automation do so on the assumption that a "human in the loop" means that an actual human "check" will take place of the results of the automated system. If the person is able to only rubber-stamp the results produced by the algorithm, then these systems should perhaps more accurately be called "quasi-automated."

This is particularly the case when the company involved spends little time or energy ensuring that staff are properly trained or prepared to make these decisions, or that they have sufficient time to make the decision themselves. This is most evident in the case of Internet content moderators, in which low-paid workers in the Philippines deal with thousands of cases of content moderation per day for a few dollars per hour (Riesewieck, 2017). However, it is also evident in the case of self-driving car drivers, where the design of the system puts drivers in a situation where they evidently do not have sufficient time to decide. This is particularly obvious in the case of amateur self-driving car drivers, who have neither the relevant training nor the experience to understand the specific types of mistakes made by automated self-driving systems.

Thus, in order for the human in the loop "claim" to be able to actually promote human rights infrastructure, it needs to ensure that it avoids some of the pitfalls discussed above. Similar decisional mechanisms already exist in the area of aviation, where a specific classification of levels of automation exists.[9] A similar

classification has also been developed by the Society of Automotive Engineers for self-driving cars (Milakis, van Arem, & van Wee, 2017). What is still under considerable discussion is how to govern the different levels of automated driving systems. In particular, in regard to many of the assumptions associated with including a human in the loop, there is a need to ensure that human beings are not used to simply rubber-stamp automated decisions in order to ensure that technical systems meaningfully promote human rights. Another similar classification has been developed by the Article 29 Working Party which suggests that "To qualify as human intervention, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision. As part of the analysis, they should consider all the available input and output data" (Article 29 Data Protection Working Party, 2017, p. 10).

Based on these assumptions, I suggest the following criteria could be used to more clearly define cases of quasi-automation:

1. **Amount of time** which the individual human operator has in relation to the task: the lower the amount of time assigned to the human operator, the more likely it is to be quasi-automated.
2. **Degree of qualification** that the human operator of the system has to fulfill the specific task: the less qualified the individual is to fulfill a specific task, the more likely it is to be quasi-automated.
3. **Degree of liability** which will be assigned to the human operator if it fails: the greater the amount of legal liability a human operator is assigned for failure, the more likely that humans are engaged in the process merely to ensure they can take liability if the automated system fails (Dannenbaum, 2010; Maurino, Reason, Johnston, & Lee, 2017).
4. **Level of support** the individual receives to conduct the task sustainably: many of the tasks involved require very high degrees of concentration and often involve making highly disturbing decisions in a short space of time. Here, higher levels of psycho-social support or other forms of support are likely to be an indicator that speaks against quasi-automation.
5. **Adaption:** the more a human operator must adapt to the system, instead of the system being designed to serve the operator, the more the system is quasi-automated.
6. **Access to information:** the human operator must have access to all relevant information in order to be able to make the correct decision.
7. **Agency:** the human operator should have sufficient "authority [. . .] to change the decision" (Article 29 Data Protection Working Party, 2017, p. 10) and should actually do so on a regular basis. If the only function of the human operator is to regularly agree with the machine and only very rarely disagrees with it, it is highly likely that the human operator's agency is insufficient.

Only if a socio-technical system does not engage in quasi-automation with a human in the loop (i.e., where they have responsibility but little agency), but

actually systematically gives humans autonomy in making decisions—only then can human rights be safeguarded.[10] Human discretion cannot exist without basic minimal conditions for that discretion to be exercised.

Assumption of Binary Liability as a Challenge

Another challenge related to understanding the role of human agency in socio-technical systems is the assumption of binary liability. In the binary liability model, either a human or a machine must necessarily be at fault, which in turn links to a social argument about the need to blame someone (Marchant & Lindor, 2012). However, this model of social blame translated into binary legal liability is unfit for a world of human-technical systems in which both equally contribute to decision making. The fact that humans are in the loop should not absolve automated systems—as is frequently currently the case—from being scrutinized legally. For example, Article 22 of the EU GDPR includes a "prohibition on fully automated individual decision-making, including profiling that has a legal or similarly significant effect" (Article 29 Data Protection Working Party, 2017, p. 9), as well as specific safeguards for fully automated decisions.

