Coppi *et al. Journal of International Humanitarian Action* (2021) 6:19
https://doi.org/10.1186/s41018-021-00096-6

Journal of International
Humanitarian Action

RESEARCH ARTICLE Open Access

Explicability of humanitarian AI: a matter of principles

Giulio Coppi[1*], Rebeca Moreno Jimenez[2] and Sofia Kyriazi[2]

Abstract

In the debate on how to improve efficiencies in the humanitarian sector and better meet people's needs, the argument for the use of artificial intelligence (AI) and automated decision-making (ADMs) systems has gained significant traction and ignited controversy for its ethical and human rights-related implications.

Setting aside the implications of introducing unmanned and automated systems in warfare, we focus instead on the impact of the adoption of AI-based ADMs in humanitarian response. In order to maintain the status and protection conferred by the humanitarian mandate, aid organizations are called to abide by a broad set of rules condensed in the humanitarian principles and notably the principles of humanity, neutrality, impartiality, and independence. But how do these principles operate when decision-making is automated?

This article opens with an overview of AI and ADMs in the humanitarian sector, with special attention to the concept of algorithmic opacity. It then explores the transformative potential of these systems on the complex power dynamics between humanitarians, principled assistance, and affected communities during acute crises. Our research confirms that the existing flaws in accountability and epistemic processes can be also found in the mathematical and statistical formulas and in the algorithms used for automation, artificial intelligence, predictive analytics, and other efficiency-gaining-related processes.

In doing so, our analysis highlights the potential harm to people resulting from algorithmic opacity, either through removal or obfuscation of the causal connection between triggering events and humanitarian services through the so-called black box effect (algorithms are often described as black boxes, as their complexity and technical opacity hide and obfuscate their inner workings (Diakopoulos, Tow Center for Digital Journ, 2017). Recognizing the need for a humanitarian ethics dimension in the analysis of automation, AI, and ADMs used in humanitarian action, we endorse the concept of "explicability" as developed within the ethical framework of machine learning and human-computer interaction, together with a set of proxy metrics.

Finally, we stress the need for developing auditable standards, as well as transparent guidelines and frameworks to rein in the risks of what has been defined as humanitarian experimentation (Sandvik, Jacobsen, and McDonald, Int. Rev. Red Cross 99(904), 319–344, 2017). This article concludes that accountability mechanisms for AI-based systems and ADMs used to respond to the needs of populations in situation of vulnerability should be an essential feature by default, in order to preserve the respect of the do no harm principle even in the digital dimension of aid.

In conclusion, while we confirm existing concerns related to the adoption of AI-based systems and ADMs in humanitarian action, we also advocate for a roadmap towards humanitarian AI for the sector and introduce a tentative ethics framework as basis for future research.

* Correspondence: gcoppi@fordham.edu; giulio.coppi@nrc.no; g.coppi@hey.com
[1]Norwegian Refugee Council, Oslo, Norway
Full list of author information is available at the end of the article

In today's increasingly technological society […], human activity cannot be properly understood without making reference to technological artifacts, which complicates the ascription of moral responsibility, using the artefacts as means to an end […]. As we interact with and through these artifacts, they affect how we perceive reality, the decisions that we make and how we make them.

*Merel Noorman and Edward N. Zalta, "Computing and moral responsibility," The Stanford Encyclopedia of Philosophy (*Noorman and Zalta 2014*), http://plato.stanford.edu/archives/sum2014/entries/computing-responsibility*

[T]the instrumental conception of technology conditions every attempt to bring man into the right relation to technology. Everything depends on our manipulating technology in the proper manner as a means. We will, as we say, "get" technology "spiritually in hand." We will master it. The will to mastery becomes all the more urgent the more technology threatens to slip from human control. But suppose now that technology were no mere means, how would it stand with the will to master it?

*Heidegger, Martin. "The question concerning technology (W. Lovitt, Trans.) The question concerning technology: and other essays (pp. 3-35)." (*Heidegger 1977*).*

Terminological notes

In this article, the concepts of "explicability" and "explainability" are used interchangeably following the approach adopted by most relevant literature. The only exception is in the final section, where the text mostly uses the term "explicability" to align with a proposal from Floridi and Cowls (2019).

Throughout the article, we refer to artificial intelligence (AI) for automated decision-making systems (ADMs) only, and any reference to either of these concepts should be considered to involve the other, unless specified otherwise.

Finally, there is no universal agreement on a taxonomy of AI definitions. For the purpose of this article, we will consider the concept of transparency as referring to the technical model in a broad manner and thus encompassing all issues related to explicability and interpretability, in alignment with the EU Ethics Guidelines for Trustworthy AI (EU High-Level Expert Group on Artificial Intelligence 2019). We instead adopt a narrow concept of transparency in our proposed framework, restricting it to the disclosure required with regard to the AI system itself (transparency in communication and traceability), but we do not delve onto it as it remains outside of our current scope of work for this article.

Introduction: towards an AI ethics framework for humanitarian research

The collective excitement for the promise of information and communication technologies has caught humanitarian actors unprepared, but the sector has shown the capacity to take this challenge in stride. Aid actors had to face the exponential mass adoption rates of mobile phone technology and integrated cameras first, and a few years later, of mobile data connection systems for internet access (Technology diffusion dataset 2004).[1] They initially reacted refusing any formal endorsement of digital communication systems to then slowly pivot towards more institutional applications (ICRC, The Engine Room and Block Party 2017).

Organizations subjected information management to the same principles inspiring all core humanitarian processes (Raymond and Card 2015), but they also showed the incapacity to align their institutional policies to the pace of technological developments (Cardia et al. 2017). The whole aid sector had a very late moment of reckoning at the 2002 Symposium on Best Practices in Humanitarian Information Management and Exchange. The event officially sanctioned the importance of formalizing through policies the aspiration to foster evidence- and data-based decision-making (Van de Walle and Comes 2015). To fully understand the extent of such delay we shall mention that the term "business intelligence" was first introduced in 1865 (Miller Devens 1865), In 1989 it was then reframed by Howard Dresner to describe "concepts and methods to improve business decision making by using fact-based support systems," and finally become common usage in late 1990s (Cebotarean 2011).

The last two decades witnessed a change in attitude and pace towards digital solutions. This led the United Nations Office for the Coordination of Humanitarian Affairs in 2013 to officially propose the recognition of information during crises—and the corresponding ability to communicate—as a basic humanitarian need (Raymond and Card 2015). This resulted in a Cambrian explosion of digital transformation initiatives within the sector. In the 2019 ICT4D conference alone, 993 participants representing 415 organizations from 81 countries got together over several days in Kampala, Uganda, to discuss digital challenges and opportunities in applying digital solutions in relief contexts (ICT4D 2019).

To bring clarity and facilitate inter-sectorial coordination, on July 12, 2018, the UN Secretary-General (UNSG) António Guterres created the High-level Panel

[1] The datasets analyzed are available in the following Github repository: Horace Dediu; Comin and Hobijn (2004) Technology diffusion dataset. https://github.com/owid/owid-datasets/tree/master/datasets/Technology%20Diffusion%20-%20Comin%20and%20Hobijn%20(2004)%20and%20others. Accessed on 06 February 2020

on Digital Cooperation. The same year, the Panel produced a set of five principles set forth in the Secretary-General's Strategy on New Technologies: protect and promote global values, foster inclusion and transparency, work in partnership, build on existing capabilities and mandates, and be humble and continue to learn (United Nations 2018); the strategy also mandated the United Nations Innovation Network (UNIN) to expand their work on frontier technologies such as blockchain, AI, and data innovation. In June 2020, the UNSG launched the Roadmap for digital cooperation, which includes 8 key areas of action, including promoting trust and security in the digital environment, ensuring the protection of human rights in the digital era, and supporting global cooperation on artificial intelligence (United Nations 2020). The fact that AI deserved a dedicated action point on such a brief list should not come as a surprise. When receiving submissions and opinions, the Panel's recommendation on the topic elicited "hundreds of responses." Responders flagged existing or future challenges in implementation posed by persisting gaps in international coordination, collaboration, and governance (United Nations 2018). In particular, the report highlighted a lack of representation and inclusiveness in global discussions, as well as the absence of a global coordination platform to bring all the initiatives dedicated to AI ethics together. It also concluded with the recommendation that "life and death decisions should not be delegated to machines," in line with the UN Secretary-General's call for a global ban on lethal autonomous weapons systems (United Nations 2018).

In addition to the UNSG initiative, on November 2019, following a decision by its General Conference UNESCO embarked on a 2-year process to elaborate the first global standard-setting instrument on ethics of AI. For this purpose, the organisation started a multidisciplinary process and launched consultations with a wide range of stakeholders, including the scientific community, people of different cultural backgrounds and ethical perspectives, minority groups, civil society, government, and the private sector. A preliminary result of this process has been the creation of UNESCO's AI Decision Maker's Toolkit that enables decision makers to respond to the challenges and opportunities of AI. The toolkit also aims to provide elements of foresight, recommendations, implementation guides, model use cases, and capacity building resources to ensure the development of a human rights-based and ethical AI throughout the AI lifecycle and across stakeholder groups (UNESCO 2019).

Many actors in the humanitarian sector are participating in individual AI initiatives to advance the UN system agenda, but these efforts are scattered and lack transparency. A good practice for the public sector has now been set by the City of New York, who published a directory of all high-priority algorithmic tools currently in use by the city administration (NYC AMPO 2020). Lately, other non-traditional stakeholders have ventured into humanitarian work by setting their own principles or initiatives related to humanitarian AI, including large technology-related private sector companies. Unfortunately, in some cases, humanitarian actors with global or local mandates have been excluded from participating in the design of such initiatives. Most importantly - and paradoxically considering the humanitarian and ethical principles that should act as framework – the process left out vulnerable population from the co-design of these new initiatives.

Introduction to ADMs in humanitarian action

Automated decision-making (ADM) is the process of making a decision by automated means without any human involvement or supervision. These decisions can be based on factual data, as well as on digitally created profiles (personas) or inferred data (ICO 2020), which is often non-statistically representative. The use of ADMs has sparked heated debates on their implications on political, social, digital, and physical security (Brundage et al. 2018); on their application by armed forces in the conduct of warfare or in other situations of violence; and on their use in humanitarian action to assist and protect the victims of armed conflict (ICRC 2019) or in sensitive topics related to social and development justice, which usually involves automated individual profiling (ICO 2020). In this paper, we focus on the use in humanitarian action and notably on the implications of using ADMs and other AI-based systems for the respect of a principled approach to humanitarian response.