Thus, by putting humans in the loop, companies can try to evade some of these limitations and safeguards. Moreover, many companies currently use human intervention in supposedly fully-automated systems—that is, to pretend an advanced AI system is transcribing your voicemail rather than a call center in the Philippines—without telling users that this is the case (Solon, 2018; Wray, 2009).

Liability is important to consider in human rights terms in regard to remedying the harms created by technical systems (Shelton, 2015). For example, as noted by United Nations Special Rapporteur on Freedom of Expression, David Kaye, in his 2018 report to the Human Rights Council, we do not yet know what such a "remedy [could] look like in the digital age" (Kaye, 2018). People cannot be blamed or held to account for just their discretion alone: the discretion of technical systems needs to be considered in equal measure to ensure that—beyond binary liability—the right to redress of individuals in the case of rights violations is adequately considered. Here, existing redress and remedy frameworks need to go beyond simply blaming a specific individual and providing damages, but rather instead need to ensure that the social-technical procedures that led to the rights violation are systematically changed.

One area where this area of regulation has developed considerably is in regard to aviation, where less binary approaches to liability have developed over time. This approach is encapsulated in Regulation (EU) No. 376/2014 on the reporting, analysis, and follow-up of occurrences in civil aviation, which calls for a "just culture" which is "a culture in which front-line operators or other persons are not punished for actions, omissions or decisions taken by them that are commensurate with their experience and training, but in which gross negligence, willful violations, and destructive acts are not tolerated."[11] This approach—also evident in several legal decisions made in the aviation

sector, which focus on organizational liability (Brüggemeier, 1991) rather than on individual liability—is increasingly common in court rulings (Schebesta, 2017), which focus on improving socio-technical systems at scale rather than on just identifying who is at fault.

This is particularly important in regard to specific limitations that currently exist on the employment of fully automated systems in different areas of society. From online content moderation to self-driving cars, there are a specific set of rules for these types of automated decision making that apply only if there is no human in the loop at all. As argued here, these rules for full automation should also be expanded to situations of "quasi-automation," where the human in the loop barely has the time or the qualifications to make an informed decision. In principle, there are some similarities between self-driving cars and the aviation industry, where complex socio-technical systems made by the combination of humans and auto-pilots are evaluated differently than those involving manually piloted aircraft (Schebesta, 2017).[12]

Based on the suggestions made above, certain minimum standards need to be set under which having a human in the loop no longer has an exculpatory function which allows companies to pretend that the systems are under active human control, rather than being essentially automated or quasi-automated. Such a basic assumption by policymakers—it is hoped—could lead companies to think differently about the way in which they treat and value those humans. While the function of such "looped-in" individuals might become less relevant as their ability to shield automated systems from liability is diminished, the employers could also choose to give them more time, proper training, and essentially greater levels of agency to avoid having their automated systems classified as quasi-automated systems (i.e., where the human is clearly just a rubber stamp). If the criteria followed above—in which employees have enough time, proper training, all relevant information, etc.—are systematically followed, it could be expected that this would also have a positive effect not just on the ability of these individuals to safeguard human rights, but also on the working conditions of these individuals themselves. Ensuring that they are able to do their jobs effectively also means ensuring that they are able to do so in a manner that does not violate their human rights to safeguard the rights of others.

More broadly, the analysis suggests that there is a need for a greater level of agency for humans in the loop in order to ensure that the liability exemption linked to their presence in the decision-making chain is not misused (Dannenbaum, 2010). This would mean that humans with greater agency also would need to accept greater liability and, at the same time, humans with low levels of agency could not be expected to shoulder high levels of liability (Crootof, 2016). While this may not directly be linked to the personal liability of the humans themselves, liability exemptions are still a tangible legal benefit granted to certain types of organizations. This should be the case only if any humans in the loop have a meaningful human decision (Roff & Moyes, 2016) to make and not if they are simply rubber-stamping the decisions already made by automated systems.

Rather disconcertingly, it is perfectly possible to interpret this article as an acknowledgment that sole human responsibility is a "mental crutch" which supports and is supported by existing ethical and legal notions about responsibility, both legal and moral, whereas the reality of quasi-automated decisions is far more messy and diverse (de Sio & Di Nucci, 2016). Importantly, the call for more human agency in socio-technical systems can only meaningfully contribute to human rights, if it is meaningful human agency and not simply a "quasi-automation" or rubber-stamping of automated decisions. This rubber-stamping of automated decisions is likely to harm human rights, as it obscures who is actually at fault while simultaneously ensuring that the actual decision maker (the automated system) does not change, given it is not recognized as the responsible party.

It is only by acknowledging that humans are not always fully agents or fully autonomous that it is possible to shift the focus of the debate on these topics to nonhuman agents which also possess forms of agency (Brey, 2005). Only by shifting the debate away from whether humans or technology are liable for mistakes—and by acknowledging that both, to a degree, are engaged in decision making and therefore also responsible for their decisions—can a more meaningful understanding of socio-technical agency be gained.

Conclusion and Paths Ahead

What conclusions can be drawn from this article? When studying human rights in Internet infrastructure, the role of decision making is seldom looked at in greater detail. This article argues that there is indeed a need for greater human decision making in Internet infrastructure, but this will only be helpful if it constitutes meaningful human decision making. If, by contrast, the rubber stamping of automated decisions increasingly takes hold in a phenomenon described here as "quasi-automation," this is likely to weaken rather than strengthen human rights. These type of quasi-automated decisions contribute to confusions around binary liability which make it more difficult for individuals to seek redress and claim access to rights, such as the GDPR's right to explanation.

At worst, keeping a human being in the loop may serve as a human fig-leaf for automated decisions made by algorithms and, thus, may even be detrimental to developing human rights Internet architecture (Wagner, 2018b). However, at best, if safeguards are implemented to prevent quasi-automation, ensuring meaningful human discretion can contribute to developing a more effective human rights Internet architecture. Once this has been clarified it may be possible to implement this systematically across both policy and technical systems. Pretending that human beings make decisions when they actually do not will not contribute to anyone's human rights and as argued here may actually harm them.

Ben Wagner, Ph.D., Assistant Professor and Director of the Privacy and Sustainable Computing Lab, Vienna University of Economics and Business, Austria [ben.wagner@wu.ac.at].

Notes

1. They need this legal and technical neutrality in order to continue to be legally classified as the managers of information platforms (like an Internet service provider) and not editors of an information service (like the New York Times).
2. The term "human rights Internet infrastructure" is used throughout this article. In the way it is used here this is meant to refer to Internet infrastructure which upholds, enables and promotes human rights, both as a legal framework and a set of values.
3. https://tools.ietf.org/html/draft-irtf-hrpc-guidelines-00.
4. For further details, see https://www.coe.int/en/web/freedom-expression/committee-of-experts-on-internet-intermediaries-msi-net.
5. See https://rm.coe.int/algorithms-and-human-rights-study-on-the-human-rights-dimension-of-aut/1680796d10.
6. See https://rm.coe.int/algorithms-and-human-rights-study-on-the-human-rights-dimension-of-aut/1680796d10.
7. This is the current state of the law in Europe, according to my conversations with people implementing the PNR regime in several EU countries, and with those involved in negotiating the PNR agreements.
8. The police officers who run these systems are able to calibrate it and insert additional variables, for example, by adding an extra variable into the algorithm ("ticket paid in cash") and seeing how effective it is ("14 drug smugglers caught today compared with only six yesterday").
9. See https://www.faa.gov/training_testing/training/fits/research/media/Det_App_Lvl_Atm.pdf.
10. Beyond the scope of the civilian context discussed here, a related concept of "meaningful human control" has been developed in a military context by Heather Roff and Richard Moyes (2016). For further details see http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf.
11. See http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX: 32014R0376&from.
12. Regulation is much stricter for aviation than for self-driving cars. For example, there is no requirement for black boxes in cars (as there is for airplanes), and the authorities often have difficulty accessing all relevant data.