The critical questions raised by experts have not deterred several humanitarian organizations from partaking in a global effort to explore the advanced automation of basic data collection and analysis processes. Most current applications of these technologies can be reconducted to a few common trends: streamlining automated processes at scale, decreasing costs and times of reaction, removing human biases from operations, and preserving agency of people affected by crisis over their data.

Notable examples of these trends are the use of ADMs in humanitarian action for (a) anticipation or prediction of a certain outcome, usually related to crisis prevention, early warning, or preparedness; (b) semi-automated or fully automated decisions regarding migratory status and resettlement of vulnerable population, namely migrants, asylum seekers, and refugees; and (c) assistance provision, including automating targeting, cash assistance provision or other forms of humanitarian assistance based on mathematical formulas (OCHA 2020; Molnar and Gill 2018; Development Pathways 2018). ADMs commonly aim to speed-up certain processes/calculations and trigger an action or suggest a decision. They present a hightened level of risk when (a) the purpose is to target, separate, or distinct a person according to certain population/group characteristics (segregation) in order to automate partially or fully a process for the sake of improving efficiencies (e.g., provide loans, cash, insurance, legal sentencing, targeting of people according to vulnerabilities) and/or (b) when they completely replace human decision-making processes, and the outcome of their decision harms directly or indirectly humans. Only in a few cases, the use of machine learning pushes farther into the realm of modeling and tries to generate predictions, where push and pull factors of human displacement are used to model a real-life situation to understand cognitive choices (Kyriazi 2019) or attempt to distill human cognitive process behind decisions. Overall, not in all cases where the main outcome is automation or support for decision-making in humanitarian action, the factors that have led to that decision have been made transparent. This leads to building ADMs that are discriminatory, inscrutable, and misleading. However, some exceptions to this might be found in the humanitarian work, for example, in OCHA's catalog for predictive analytics in humanitarian action (OCHA 2019), setting a peer-review mechanism that aims to transparentize and scrutinize the building of such systems in the humanitarian sector.

It is important to denote that not all ADMs use AI-based mechanisms—some might use simple mathematical or statistical formulas (Development Pathways 2018) to support calculations for decisions. Similarly, not all AI-based systems are ADMs, as they are not supporting decisions or attempting to automate them. Nevertheless, for those ADMs that are based on AI—that either support or replace partially or fully humanitarian decisions—some systematic due diligence should be put in place, as they are as fallible as the processes led by humans and bring with them the risk of similarly catastrophic consequences. But chasing algorithmic automation carries an additional risk. Our analysis shows that the digitalisation of core functions influencing decision-making processes can have significant—and potentially disruptive—impact on the nexus between humanitarian ethics and the implementation of humanitarian action, in addition to the potential negative impacts on the rights of individuals (Greenwood et al. 2017). This risk is even more poignant as newer systems aim to go beyond the automation of core existing models, announcing the progressive establishment of entirely new decision-making processes unlocked purely or mostly by emerging technologies.

In this article, after exploring the promises and pitfalls of AI in ADMs, we introduce the problem of opacity. We then analyze the challenges met by the concept of principled humanitarian action in an increasingly digitalized environment, and proceed to identify a set of critical issues representing the major points of friction between the current humanitarian ethics framework and the use of AI in ADMs. We conclude with a tentative roadmap towards principled humanitarian AI, including a research proposal to explore a set of proxy metrics and an explainability matrix.

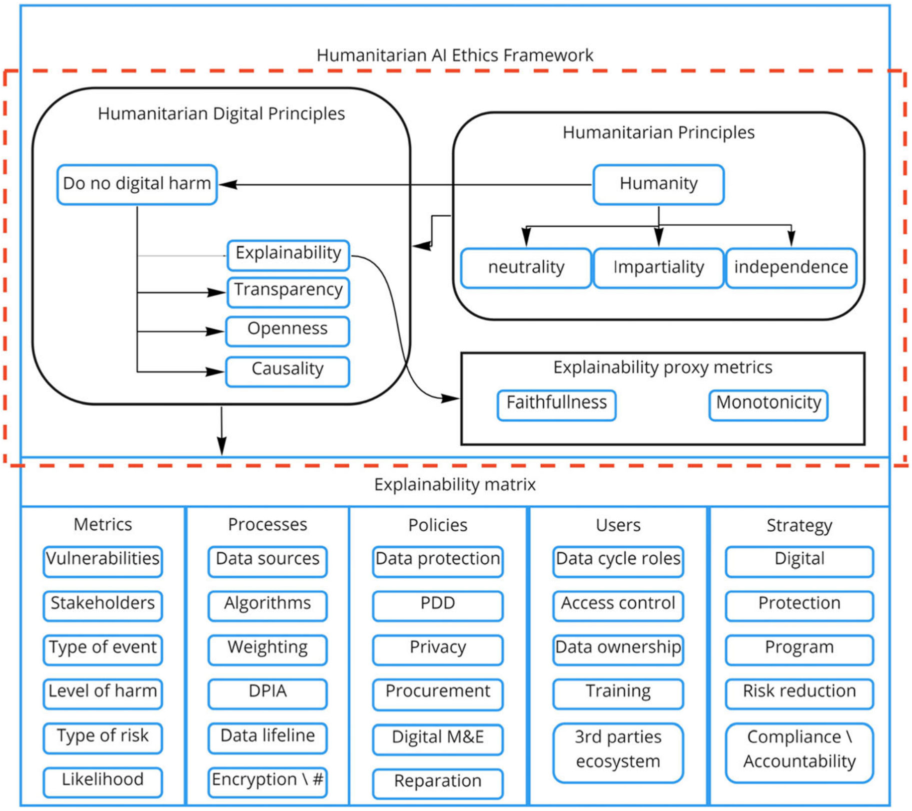
The scope of this article is thus to propose a theoretical framework that we believe could help humanitarians and tech actors in navigating the design and implementation of AI for ADMs (red dotted box in Fig. 1), with a special attention to the introduction of a tenet dedicated to explainability and to suggest an agenda for future research (represented by the whole diagram).

Machine learning, deep learning, artificial intelligence, and ADMs

AI has been defined in many ways, and there is not only one accepted general definition of it. In this paper, we will accept the definition used by Russell and Norvig (2010) that define AI as systems that ideally could (1) act like humans (e.g., interact with humans or imitate their acting); (2) think like humans (e.g., imitate the cognitive process of humans); (3) think rationally (e.g., using logic to solve problems, such as classification tasks); and (4) act rationally (e.g., automating intelligent behavior).

The field of AI has significantly evolved since Turing asked, "Can machines think?" (Turing 1950), but most of its accomplishments are commonly attributed to the exponential increase in computer processing power rather than advances in AI (Copeland 2019; Dreyfus 1992). This area of research has attracted attention also from within the humanitarian sector, especially in relation to AI-based ADMs. Within the broader spectrum of AI, ADMs refer to a particular class of technologies that either assist or replace the judgment of human decision-makers. Throughout the article, we refer sometimes to one or the other, but always in relation to one another unless otherwise specified. ADMs based on AI are systems that are expected to think and act rationally, as well as systems that act like human, replacing human judgments to respond to human problems (Russell and Norvig 2010). These systems draw from fields like statistics, linguistics, and computer science, and use techniques such as regression, rule-based systems, predictive analytics, machine learning (ML), deep learning (DL), and neural networks (NN), often in combination with one another (Molnar and Gill 2018).

AI technologies are born out of the radical decision of turning computer programming on its head. In the case of some ML techniques, instead of a programmer writing the rules to generate an algorithm to solve a



Metrics	Processes	Policies	Users	Strategy
Vulnerabilities	Data sources	Data protection	Data cycle roles	Digital
Stakeholders	Algorithms	PDD	Access control	Protection
Type of event	Weighting	Privacy	Data ownership	Program
Level of harm	DPIA	Procurement	Training	Risk reduction
Type of risk	Data lifeline	Digital M&E	3rd parties ecosystem	Compliance \ Accountability
Likelihood	Encryption \ #	Reparation		

Fig. 1 Scope of article (dotted red line) and proposed matrix for future research

problem, the program generates its own algorithm based on selected techniques and training data to generate a desired output (Knight 2017). This is how DL, which is a subset of ML, multi-layered neural networks—modeled to work like the human brain—"learn" from large amounts of unstructured data. While all machine learning can work with and learn from structured, labeled data, deep learning can also ingest and process unstructured, unlabelled data (IBM 2020). As an example, a DL system mathematically approximates the way human neurons and synapses learn by forming and strengthening connections. This is done by feeding training data to a neural network, which is gradually adjusted until it responds in the correct way (Knight 2019). When DL is used in computer vision in cancer screening, the machine is provided with a full raw dataset of images of an organ. It is then requested to identify an object within the image (e.g., cell anomalies) without being shown any previous example of how this looks like. The machine will later find similar anomalies in any new organ image, when present.

Just like cancer screening in DL, most areas within AI are in their early development and still require significant improvement. For example, DL-based algorithms can generalize and correlate similar inputs to outputs, but they perform much worse when applied beyond their training distribution (Bengio et al. 2019). They also hardly capture the effective potential for correlation between phenomena, and often struggle to attribute causation (Knight 2019). This leads, for example, to misclassification of objects within an image (e.g., misdiagnosis or mistargeting) or misidentification of individuals that could lead to more severe (legal or humane) consequences.

While some types of AI are already relatively transparent, others can be rendered transparent by explanations at a minimum of three levels: at the level of the entire model (Pizzi et al. 2020), at the level of individual components (e.g., parameters), and at the level of a particular training algorithm (Lepri et al. 2017). In the case of random forest algorithms, for example, the output results from combinations of other trees' outputs, and transparency is achieved by understanding what parameters were used to decide a certain output (branches variables) and the path that led to a final prediction outcome.

Many others (e.g., convolutional neural networks, hereafter CNNs) pose important challenges in understanding the causal linkages leading to their outputs (Holm as quoted in Gent 2019). CNNs process inputs (e.g., images) by assigning a weight and a level of importance to each incoming input, based on those processed previously. This means that the evaluation criteria are constantly changing, often in seemingly arbitrary ways. CNNs present two equally relevant problems related to the overall issue of transparency. These systems raise questions related to explainability, which comprises the focus on why a certain output is generated, and the concept of interpretability, which seeks to understand —without necessarily looking in the AI black box—how much can we trust the result to be equally reliable if another, different case is presented in the future (Choudhury 2019).

Both issues are key challenges in the evolution from evidence-based analysis to automation through artificial intelligence, a broad trend in the field of computer science. More broadly digital systems are transitioning from a situation of complication (a system that, despite the elevated number of its components, can still be given a complete description in terms of its individual constituents) to a state of increasing complexity (Page et al. 2018).

According to the definition developed by Cilliers, in a complex system, "the interaction among constituents of the system, and the interaction between the system and its environment, are of such a nature that the system as a whole cannot be fully understood simply by analysing its components. Moreover, these relationships are not fixed, but shift and change, often as a result of self-organisation" (Cilliers 2002). This definition fits perfectly the reality of humanitarian settings, where relationships of event, actors, and rules are always changing according to the specific operational context. In a rapid onset crisis, the emerging situation (e.g., internal displacement) and appropriate response (e.g., cash assistance and protection) are much more than just a function that combines their factors. In these cases, intersectional levels of vulnerability and power can overlap and evolve based on cultural, sociological, political, ethnologic, socioeconomical, and even historical factors linked for example to colonialism **and social justice**, exacerbating the complex dynamics of the system.

Arguably, in the overall balance between confirmed information available and unknown or unconfirmed data, the latter very often prevails, especially in hard to reach areas. The humanitarian complexity makes it extremely difficult to automate or simulate even part of the operational response cycle. Information systems would have to be capable to immediately adapt to often unclear new requirements and challenges, to be able to perform adequately (Ramaraj 2010). Any option given to the decision maker by a black box requires trusting a very broad probabilistic classifier or a network of functions, with very limited capacity to understand how the changing factors will influence the option given to them. But existing research on the topic leaves little to no room for trust in a tech culture that has often been accused of being opaque by design and not by necessity (Pasquale 2016).

Too much attention to the challenges of complexity would, however, be misplaced. Despite being used to assist in operations deployed in complex environments, most current humanitarian applications of algorithmic automation, including the use of ADMs to assist humanitarian decisions, do not actually fit the strict definition of technological complexity given by Cilliers. Differently from the research and commercial fields where most AI system are leveraging a potentially immense number of interacting components, all current humanitarian iterations have been deliberately kept from reaching such a level of sophistication.[2] There is a form of cognitive dissonance in the humanitarian sector pursuit of technology solutions that are designed to be complex—and by natural evolution, to be increasingly complex over time—and its constant downsizing of most concrete applications due to concerns about potentially losing control over it.

But even this cautious attitude will not shield humanitarians from the challenges posed by complex technological systems in the future. As commercial AI platforms become more efficient and ubiquitous, aid actors will eventually partake in a similar level of complexity by accessing (more or less knowingly) solutions powered by major tech providers, thus becoming themselves an additional cog in their vast list of components.

The problem of opacity

When platforms are so complex that their inner workings become unintelligible, researchers define the result as system opacity, which is the overall obfuscation of key processes leading to a certain output.

Burrell identifies "three distinct forms of opacity include: (1) opacity as intentional corporate or institutional self-protection and concealment and, along with it, the possibility for knowing deception; (2) opacity stemming from the current state of affairs where writing (and reading) code is a specialist skill and; (3) an opacity that stems from the mismatch between mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation" (Burrell 2016).

Concerns about opaque systems have grown in parallel with the interest generated by ADMs. Their promise to apply decision support systems for well-structured environments (Keller et al. 2004) found already concrete

[2] Humanitarian applications of AI have rather shown the marks of non-linear interaction by a relatively small number of equations, a state defined by Cilliers as "chaotic behaviour" or "deterministic chaos" (Cilliers 2002).

applications in supporting traditional functions such as logistics or payroll management systems. ADMs can either make use of embedded AI on the processing of the data or **reflect** completely rule-based systems. When rules are simple and well-structured, responding to both explainability and interpretability, ADMs have shown capacity to improve efficiencies and assist humans in decision-making.

But only rarely human or social environments deal with simple and well-structured social or professional environments. In less than 2 years, the number of business executives expressing concern about how to demonstrate that AI-powered processes fall within regulatory requirements has grown from around 29 to 60% out of a sample of 5000 informants (Brenna et al. 2018). As a result, several companies have started discussing publicly their policies, strategies, and even challenges in dealing with the complexity of AI (Castellanos and Nash 2018).

The tech sector has rapidly received the message from their commercial audience: establishing a generic causal link between inputs and output is no longer enough, even more important is the ability to examine the process end to end. Researchers proceeded then to develop new systems, including for example, action-to-outcome maps (ATOMs), visual representations of the whole project action explaining how the system expects to cause certain impacts (Perdicoulis 2016). The design of causality diagrams aims to provide a panoramic view of the project and even allow forecasting or future impact assessments through simulation ATOMs showing the results of qualitative simulation. (Perdicoulis 2016).

Other proponents have proven that some neural networks can be distilled into a soft decision tree, thus offering a visual representation of the pathways that led from inputs to outputs (Frosst and Hinton 2017). More recently, researchers proposed using the speed of adaptation to a modified distribution as a meta-learning objective, to determine the cause-effect relationship between two observed variables. This would create a training signal to find a way to factorize knowledge into components and mechanisms that match the assumption of small change (Bengio et al. 2019).

The list goes on, as progress has been made to develop algorithms for machine-learning models that can be understood by humans not only at protocol level, but also by identifying specific explanation methods, as we will see in further depth in the final section of this article. Molnar (2020a, 2020b) refers to both explanation methods (expressive power, translucency, portability, and algorithmic complexity) as well as the individual explanations (accuracy, fidelity, consistency, stability, certainty, degree of importance, novelty, and representativeness) for models and its predictions to be understood by humans. She suggests the use of model-agnostic methods (e.g., local interpretable model-agnostic explanations, LIME) which weigh for example the proximity of the sampled instances (data points) to the instance of interest (outcome or data point). Similar approaches can be taken to humanitarian applications in the design process, particularly for more opacity-prone applications.

It appears thus that the concept of explainable AI (XAI) has gained a solid foothold in the discussion over AI and ethics. Projects from Microsoft, Google, the World Economic Forum, and the draft AI ethics guidelines for the EU commission include different nuances of this same principle (Robbins 2019). Tech companies such as IBM (Bellamy et al. 2018) and tactical research agencies such as the US Defence Advanced Research Projects Agency, made of XAI an important part of their research programs. The common objective is to create a suite of machine-learning techniques that produce more explainable models that users can understand, manage, and trust, while maintaining a high level of learning performance (Turek 2016; Fig. 2).

As commercial ADMs are gaining traction and adoption, most early deployments of ADMs in complex applications with societal implications (e.g., assistive banking loans decisions, fraud detection, school admissions, or criminal sentencing) have been marked by unfair practices. Regardless of the technique they use—purely mathematical formula, machine learning, or deep learning—there is strong evidence that ADMs carry with them the risk to automate and reinforce inequality, discrimination, and bias (Eubanks 2018; Floridi, Cowls, et al. 2020).

In some cases, such as Predpol, a predictive surveillance system to inform tactical efficiency in patrol planning by law enforcement agencies, the competing public interest for safety has been argued to be sufficient cause to justify the risk of stigmatization and over-policing of some disadvantaged areas (Rizzi and Pera 2020). This, despite evidence that when compared to a simulation based on estimated drug use, Predpol resulted to contradict the distributed results of the scenario, rather pointing authorities towards predominantly Black neighborhoods at twice the rate as white neighborhoods (Lum and Isaac 2016).

In other situations, the AI system falls definitely short from this balancing exercise. It is the case of COMPAS, a software using algorithms to help judges in evaluating the risk of potential relapses. Reports based on the balancing test found COMPAS to be unexplainable as the algorithms are a commercial secret, partial in its judgment as influenced by human and societal biases during its training phase, and not having clear added value compared to simpler linear systems not requiring gender, ethnic, or racial data analysis (Rizzi and Pera 2020; Ingwin et al. 2016).

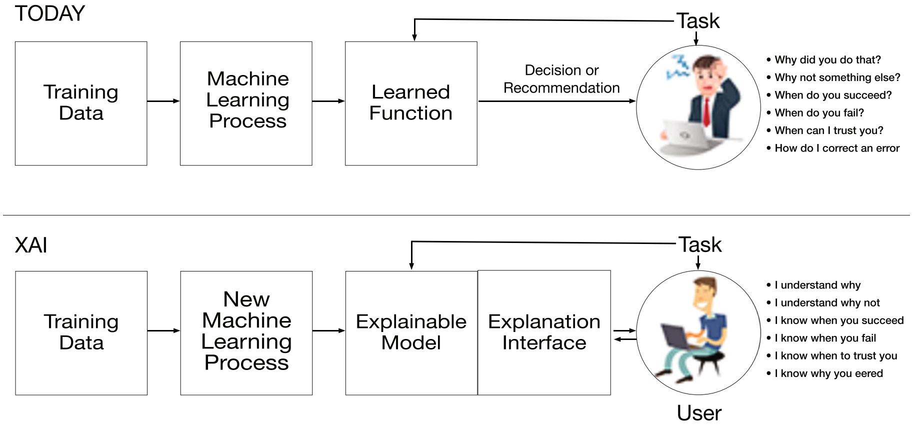


Fig. 2 XAI concept by Turek (2016)

A case study that could have had catastrophic humanitarian consequences was raised by the Citizen Lab and the International Human Rights Program at the University of Toronto's Faculty of Law, that successfully opposed the introduction of an ADM aimed at automating part of the Canadian immigration process. The case study, Bots at the Gate (Molnar and Gill 2018), reports AI experiments by Canada's government aimed at automating certain activities currently conducted by immigration officials and at supporting the evaluation of some immigrant and visitor applications. These can include decisions on a spectrum of complexity, including whether an application is complete, whether a marriage is "genuine," or whether someone should be designated as a "risk" (Keynon 2018). Even though the adoption of ADMs in the immigration systems is supposedly inspired by ethical principles, their application to profiles at "risk" could lead to automated learning-based errors and result in the *refoulement* of asylum seekers and refugees in a manner that would be illegal under international law.