References

Anderson, S. 2012. "Why Did Facebook Censor This Photograph?" *Foreign Policy*. https://foreignpolicy.com/2012/11/14/why-did-facebook-censor-this-photograph/.

Article 29 Data Protection Working Party. 2017. G*uidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679*. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053.

Billings, C.E. 1991. *Human-Centered Aircraft Automation: A Concept and Guidelines*. https://ntrs.nasa.gov/search.jsp?R=19910022821.

Bouhs, D. 2016. "Hasskommentare Auf Facebook & Co—Die Online-Müllabfuhr Des Internets." *Deutschlandfunk*. http://www.deutschlandfunk.de/hasskommentare-auf-facebook-co-die-online-muellabfuhr-des.761.de.html?dram:article_id=352951.

Brey, P. 2005. "Artifacts as Social Agents." In *Inside the Politics of Technology: Agency and Normativity in the Co-Production of Technology and Society*, ed. H. Harbers. Amsterdam, the Netherlands: Amsterdam University Press, 61–84.

Brüggemeier, G. 1991. "Organisationshaftung: Deliktsrechtliche Aspekte Innerorganisatorischer Funktionsdifferenzierung." *Archiv Für Die Civilistische Praxis* 191 (H. 1/2): 33–68.

Cath, C., and L. Floridi. 2017. "The Design of the Internet's Architecture by the Internet Engineering Task Force (IETF) and Human Rights." *Science and Engineering Ethics* 23 (2): 449–68.

Contissa, G., F. Lagioia, and G. Sartor. 2016. "The Ethical Knob. SSRN Scholarly Paper." ID 2881280. Rochester, NY: Social Science Research Network.

Crane, D.A. 2011. "Search Neutrality as an Antitrust Principle." *George Mason Law Review* 19: 1199.

Crootof, R. 2016. "A Meaningful Floor for Meaningful Human Control." *Temple International & Comparative Law Journal* 30: 53.

Dannenbaum, T. 2010. "Translating the Standard of Effective Control Into a System of Effective Accountability: How Liability Should Be Apportioned for Violations of Human Rights by Member State Troop Contingents Serving as United Nations Peacekeepers." *Harvard International Law Journal* 51: 113.

Degeling, M., and B. Berendt. 2017. "What Is Wrong About Robocops as Consultants? A Technology-Centric Critique of Predictive Policing." *AI & Society* 33 (3): 347–56.

Denardis, L. 2011. "The Privatization of Internet Governance." Paper presented at the Fifth GigaNet Annual Symposium, September 13, 2010, Vilnius, Lithuania. https://www.giga-net.org/2010-annual-symposium/.

Edwards, L., and M. Veale. 2017. "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For." SSRN Scholarly Paper. ID 2972855. Rochester, NY: Social Science Research Network.

Ekbia, H., and B. Nardi. 2014. "Heteromation and Its (Dis)Contents: The Invisible Division of Labor Between Humans and Machines." *First Monday* 19 (6).

Ensign, D., S.A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. 2017. *Decision Making With Limited Feedback: Error Bounds for Recidivism Prediction and Predictive Policing*. https://pdfs.semanticscholar.org/03d9/cc7e2750bcf84d6e26292b4ae13245c55470.pdf.

Fergusson, J. 2014. *Twelve Seconds to Decide in Search of Excellence: Frontex and the Principle of 'Best Practice*.' Luxembourg: Publications Office.

Fitts, P.M. 1951. *Human Engineering for an Effective Air-Navigation and Traffic-Control System*. Oxford, England: Division of National Research Council.

Gill, L., D. Redeker, and U. Gasser. 2015. "Towards Digital Constitutionalism? Mapping Attempts to Craft an Internet Bill of Rights." Berkman Center Research Publication No. 2015-15. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2687120.