However, not everyone sees the black box as a new problem and, in some cases, as a problem at all. As it has been noted already, even before AI—and most definitely in the pre-digital era—humans already based many decisions on judgment and experience resulting from their own natural deep-learning processes (Holm as quoted by Gent 2019). Opacity would then be something we already embrace and accept as part of our human nature. Around the concept, communities have developed governance structures to ensure consistency of processes and mitigation of their potential pitfalls.

It has even been objected that requiring explicability would hinder potentially ground-breaking applications, drawing parallels with some chemical or physical processes such as aerodynamic lift. Although still somehow scientifically unexplained, the mere act of flying is a positive and essential component of our collective experience that was harnessed through decades of trial and error (Regis 2020). Recalling Aristotle, supporters of the empirical approach in AI affirm that, "when our knowledge of causal systems is incomplete and precarious (…) the ability to explain how results are produced can be less important than the ability to produce such results and empirically verify their accuracy". Adopting a pragmatic and utilitarian focus, the empirical approach sees a blanket requirement that machine learning systems in—for example—medicine be explainable or interpretable as not only unneeded, but unfounded and potentially harmful (London 2019). Others, adopting a more relativistic approach, have noted how "there seem to be many implementations of AI in situations of low to no risk (in terms of harm)" and that it "is unreasonable that the decisions resulting from AI in these situations should be required to provide explanations" (Robbins 2019).

While the former argument is affected by the underlying logical fallacy of assuming undefined and unproven vital benefits for the common good as a reason to rein in doubts and hesitations, the latter argument builds on an extremely narrow concept of responsibility, more related to liability than to ethics. According to this theory for example, racial discrimination resulting from a biased dataset used to train an efficient medical algorithm would not be an issue pertaining to opaque AI. It would rather just be a sign of poor performance of the system once it is proven that race is not a key feature determining the output by design (Robbins 2019). A clear example might be an AI system providing early detection of skin cancer, that is trained only or mostly on datasets from people with light skin and European upbringing, thus failing to detect essential symptoms on darker skin and increasing their likelihood of being detected only at later stage when compared to other skin types. Based on the pragmatic approach, this issue would not be imputable to the opacity of the AI system, but only to a failure in designing the training dataset. The correct response in this view would be to not pull the system, but to correct the training for further use. This interpretation might appear reasonable when the primacy is placed on the overall wellbeing or prosperity of the public collective, but results inacceptable when the main objective is the dignity, safety, and security of any individual in situation of need or vulnerability.

More specifically, this argument builds on the unproven assumption that "low to no risk" applications could be identified and defined in an abstract way, just by their original design. The argument that "[t]he property of requiring explicability should attach to a particular action or decision rather than the entity making that decision" (Robbins 2019) fails to explain how tech development companies could possibly preventively identify all actions that are intrinsically harmful, and only design AI systems around those that are not. Humanitarians and human rights practitioners cannot in good conscience deploy a solution that has only some chances of being non-harmful. Most digital and non-digital solutions have been weaponized in various and often unexpected ways in the past.

Similar conclusions have been reached in sectors more or less adjacent to humanitarian technology. A joint UNICRI-INTERPOL report on AI and Robotics for Law Enforcement states that their use in law enforcement should be characterized by fairness, accountability, transparency, and capability of being explained (UNICRI-INTERPOL 2019). In the UK, a review by the Committee on Standards in Public Life found that AI "has the potential to revolutionise the delivery of public services". The same report, however, warns that AI also poses challenges to at least three of the Nolan Principles, which constitute the basis of the ethical standards expected of public office holders in the country: openness, accountability, and objectivity (Committee on Standards in Public Life 2020). This review follows a poignant report by the Special Rapporteur Philip Alston who first noted that "[t]he British welfare state is gradually disappearing behind a webpage and an algorithm, with significant implications for those living in poverty" (Alston 2019a, 2019b). The Rapporteur explained that the concept of transparency covers not just the mere existence but also the inner workings of automated systems. He also stated that in its absence "the rights to contest an adverse decision and to seek a meaningful remedy are illusory" (Alston 2019a, 2019b). On a similar note, the OECD released a review into bias in algorithmic decision-making recognizing that "[i]t is well established that there is a risk that algorithmic systems can lead to biased decisions," especially when existing human biases are encoded into algorithmic systems. According to the OECD report, system owners should "ensure that decisions can be scrutinised, explained and challenged so that our current laws and frameworks do not lose effectiveness, and indeed can be made more effective over time" (OECD 2020).

In practice, this argument had found regional normative strength already in 2018 through the introduction by the European Union of Article 22 under section 4 of the General Data Protection Regulation (GDPR). According to article 22, if a user sees their claims rejected based on scores from automated intelligent processing software, the interested party has a right to demand an explanation. Any incompliance can be sanctioned up to €20 Mn or 4% of the company's global annual turnover (Zomignani Barboza et al. 2020; EU General Data Protection Regulation 2016; Complete guide to GDPR compliance 2020).

The same balancing exercise of efficiency versus guiding ethical principles represents a challenge in the deployment of ADMs for assisting decision-making processes by humanitarian organizations and especially those under GDPR jurisdiction.

Humanitarian principles in a digital world

Humanitarian ethics are principle-based, building on four core principles (humanity, neutrality, impartiality, and independence) and an environment of around thirty-three overall principles that are routinely used in the pursuit of humanitarian action (Slim 2015).

Despite being originally action-guiding, the role of humanitarian principles goes beyond their operational value. Their importance in framing the space for humanitarian policy and action is widely recognized, including by official public policies (Norwegian Ministry of Foreign Affairs 2019). Among the four core principles, only two are generally considered to be absolute and constitute exceptionless norms: the principles of humanity and impartiality (Slim 2015).

The principle of humanity is supposed to drive any organization whose "purpose is to protect life and health and to ensure respect for the human being" (Pictet 1979).[3] This principle "enables the institution to define its tasks, to outline the field for its intervention and mark its limits (…). Although it is the purpose […] to make the world a better place, it can do so only in certain respects. It cannot undertake every activity regarded as benevolent but must concentrate on specific responsibilities. Only in so doing will it guard itself from a dangerous dispersal of effort" (Pictet 1979). As later formalized through the extension of the Hippocratic Oath of Do no harm to the ethics of aid, the restorative action of alleviate suffering must be accompanied by preventive action (Pictet 1979).

[3] Although it has been highlighted by Slim that this is a formulation of objective, not value. It "states what humanitarian action wants to do, but it does not explain why it is good to do it" (Slim 2015).

Humanitarians are also required to act impartially, assisting solely based on need. The principle of impartiality encompasses three subprinciples, namely non-discrimination, proportionality, and impartiality in its narrower meaning. Pictet recalls that "[f]rom 1864 onwards, non-discrimination found expression in the Geneva Conventions and, later on, in international or regional human rights and humanitarian legal frameworks. It is also a principle of long standing in the field of medical morality and ethics" (Pictet 1979).

While the subprinciple of non-discrimination is restrictive—mostly focusing on defining what should not be done—the concept of proportionality is positive as it requires aid workers to provide assistance consistently with the degree of the suffering and based on their degree of urgency (Pictet 1979). The last subprinciple, impartiality, instructs humanitarians to act based on existing rules and principles, and notably the substantive principles of humanity, non-discrimination, and proportionality.

The remaining principles of neutrality and independence are considered obligatory but not absolute. They represent strong obligations but can tolerate exceptional circumstances (Slim 2015). The principle of neutrality requires an abstention from judgment, as long as this does not worsen the situation of persons affected by the crisis. Neutrality is never applied to those who suffer but only to belligerents and only to make sure conditions are met "to continue to enjoy the confidence of all" (Pictet 1979). Humanitarians are also required to be independent and operate accordingly, which translates in their sovereignty over decision-making involving political engagement, religion, and economic issues. Despite being a derivative principle, the adherence to the concept of independence is also key to maintain neutrality (Pictet 1979).

Although it is common understanding that information technology (ICT) now being part of humanitarian action should be guided by the four humanitarian principles (Vonèche Cardia et al. 2017; Raymond and Card 2015), ICT design, adoption, and deployment in situation of crisis are not often approached with the principles in mind (Vonèche Cardia et al. 2017).

As recalled by Slim, in applied ethics, principles are used for three main purposes: (1) to affirm moral norms; (2) to act as constant operational guides to ethical decision making; and (3) to generate specific rules (Slim 2015). For the scope of this article, we focus mostly—albeit not exclusively—on the second aspect, exploring how these guides behave when abstracted into digital systems beyond human control, or when such control moves away from humanitarian actors.

Some actors providing humanitarian services or doing business in humanitarian contexts have objected to their subjection to such guidance, claiming that their mandate is not to align on philanthropic ideals of NGOs they cooperate with, but rather to make profit to fulfill their statutory role and commercial nature. In this perspective, putting humanity first would not be a strict requirement for their engagement, even if they engage in socially worthwhile initiatives (Friedman 1970). Fifty years later, these theories might be less loud but are still very present. Just recently, the CEO of Silicon Valley-based cryptocurrency exchange and broker Coinbase affirmed in an open letter that staff should avoid distractions, focus on their respective jobs, and work toward making their employer a great company. Achieving the company's mission is presented as "the way that we can have the biggest impact on the world". In his words, the company will have an impact by focusing on building and being transparent about what our mission is and is not with engagement in politics and championing of social issues both falling in the latter category (Kelly 2020).

Among humanitarian researchers, however, there is limited controversy on the matter. Although in the past the principle of humanity was seen as limiting to "a consecrated priesthood of relief agencies and their relatively small range of relief activities in war," the same principle has evolved into a cosmopolitan or universal ethic. Humanitarian responsibility extends to all parties involved in war and with war including those with indirect stakes such as international businesses and especially technology companies providing services related to humanitarian action (Slim 1998). The intertwining between ethical factors driving technological advancement and humanitarian principles is evident in the work of Dodgson et al., introducing how eight key AI principles emerging from current literature translate in the humanitarian do no harm framework (Dodgson et al. 2020). Despite looking a seemingly abstract exercise, this debate has very concrete consequences: private actors and third parties engaging in the so-called war economy or providing services to humanitarian organizations must respect most of these guides, if they want to avoid being considered a legitimate military target by the warring parties (ICRC 2006).

Even on the purely humanitarian side of the spectrum, this is far from being an intellectual speculation on the collective and shared ethical responsibilities in situations of natural disaster or violence. The adoption of ICT systems, including the first implementations of AI, has been marred by prevalent biases, security risks, and issues with consent that can undermine the role of humanitarian actors in crisis contexts by leaving aid recipients at further risk of vulnerability. It has also been affirmed that the negative impact of AI and ADMs could indirectly affect the maintenance of international humanitarian and human rights legal frameworks (Wright and

Verity 2020), by undermining existing protection and accountability mechanisms.

Critical issues in the adoption of AI for principled humanitarian action

AI model training and humanitarian experimentation

Differently from human reasoning, any technique currently used to build ADMs cannot analyze, predict, or transfer knowledge to anticipate potentially harmful consequences, if it has not already recorded and studied the same or similar combination of cause and effect several times in the past. As it has been noted, to understand that dropping objects causes them to break, a robot needs to toss dozens of vases onto the floor and see what happens (Knight 2019). When looked at from the lens of the humanitarian principles, this approach falls within the notion of humanitarian experimentation, a practice that is incompatible with the "do no harm" imperative (Sandvik et al. 2017). An example could be the use of biometrics and other demographic identifiable information in a predictive model for fraud prevention, where an untested technology could be deployed and refined on unaware and disempowered individuals in situation of vulnerability outside of protective legal frameworks or accountability mechanisms. The risk of exploiting human suffering to improve digital systems is a first major obstacle to the ethical implementation of ADMs in humanitarian settings, especially as it exposes these communities to a high risk of system failure. Such risk often comes with no real option to opt out, contest, appeal, reparate, redress, nor a promise to obtain a concrete direct benefit that would not be achievable with a more established solution.

Some mitigating measures could prove effective, such as using exclusively historical data, anonymized and cleaned to ensure people's protection and dignity, particularly those who are most vulnerable. However, to be effective over time, AI algorithms require regular refreshing of the training model to match changing conditions (Chui et al. 2018), a requirement that seems inevitable in any humanitarian context. The need for updates of large-scale datasets on a yearly, monthly, or in the example of the fraud prevention mechanism mentioned above, even daily basis would rapidly require humanitarian organizations to feed almost real-time data to the model, an operation that can only be satisfied by stretching an already-overwhelmed technical capacity for data collection or even overriding risk-reduction policies.

The use of humanitarian-related data to improve training models poses a further ethical problem when adopting third parties' systems, even if implementation happens within the humanitarian mandate. Most common commercial AI algorithms generate an enormous return on investment for companies, contributing to an estimate of $3.5 trillion and $5.8 trillion in value annually across nine business functions in 19 industries (Chui et al. 2018). Feeding data and metadata generated from processing activities involving people experiencing humanitarian distress — often with poor acquisition and processing quality — in order to train the model used to refine such a profitable business model constitutes part of a broader dilemma that extends to the fields of messaging, cloud-based systems, big data models, or even cash-transfer programs and social media (ICRC and Privacy International 2018), particularly when this data is the result of aid donations or public funding. This raises significant dilemmas especially as the direct added value of the digital system for the individuals in situation of vulnerability is often not evident prima facie, as shown by the criticism that followed the announcement of a partnership between the World Food Program and Palantir, a data software company known for its work in intelligence and immigration enforcement. The partnership, worth $45 million, raised concerns as it involves a data integration that would include records of distributions to program participants by the aid actor with the company that has been criticized for "secrecy, profiling bias, enabling human rights violations, and the wholesale harvesting of personal data" (Parker 2019; Mijente 2019).

A clash of opacities: translating humanitarian protocols into ADMs

The disruption of the causal link between human observation, analysis, and decision-making was already affecting the aid sector in the pre-digitization era. The humanitarian sector has been defined as historically "bad at connecting information that it gathers to decisions that it makes" (Humanitarian Congress Berlin 2018). In this sense, the increased attention given to automated decision-making systems compared to the similar issue of opacity in human-controlled decision-making systems is again another peculiar form of cognitive dissonance.

This skewed perception is not however completely without basis. As we have seen already, there is now broad public awareness among managers about the sudden potential to fall out from compliance with ethics at scale without them noticing, being able to explain why this is happening, or even do anything to prevent it. The private sector already offered a series of cautionary tales, starting from the inquiry opened by New York State regulators on the algorithms used by Apple Card to determine the creditworthiness of applicants, after many prominent figures publicly complained about gender discrimination (Vigdor 2019). The friction between concern and aspiration is worsened by the pressure that the international community puts on the humanitarian system to deliver quicker results, and even to recur to anticipatory humanitarian action[4] to improve the efficient use of resources. While the problem of opacity is not new to the sector, digitizing it into an AI-powered system could add a further layer of complexity to it. The use of AI could institutionalize opacity and make it structural by embedding it in digital transformation processes. As recalled by Rizzi and Pera, we "do not count, at least for now, with a way of trespassing axiological values to exact value units which can be introduced inside an algorithm, nor a method to conjugate in it any reference of principles" (Rizzi and Pera 2020).

To tackle the concept of causality in the AI dimension, the development team must then first standardize the wealth of processes that drive decision-making or at least design a neural system that could reach a similar result. In humanitarian contexts, this implies translating the ethical frameworks underpinning the delivery of assistance and protection to persons affected by situations of crisis, and notably the principles defined in the previous section, in software modules capable of constructing or assisting in decision-making processes. While humanitarian experts drafting these principles appreciate a large degree of vagueness and freedom of interpretation as strengths in dealing with ever changing and unpredictable situations (Gisel 2016; Labbé and Daudin 2015), the opposite is true for algorithmic systems, where rule-based models are currently essential in ensuring algorithmic interpretability (ICRC 2019).

While this calls for caution in deploying ADMs, it may also open opportunities to embrace an open-ended attitude towards unexpected and surprising outcomes. In a way, and with the caveats that algorithms themselves carry with them their own set of biases infused from their human designers and operators, algorithmic assistive systems could be harnessed to mitigate or compensate forms of human-specific bias in decision-making. This is the case—for example—of confirmation bias, a high-risk factor affecting the humanitarian sphere "given the strong role of humanitarian narratives, and the reliance on closed social networks, motivational and cognitive elements" (Comes 2016). An early example is the effort done by UNHCR to try to remove or mitigate any type of bias in their recruitment process through project ARiN (Brookland 2019).

Opacity as disconnect from humanitarian principles in ethical decision-making

In the public discourse, AI systems are accompanied by an aura of enormous potential, overlooking the countless ways in which these systems can fail. Shankar et al. have counted over 200 journal entries published over just 2 years describing adversarial attacks on the algorithms and data, a number increasing even more when including also non-adversarial failure modes. Their work resulted in a taxonomy of machine learning pathologies, categorizing failures and their consequences so that policy makers can begin to draw distinctions between causes which will in turn inform public policy initiatives to promote ML safety and security (Shankar et al. 2020).

As mentioned, the accountability gap resulting from lack of evidence-based decision-making is something that is well-known in the humanitarian sector and whose ramifications have been object of thorough research and experimentation. Even considering this, the three forms of algorithmic opacity defined by Burrell present unprecedented risks for humanitarian ethics, resulting in forms of abdication of the centrality of humanitarian principles in decision-making processes, combined with the potential harm multiplier effect of AI systems (Brundage et al. 2018).

When relying on proprietary code or whenever being precluded from auditing backend processes managed by partners or third-party providers, humanitarians make themselves vulnerable to errors or manipulation. Errors could go undetected if the organization has no means to tell if the algorithm is valid or if it is actually better than other existing models (Handelman et al. 2019).

Errors could also be derived from the inability to understand why (or which) inputs generate a certain output, resulting in unchallenged assumptions becoming operational decisions in life-threatening situations. For example, an ADM generating needs assessment and response planning for assistance distribution in an area of displacement where multiple communities are affected, the system might orient field teams in prioritizing the wrong group based on incorrect data training, modeling, processing, or analysis. In addition to constitute a breach to the principle of impartiality, the inability of local teams to understand the error and mitigate its consequences could increase tensions among affected groups and potentially fuel additional conflict. Such a situation could be due to a wide array of factors, from the so-called shadow AI introducing automated decision systems outside the oversight of the institutional IT department (Cearly et al. 2019),[5] to the incorrect integration of

[4] See, e.g., the Core Responsibility number 4: Change people's lives: from delivering aid to ending need, endorsed as part of the Agenda for Humanity at the World Humanitarian Summit by 180 Member States of the United Nations over 700 local and international NGOs, the private sector. The Agenda for Humanity is a five-point plan that outlines the changes that are needed to alleviate suffering, reduce risk, and lessen vulnerability on a global scale (https://www.agendaforhumanity.org/cr/4. Accessed on 10\02\2020).

[5] Research by Gartner suggest that by 2022 around 30% of organizations deploying AI for decision-making will have to face the phenomenon of shadow AI as a major risk to effective and ethical decision-making (Cearly et al. 2019).

those systems with the local decision-making environment.

But humanitarians could also be instrumental to abuses by external actors profiting from the data and metadata generated in the process or intervening in the mathematical manipulation that happens in between weighted inputs and classification outcomes (Burrell 2016). Kaspersen and Lindsey-Curtet provided an example of how neutrality—or rather the perception of it by affected communities—could be compromised by a phone hack leading to a military attack against a location visited by an unsuspecting humanitarian team doing protection work (Kaspersen and Lindsey-Curtet 2016). While this specific scenario does not mention the use of AI, the same risk applies to the use of deep learning technologies even without the need for an unlawful electronic intrusion in the humanitarian digital kit. When generating data and metadata in a cloud-based, proprietary, and third party-provided system, the information is processed, mixed, and shared in potentially countless training datasets and databases for all sorts of purposes. It is highly probable—considering that military and intelligence actors are expected to be among the major investors and users of autonomous and advanced technologies (MarketResearch.biz 2020)—that some of that data will contribute to invisible processes leading to targeting in law enforcement or military operations. This is also true for potential surveillance of vulnerable populations in certain already-difficult contexts (Singh 2019).