Gillespie, T. 2014. "The Relevance of Algorithms." In *Media Technologies: Essays on Communication, Materiality, and Society*, eds. T. Gillespie, P.J. Boczkowski, and K.A. Foot. Cambridge, MA: MIT Press, 167–94.

Goodman, B., and S. Flaxman. 2016. "European Union Regulations on Algorithmic Decision-Making and a Right to Explanation." In *2016 ICML Workshop on Human Interpretability in Machine Learning*. New York, NY: ArXiv e-prints.

Irani, L. 2015. "Difference and Dependence Among Digital Workers: The Case of Amazon Mechanical Turk." *South Atlantic Quarterly* 114 (1): 225–34.

Johnson, D.R., and D.G. Post. 1996. "Law and Borders—The Rise of Law in Cyberspace." *First Monday* 1 (1).

Karppi, T. 2018. "'The Computer Said So': On the Ethics, Effectiveness, and Cultural Techniques of Predictive Policing." *Social Media+Society* 4 (2): 2056305118768296.

Kaye, D. 2018. "Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression." A/HRC/38/35. Geneva, Switzerland: United Nations.

Kelsey, J.T.G. 2008. "Hacking Into International Humanitarian Law: The Principles of Distinction and Neutrality in the Age of Cyber Warfare." *Michigan Law Review* 106 (7): 1427–52.

Kirkpatrick, K. 2017. "It's Not the Algorithm, It's the Data." *Communications of the ACM* 60 (2):21–3.

Kitchin, R. 2017. "Thinking Critically About and Researching Algorithms." *Information, Communication & Society* 20 (1): 14–29.

Kitchin, R., and M. Dodge. 2011. *Code/Space Software and Everyday Life*. Cambridge, MA: MIT Press.

Kushner, S. 2013. "The Freelance Translation Machine: Algorithmic Culture and the Invisible Industry." *New Media & Society* 15 (8): 1241–58.

Kyriakidis, M., J.C.F. de Winter, N. Stanton, T. Bellet, B. van Arem, K. Brookhuis, M.H. Martens et al. 2017. "A Human Factors Perspective on Automated Driving." *Theoretical Issues in Ergonomics Science* 0 (0): 1–27.

Levin, S. 2016. "Uber Blames Humans for Self-Driving Car Traffic Offenses as California Orders Halt." *The Guardian*. https://www.theguardian.com/technology/2016/dec/14/uber-self-driving-cars-run-red-lights-san-francisco.

Lu, Z., X. Coster, and J. de Winter. 2017. "How Much Time Do Drivers Need to Obtain Situation Awareness? A Laboratory-Based Study of Automated Driving." *Applied Ergonomics* 60: 293–304.

Luckerson, V. 2015. "Google Blames Humans for Accidents Involving Its Self-Driving Cars." *Time*. http://time.com/3854528/google-self-driving-cars-accidents/.

Marchant, G.E., and R.A. Lindor. 2012. "The Coming Collision Between Autonomous Vehicles and the Liability System." *Santa Clara Law Review* 52: 1321.

Mathew, A.J. 2014. *Where in the World Is the Internet? Locating Political Power in Internet Infrastructure*. PhD Dissertation, University of California, Berkeley.

Maurino, D.E., J. Reason, N. Johnston, and R.B. Lee. 2017. *Beyond Aviation Human Factors: Safety in High Technology Systems*. Abingdon, Oxon; New York, NY: Routledge.

Merat, N., A.H. Jamson, F.C.H. Lai, M. Daly, and O.M.J. Carsten. 2014. "Transition to Manual: Driver Behaviour When Resuming Control From a Highly Automated Vehicle." *Transportation Research Part F: Traffic Psychology and Behaviour* 27 (Part B): 274–82.

Milakis, D., B. van Arem, and B. van Wee. 2017. "Policy and Society Related Implications of Automated Driving: A Review of Literature and Directions for Future Research." *Journal of Intelligent Transportation Systems* 21 (4): 324–48.

Mueller, M. 2010. *Networks and States*. Cambridge, MA: MIT Press.