The risk of mathematical manipulation is more subtle, but just as dangerous. This could result—for example—in the deliberate downscaling of the protection risk for a specific ethnic group or, on the contrary, inflating the risk factor for a less vulnerable community enjoying favorable political connections or ongoing humanitarian assistance (e.g., assistance targeting based on mathematical/statistical formulas). In some cases, the distortion in the parameters or systems could be due to bad faith or manipulation by the same humanitarian actor, be it intentionally (modeling inputs or tweaking the algorithm to confirm a preconceived notion or decision, or to cover up a mistake) or unintentionally (e.g., due to poor data quality or through confirmation bias, as the dataset used by the algorithm could be skewed towards those situations or communities more frequently visited or monitored in the past or those whose voice is stronger in the community leading to misrepresentation). Some of the examples mentioned reflect what we could tentatively define as "functional opacity," a condition where the lack of visibility and control over the inner wirings of an AI system applies only to those parties involved in the operational use of the solution towards the implementation end of the data pipeline.

Functional opacity could also result from the limited access of humanitarian organizations to the professional profiles required to master artificial intelligence. This scenario would expand the risk profile also to organizations using open code or non-proprietary solutions and is likely to affect in a particular way local charities with limited funding and working in volatile environments. On the epistemic level, the introduction of a super-humanitarian holding the technical skills required to understand, run, and oversee these algorithms would increase the challenges in realizing the localization agenda and make access barriers for direct action by the broader spectrum of small local organizations even harder. Considering that AI systems have been proven to benefit from an almost irrational level of trust from non-technical users to the point of generating behavioral influences in their choices or perceptions (Warshaw et al. 2015; Springer et al. 2017), the concentration of AI skills in the hands of few Western organizations would revive power dynamics based on blind trust, dependency, and authority typical of what has been defined as technocolonialism (Madianou 2019).

Finally, in relation to the last shade of algorithmic opacity identified by Burrell, there is an irreconcilable disconnect between human and machine reasoning, as these two realities respond to mechanisms and logic that are very distant from each other. In neural networks, where "an algorithm does the 'programming' (i.e. optimally calculates its weights) […] it logically follows that being intelligible to humans (part of the art of writing code) is no longer a concern, at least, not to the non-human 'programmer'" (Burrell 2016). Most AI systems are in fact designed to evolve so that the implementation process is increasingly abstracted away, their validity being only judged by the quality of its inputs[6] and—especially—the correctness of its outputs (Venkatasubramanian 2019).

But all of the non-absolute humanitarian principles are interpretive concepts, which means that their implementation needs specification in a particular situation. Lacking this, they can result in moral conflicts due to competing principles, or even moral paradoxes, leading to harm as a result of a formally correct application of a principle (Slim 2015). Unfortunately, in the immediate future, humanitarians can rely on limited help from their technical partners. As noted by Venkatasubramanian, "[e]ven the unit tests we build for software test inputs and outputs, rather than process" (Venkatasubramanian 2019).

[6] Which is in and by itself a serious issue as datasets and data training strategies are mostly tailored on what are commonly defined as "Caucasian" men profiles and experiences, as highlighted for example by Balsari (2019).

Noise in the AI ethics panorama

The review of existing literature highlighted an overarching framework consisting of five core principles for ethical AI, four of which are core principles commonly used in bioethics: beneficence, non-maleficence, autonomy, and justice. In addition to these, Floridi and Cowls propose an expanded version of the pre-existing concept of explicability as intelligibility. The objective is to move beyond the already seen questions "how does it work?" and "how much can we trust its consistency in implementation?" This broader version of the principle of explicability incorporates "both the epistemological sense of intelligibility and the ethical sense of accountability (as an answer to the question: 'who is responsible for the way it works?')" (Floridi and Cowls 2019).[7] The principle of explicability states that "for AI to promote and not constrain human autonomy, our 'decision about who should decide' must be informed by knowledge of how AI would act instead of us" (Floridi and Cowls 2019).

Reaching broad agreement on this interpretation of the principle of explicability would definitely be a step in the right direction. A step that, however, risks to have limited impact if it remains just another entry in the endless stream of guiding documents dedicated to ethics in AI.[8] The ethics landscape of AI seems to suffer from the same deterministic chaos of obscure algorithms. As it has been noted, the problem with this technology is not so much the lack of principles but an uncontrolled proliferation that undermines their authority (Floridi and Cowls 2019; Wright and Verity 2020).

The continuous growth of proposed soft tools in the AI ethics environment is hampering the establishment of a bedrock of rules and principles where both researchers and practitioners find a shared agreement. This in turn reduces the capacity of humanitarian actors to engage with peace of mind, as they lack the capacity to trust that by adopting a certain solution they are also buying into a common set of values. But ethics are not the only framework of reference, as the humanitarian sector is constantly called to make complicated trade-offs between the flexibility of unenforceable and fleeting ethics guidelines, policies, and codes of conduct, and the slow-moving rigidity of rights-based frameworks (Gruskin and Dickens 2006). In the deafening noise of light policy documents and frameworks, a clear signal has been instead given by the normative sphere.

The General Data Protection Regulation introduced in 2016 stated unambiguously the need for transparent algorithmic decision-making. In Art. 22 it envisions a "Right to Explanation" (EU General Data Protection Regulation 2016 Art. 22; Goodman and Flaxman 2019) which represents a welcome development in providing enforceable guidance. The recent decision of the District Court of the Hague in the Netherlands in the System Risk Indication (SyRI) case (NJCM cs/ De Staat der Nederlanden) showed that the most effective response might actually lie in the interplay between GDPR-like normative documents, Human Rights treaties, and national law. SyRI was a program collecting 17 categories of government data from residents living in low-income and immigrant neighbourhoods assigning each household a value through a predictive algorithm to indicate the level of risk to benefit agencies. The court, building also on an Amicus Curiae brief by the UN Special Rapporteur on Extreme Poverty and Human Rights (Alston 2019a, 2019b—Brief), found the program in violation of the European Convention of Human Rights (as it assumed that people in some neighbourhoods had higher chances of committing crimes) and data protection (as GDPR prohibits a mass collection of personal data without explanation or consent) (Alston 2019a, 2019b—Brief; Burack 2020).

Humanitarian governance and algorithmic decision-making

Humanitarian organizations officially adopt conservative approaches to the use of unfamiliar digital systems,[9] an attitude due in equal parts to protection concerns and limited resources. The same cautious approach do not always find consistent application when organisations are faced with the suasion of potential implementation of technological solutions in seemingly intractable onset crises (Sandvik, Jacobsen and McDonald 2017).

The analysis of the policies made publicly available by humanitarian institutions shows the abundance of digital device guidelines, data collection methods, soft policy contributions, GDPR compliance statements, and internal reactive press tool protocols. But it also shows the absence of official enforcement, governance or redress policies and standards for harm done to individuals for breaches to their privacy, data protection, or physical integrity as a result of technological failures.[10] According to the risk framework developed by Metcalfe et al., it appears that organizations often consider digital risks as institutional rather than programmatic (Metcalfe, Martin, and Pantuliano 2011). While programmatic risk

[7] Recently, yet another principle has been proposed, inspired by the concept of solidarity in redistributing wealth, resources, or even increased productivity originating from the introduction of advanced AI systems (Luengo-Oroz, 2019).

[8] An attempt to reflect the amplitude of existing guidelines and frameworks can be found in the "AI Ethics Guidelines Global Inventory". Available at: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/ ;

[9] See for example the approach of ICRC to AI (ICRC 2019).

[10] "AI Ethics Guidelines Global Inventory". Available at: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/; the authors also inspected 21 websites of international and local organizations looking for mentions of policies on redress for harm from the use of digital or technology. Full list available in Appendix .

includes the “[r]isk of causing harm through intervention” (Metcalfe, Martin, and Pantuliano 2011), institutional risks are defined as “[r]isks to the aid provider (security, fiduciary failure, reputational loss, domestic political damage)” (Metcalfe, Martin, and Pantuliano 2011).

Common approaches to digital risk mitigation appear thus aimed at setting off reputational risk, resulting in brand protection strategies to shield the organization from accusations of partisanship or partiality from parties to a conflict. In this way, organizations adopt a liability lens to translate the principles of neutrality, impartiality, and independence to their digital dimension.

It is hardly possible to overstate the importance that reputation plays in allowing safe and effective access to the most hard-hit areas of the world. It is not by chance that the emblems of the Red Cross and Red Crescent movement (including the ICRC, the organization entrusted by the Geneva Conventions with the task of monitoring compliance of warring parties with IHL) enjoy special attention under international law as protected symbols when used in their operational function (Rolle and Lafontaine 2009; ICRC 2020).

However, with the increasing pervasiveness of advanced digital solutions in the first line of humanitarian action, the balance between brand protection and individual agency requires enhanced scrutiny due to its potential to do harm both individually and at scale (Greenwood et al. 2017; Wright and Verity 2020; Dodgson et al. 2020).

Implementing the principles of neutrality, impartiality, and independence with a liability focus is likely to create a disconnect with the principle of humanity, the essential principle “from which all the other principles flow” (Pictet 1979). As it happens, for any action to be defined as humanitarian, humanity “obviously has to stand in first place” (Pictet 1979; Greenwood et al. 2017). Even assuming that a liability approach would not aprioristically negate the primacy of the principle of humanity, the issue of whom the humanitarian system is liable to becomes then the key factor in defining this question.

A roadmap to humanitarian AI

Adopting explicability and its proxies as a humanitarian digital tenet

We saw how Floridi and Cowls (2019) proposed expanding the five core principles for ethical AI to add a broader version of the principle of explicability, that includes both intelligibility and accountability. This proposal is consistent with the example set by the European regulators through Art. 22 GDPR that as mentioned introduced the right to demand an explanation and built the basis for the first legal actions against the unfair deployment of ADMs.

To align with this trend, we strongly advise that humanitarian actors move from adhering to sector-wide platforms such as the Digital Principles (Principles for Digital Development 2015) to adopt more granular policies on technology development and/or human rights-based frameworks applied to AI. So far, endorsement of broad principles such as transparency, openness, and causality could have sufficed to mirror their commitment to traditional humanitarian principles. This, however, is no longer enough when dealing with complex systems such as AI platforms. The first step towards a principled humanitarian AI should be the adoption of an ethics charter including explicability as a core tenet to the principle of do no digital harm.

We also propose to adopt a method based on faithfulness and monotonicity (Das and Rad 2020) to improve human understandability of explainability method results (see Fig. 1). A faithful interpretation is one that accurately represents the reasoning process behind the model’s prediction. In line with the proposal by Jacovi and Goldberg, this judgment should not happen in a binary manner (*faithful–not faithful*) but rather allowing the evaluation of a system on a spectrum (Jacovi and Goldberg 2020). A monotonic model is a model that has some set of features (monotonic features) whose variation always leads the model to consistently adjust its output (Das and Rad 2020). In humanitarian terms, we can imagine that an AI system is poised to indicate the short path communities shall follow to reach an area of distribution. If suddenly an information about the potential presence of landmines on the same path is added to the system, a non-monotonic model would start weighting the different factors before taking a decision based on a rationale that would be hardly predictable in advance. On the other hand, when using a monotonic model, even the barest minimum signal flagged as unacceptable would always suffice to activate a safety protocol even in the presence of a large amount of other signal (Tsukerman n.d.).

Defining a set of metrics for forward engineering a humanitarian AI

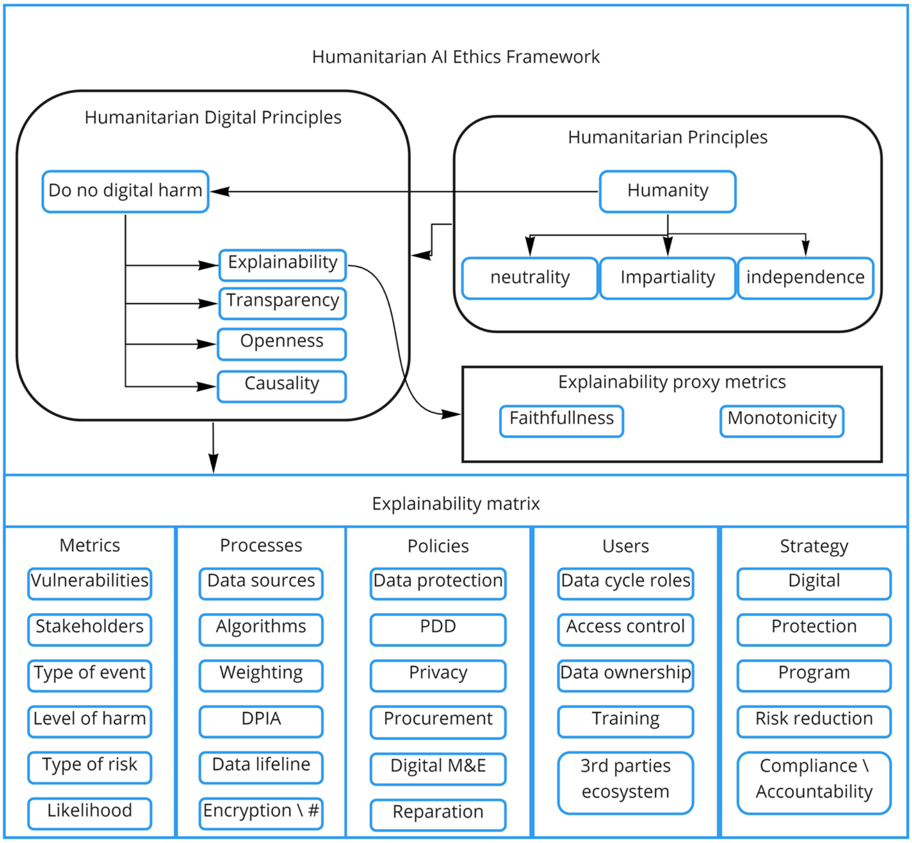
Once humanitarians have embraced a set of digital principles specific to AI that include the principle of explainability and its proxies, the problem of how concretely to design an ADM prevented from or uncapable of doing harm still remains largely untouched. The model will need to be designed and set on the right parameters in a process that cannot be purely retrospective. The traditional approach of trial and error to adjust the factors and improve performance by retrofitting the system simply would not be ethical in a social or humanitarian context. It would also correct issues in processing data without however removing eventual structural biases or errors at the source code or data level, in what Polack defined “algorithmic reformism” (Polack 2020).

Building on the theory of “Forward engineering” developed by Polack, and irrespective of the algorithmic solutions that are to be implemented, at algorithm design stage, we can identify how “relationships between design constraints lead to design implications: technical limitations, dependencies, and design compromises that are not made explicit by algorithmic frames but emerge in the process of forward engineering them” (Polack 2020; see also Theodorou et al 2017). We propose as an item for further research the breakdown of those implications based on the initial matrix presented in Fig. 3 (within the dotted red box). The objective is to start identifying a data model on top of which the forward engineering approach and the monotonicity triggering factors can be built. The same factors could also be used to run what Watson and Floridi called the “explanation game,” a formal framework for conceptualising the goals and constraints of explainable AI systems (Watson and Floridi 2020).

Enforcing the principle of precaution while building explicable AI systems

While waiting for reliable and effective explicable AI systems, some temporary preventive measures can already be taken. Two key purposes of humanitarian principles are to affirm moral norms and generate specific rules (Slim 2015). We could then envisage that “if the expression of norms is in the form of a specific process or implementation, then we need to institute ways to freeze that implementation—or at least continually audit it—in ways that we don’t typically do with software” (Venkatasubramanian 2019). This could take the form of an inquisitorial model of quality control to achieve technological due process (Keats Citron 2007), proposed together with the standard that research studies of ML algorithms should include in the end product, the predictive algorithm developed (Handelman et al. 2019).

While promising, early experiences have shown the current limits of algorithmic accountability. Despite sharing the same name with well-established practices in the tax and financial sectors, algorithmic accountability seems to suffer from the lack of incentives to function as



Metrics	Processes	Policies	Users	Strategy
Vulnerabilities	Data sources	Data protection	Data cycle roles	Digital
Stakeholders	Algorithms	PDD	Access control	Protection
Type of event	Weighting	Privacy	Data ownership	Program
Level of harm	DPIA	Procurement	Training	Risk reduction
Type of risk	Data lifeline	Digital M&E	3rd parties ecosystem	Compliance \ Accountability
Likelihood	Encryption \ #	Reparation		

Fig. 3 Diagram of the proposal for a humanitarian AI ethics framework and explainability matrix

a check on AI applications. This problem was made evident in the debate that engulfed HireVue, whose claims of audited fairness of its AI system to analyze facial features and movements during job interviews have been revealed to mischaracterize the results and scope of the audit (Engler 2021).

Alternatively, different AI models could be explored that do not focus on task automation. These models would rather aim to provide a person with augmented control over decision-making, seen as a creative process. Applications of this theory, repurposing the internal representations of neural networks learned as tools, have so far been tested in images and music and named activation atlases. These algorithms form a collection of simple, atomic concepts that are combined and recombined to form much more complex visual ideas. Using something like an activation atlas as a palette, they allow the user to "dip a brush into a "tree" activation and paint with it," using a palette of concepts rather than colors to create an array of machine learned, but human interpretable, languages for images, audio, and text (Carter et al. 2019).

Similar findings were also reflected in a set of four guiding principles that emerged from extensive investigations with relevant humanitarian experts, summarized in: "Avoid AI if possible, Use AI systems that are contextually-based, Empower and include local communities in AI initiatives, and Implement algorithmic auditing systems" (Wright and Verity 2020).

Framing the ambitions and concerns of the sector is a concrete first step to normalize the discussion on the potential contribution of this technology. It can also help in finding concrete applications that might represent a first, safe and secure step to start experimenting adequate risk mitigation frameworks and audit methodologies. In the meantime, knowing that most AI products are still barely scratching the surface of narrow AI, we could follow the distinction proposed by Robbins between explicability of steps and processes and explicability of a certain outcome. In his perspective, at this stage, the "how" did the system reach a certain conclusion is less important than "why" this conclusion was deemed valid. (Robbins 2019). Robbins' argument can be turned into a litmus test for humanitarian AI, temporarily abstaining from the use of opaque AI systems for all those specific decisions that require explicability by human standards to avoid harm (assuming that it is actually possible to define them in advance) (Robbins 2019; Wright and Verity 2020). In these cases, automation can still be an option; "however, this should be restricted to the old-fashioned kind of automation whereby the considerations are hard-coded into the algorithm" (Robbins 2019). Alternatively, humanitarians could settle for simpler versions of AI systems, where the trade-offs between efficiency gain and loss of control do not include the risk of harmful consequences for vulnerable persons.

Promoting improved legal frameworks

It seems inevitable, as foreseen by Schuppli, that

> [d]ecision-making by automated systems will produce new relations of power for which we have as yet inadequate legal frameworks or modes of political resistance and, perhaps even more importantly, insufficient collective understanding as to how such decisions will actually be made and upon what grounds [...] demands for public accountability and oversight will require much greater participation in the epistemological frameworks that organize and manage these new techno-social systems, and that may be a formidable challenge for all of us (Schuppli 2014).

While Schuppli fears the "closure of a certain 'epistemology of facts'[...] cloaked under a veil of secrecy called 'national security interests'" (Schuppli 2014), a similar concern also applies to the epistemology of principles within the humanitarian sector.

We consider reasonable, as proposed by some scholars, to envisage that these new relations of power are preventively regulated by adequate rules of engagement with a projective sense of the law and inspired by the Geneva Conventions, instead of adopting the frame of The Hague Conventions (Schuppli 2014; Lapadula 2019). Modifying international legal instruments (such as the Conventions) to add references to technical features would be a time-consuming and politically sensitive process. In addition, it would also risks carrying within itself the mark of obsolescence that comes from entrusting protection from an extremely technical and obscure risk generated by a fast-developing technology to a slow-moving, policy-oriented system.

To mitigate this problem, we recommend that humanitarian organizations endorse enforceable standards maintained by professional organizations. An example is the P 7001 currently being explored by IEEE (Bryson and Winfield 2017), aiming to create a standard for measurable, testable levels of transparency, so that autonomous systems can be objectively assessed and levels of compliance determined. Or the TR 24368 being proposed by ISO (Naden 2019), designed to provide an overview of ethical and societal concerns of AI. Some, like legal scholar Andrew Murray, invoke international cooperation on the standards of regulation under a UN-like global standard-setting body to avoid standards being designed to be beneficial to regional industries rather than communities and individuals (Van Den Meerssche 2020).