Pasquale, F. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Pasquale, F.A. 2016. "Platform Neutrality: Enhancing Freedom of Expression in Spheres of Private Power." SSRN Scholarly Paper. ID 2779270. Rochester, NY: Social Science Research Network.

Pritchard, J. 2015. "Google Acknowledges 11 Accidents With Its Self-Driving Cars." *Associated Press*. http://bigstory.ap.org/article/297ef1bfb75847de95d856fb08dc0687/ap-exclusive-self-driving-cars-getting-dinged-california.

Riesewieck, M. 2017. *Digitale Drecksarbeit wie uns Facebook & Co. von dem Bösen erlösen*. Munich: Dtv.

Roberts, S.T. 2014. *Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation*. PhD Dissertation, University of Illinois at Urbana-Champaign.

Roberts, S.T. 2016. "Commercial Content Moderation: Digital Laborers' Dirty Work." In *The Intersectional Internet: Race, Sex, Class and Culture Online*, eds. S.U. Noble and B. Tynes. New York: Peter Lang Publishing, 147–60.

Roff, H.M. and R. Moyes. 2016. "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons." Briefing Paper Prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons, Geneva, Switzerland.

Saunders, J., P. Hunt, and J.S. Hollywood. 2016. "Predictions Put Into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot." *Journal of Experimental Criminology* 12 (3): 347–71.

Schebesta, H. 2017. "Risk Regulation Through Liability Allocation: Transnational Product Liability and the Role of Certification." *Air and Space Law* 42 (2): 107–36.

Shelton, D. 2015. *Remedies in International Human Rights Law*. Oxford: Oxford University Press.

Sheridan, T.B. 2000. "Function Allocation: Algorithm, Alchemy or Apostasy?" *International Journal of Human-Computer Studies* 52 (2): 203–16.

de Sio, F.S., and E. Di Nucci. 2016. "Drones and Responsibility." In *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on Remotely Controlled Weapons 1*, eds. F.S. de Sio and E. Di Nucci. London: Routledge, 1–14.

Solon, O. 2018. *The Rise of "Pseudo-AI": How Tech Firms Quietly Use Humans to Do Bots' Work*. http://www.theguardian.com/technology/2018/jul/06/artificial-intelligence-ai-humans-bots-tech-companies.

Suzor, N. 2018. "Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms." *Social Media + Society* 4 (3): 1–11.

Wachter, S., B. Mittelstadt, and L. Floridi. 2016. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." SSRN Scholarly Paper. ID 2903469. Rochester, NY: Social Science Research Network.

Wagner, B. 2012. "After the Arab Spring: New Paths for Human Rights and the Internet in European Foreign Policy." Brussels, Belgium: European Union.

Wagner, B. 2016. *Global Free Expression: Governing the Boundaries of Internet Content*. Cham, Switzerland: Springer International Publishing.

Wagner, B. 2018a. "Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regularoy Implications." DGI(2017)12. Strasbourg, France: Council of Europe.

Wagner, B. 2018b. "Ethics as an Escape From Regulation: From Ethics-Washing to Ethics-Shopping?" In *10 Years of Profiling the European Citizen*, eds. E. Bayamlioglu, I. Baraliuc, L.A.W. Janssens, and M. Hildebrandt. Brussels, Belgium: Vrije Universiteit Brussel (VUB), 108–15.

Wagner, B. 2018c. "Free Expression?—Dominant Information Intermediaries as Arbiters of Internet Speech." In *Digital Dominance: Implications and Risks*, eds. M. Moore and D. Tambini. Oxford: Oxford University Press, 219–40.

Wolverton, T. 2018. "Mark Zuckerberg Says AI Won't Be Able to Reliably Detect Hate Speech for 'Five to 10' Years." *Business Insider Deutschland*. https://www.businessinsider.de/facebook-ceo-zuckerberg-says-hate-speech-stumps-ai-2018-4.

Wray, R. 2009. "SpinVox Boss Defends Her Company Against BBC's Allegations." *The Guardian* (July 23).