In the meantime, in some regions, individuals are not left completely without protection. We already mentioned that recently, the District Court of The Hague recognized and actioned the interplay of GDPR and European Convention on Human Rights in protecting rights of individuals exposed to automated digital systems (Rizzi and Pera 2020). We believe that while the other steps are taken, humanitarian organizations can safely align their behavior to these existing legal frameworks to maximize the protection of the individuals covered by their digital systems.

Designing humanitarian digital accountability for a complex ecosystem

As already mentioned, the concept of explicability as devised by Floridi and Cowls adds an accountability layer on top of the epistemological problem (Floridi and Cowls 2019). However, the practical implementation of such a system, a mechanism bringing accountability at each and every step of the complex of chaotic behaviour of algorithms, is still being investigated.

Semi-autonomous systems feature a complex variety of components, be them physical (e.g., project managers) or immaterial (e.g., industrial practices and legacies), thus making it hardly imaginable to hold a single individual or entity accountable when something goes wrong (Ganesh 2018). As noted by Schuppli, "[c]omplex systems are rarely, if ever, the product of single authorship; nor do humans and machines operate in autonomous realms" (Schuppli 2014). Lessons can be taken from existing industries, such as aviation, where shared and distributed accountability for errors in complex technical systems is accepted and regulated (Galison and Roland 2000; Vaughan 1997; both referenced in Ganesh 2018).

Much more complicated is the balancing act of computer agency and moral responsibility. In the field of computer sciences and automation, it has been observed that if computer systems can diminish users' senses of their own moral agency and responsibility, "this would lead to erosion of accountability" (Cummings 2006). In this case, the inherent complexity of socio-technical systems can result in a moral buffering effect on the user decision maker, as decision support systems that integrate higher levels of automation can possibly lead them to perceive the computer as a legitimate authority, diminish moral agency, and shift accountability to the computer (Cummings 2006). In some instances it can also accentuate [human] confirmation bias and [machine] automation bias (Goddard et al 2012) leading humanitarians to over trust machine results based on their own biases. In some other cases, the effect can be accentuated by user interface choices, user experience journeys, or even dark patterns, that under the pretext of relieving users from the stress of their tasks, can accentuate a sense of levity in taking decisions that can then result in potential harm to people.

Examining algorithms and AI-based systems from an anthropological viewpoint allowed to identify such risks,[11] often hidden in full sight under the folds of commonly accepted practices among technology developers and users. This has led to a critique of the simplistic human/machine dichotomy and instead proposed non-binary lenses for examining AI that could be relational, communal, or intersectional (Kelliher et al. 2018). We consider that a change in the narratives of AI, framed under the proposed concept of speculative AI (a form of speculative and critical design), is thus required to create "situated communal AI knowledge systems, with distributed loci of control, access, and accountability" (Kelliher et al. 2018).

Beyond the different ways in which complexity in AI and ADMs contribute to reducing **both humanitarian and algorithimic** accountability, there is one aspect of it that does not depend on technical designs or agents' perceptions and behaviors. As already noted, the humanitarian sector does not represent a virtuous example of transparency in policies for redress. In line with the recommendations by the EU High-Level Expert Group on Artificial Intelligence (EU High-Level Expert Group on Artificial Intelligence 2019), we believe that establishing clear, public mechanisms for compensation, redress, reparation, restitution, and recognition of eventual harm done to individuals or communities is a necessary step in the direction of upholding the spirit of the humanitarian principles. We also register that these accountability policies do not and shall not depend on advances in explainability of AI systems, nor in the development of additional legal instruments, but on the political will of each organization to set up adequate systems.

Appendix

List of websites analyzed for policies on redress for harm from the use of digital or technology

Adeso. https://adesoafrica.org/who-we-are/mission-vision-values/index.htm. Accessed on 30\12\2020

Ashoka. https://www.ashoka.org/en-se/organizational-accountability. Accessed on 30\12\2020

Care. https://www.care.org/accountability-and-transparency. Accessed on 30\12\2020

Charity Water. https://www.charitywater.org/about. Accessed on 30\12\2020

Danish Refugee Council. https://drc.ngo/relief-work/concerns-complaints/code-of-conduct. Accessed on 30\12\2020

[11] See also the research done on algorithmic impact assessments, where impacts are constructed as close as possible to actual harm (Metcalf et al. 2021).

FHI360. https://www.fhi360.org/about-us/compliance-office. Accessed on 30\12\2020

Heifer. https://www.heifer.org/about-us/inside-heifer/index.html. Accessed on 30\12\2020

International Committee of the Red Cross. https://www.icrc.org/en/document/code-conduct-employees-icrc. Accessed on 30\12\2020

International Rescue Committee. https://www.rescue.org/page/our-code-conduct. Accessed on 30\12\2020

Médecins Sans Frontières. https://www.msf.org/who-we-are. Accessed on 30\12\2020

NEAR. http://near.ngo/. Accessed on 30\12\2020

Norwegian Refugee Council. https://www.nrc.no/who-we-are/accountability/. Accessed on 30\12\2020

Oxfam. https://www.oxfam.org/en/what-we-do/about/safeguarding. Accessed on 30\12\2020

Plan International. https://plan-international.org/organisation/accountability-policies-commitments. Accessed on 30\12\2020

Save the Children. https://www.savethechildren.net/about-us/accountability. Accessed on 30\12\2020

Seeds. https://www.seedsindia.org/policies/ . Accessed on 30\12\2020

UN Women. https://www.unwomen.org/en/about-us/accountability. Accessed on 30\12\2020

UNHCR. https://www.unhcr.org/5e21d0cb4. Accessed on 30\12\2020

UNICEF. https://www.unicef.org/innovation/what-we-do-new. Accessed on 30\12\2020

World Food Program. https://www.wfp.org/oversight. Accessed on 30\12\2020

World Vision International. https://www.wvi.org/accountability . Accessed on 30\12\2020

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41018-021-00096-6.

Additional file 1: Annex 1. – An early proposal for a tentative humanitarian AI Ethics framework for future research and development.

Acknowledgments

We thank Dragana Kaurin, Chris Earney, and Shireen Al-Chalabi for their comments to the first draft and to the reviewers for their substantial contribution in improving this article. We are also thankful for the information sharing on AI and ethics done by the online community of experts including Dorothea Baur, Abeba Birhane, Rumman Chowdhury, Joanna J Bryson, and many others. Your tireless and selfless dedication to sharing knowledge was instrumental in laying the ground for many of the questions and sources explored in this article.

Authors' contributions

All authors read and approved the final manuscript.

Funding

No funding was received for this research. All authors are supported by their respective institutions.

Availability of data and materials

The datasets analyzed during the current article are available in the following Github repository: Technology Diffusion dataset. https://github.com/owid/owid-datasets/tree/master/datasets/Technology%20Diffusion%20-%20Comin%20and%20Hobijn%20(2004)%20and%20others. Accessed 06 February 2020,

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Norwegian Refugee Council, Oslo, Norway. [2]UNHCR Innovation Service, Geneva, Switzerland.

Received: 14 July 2020 Accepted: 9 April 2021
Published online: 06 October 2021

References

Alston P (2019a) Brief as amicus curiae before the District Court of the Hague on the case of NJCM c.s./De Staat der Nederlanden (SyRI), case No. C/09/550982/ HA ZA 18/388. https://www.ohchr.org/Documents/Issues/Poverty/Amicusfinalversionsigned.pdf. Accessed on 30 Dec 2020.

Alston P (2019b) Extreme poverty and human rights. Report submitted in accordance with Human Rights Council resolution 35/19, UN - General Assembly. https://undocs.org/pdf?symbol=en/A/HRC/41/39/Add.1 . Accessed on 30 Dec 2020.

Balsari S (2019) Will AI help universalize health care? the BMJ. https://blogs.bmj.com/bmj/2019/09/23/satchit-balsari-will-ai-help-universalize-health-care/. Accessed 17 Mar 2021

Bellamy R, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilovic A, Nagar S, Ramamurthy K, Richards JT, Saha D, Sattigeri P, Singh M, Varshney K, Zhang Y (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. ArXiv, abs/1810.01943

Bengio Y, et al. (2019) A meta-transfer objective for learning to disentangle causal mechanisms. ArXiv e-print. https://arxiv.org/abs/1901.10912. Accessed on 10 Feb 2020.

Brenna F, Goyal M, Danesi G, Finch G, Goehring B (2018) Shifting toward Enterprise-grade AI - Resolving data and skills gaps to realize value. IBM Corporation. https://www.ibm.com/downloads/cas/QQ5KZLEL. Accessed 23 Mar 2021

Brookland J (2019) Revolutionising recruitment: a test for AI in the United Nations. UNHCR Innovation Service. https://medium.com/unhcr-innovation-service/revolutionising-recruitment-a-test-for-ai-in-the-united-nations-4456df0b1431. Accessed on 07 Feb 2021.

Brundage M, et al. (2018) The malicious use of artificial intelligence: forecasting, prevention, and mitigation. University of Oxford. https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf. Accessed on 21 Mar 2020

Bryson J, Winfield A (2017) Standardizing ethical design for artificial intelligence and autonomous systems. Computer 50(5):116–119. https://doi.org/10.1109/MC.2017.154

Burack J (2020) Addressing algorithmic discrimination in the European Union. A path for Europe. https://pathforeurope.eu/addressing-algorithmic-discrimination-in-the-european-union/. Accessed on 30 Dec 2020.

Burrell J (2016) How the machine "thinks": understanding opacity in machine learning algorithms. Big Data Soc 3(1):205395171562251. https://doi.org/10.1177/2053951715622512

Cardia et al (2017) Towards a principled approach to humanitarian information and communication technology. In: ICTD '17: proceedings of the ninth international conference on information and communication technologies and development, article no.: 23. https://doi.org/10.1145/3136560.3136588

Carter S, et al. (2019) Activation Atlas. Distill. https://distill.pub/2019/activation-atlas/. Accessed on 12 Feb 2020.

Castellanos S, Nash K (2018) Bank of America confronts AI's "Black Box" With Fraud Detection Effort, Wall Street Journal. https://www.wsj.com/articles/bank-of-america-confronts-ais-black-box-with-fraud-detection-effort-1526062763. Accessed 23 Mar 2021

Cearly D (2019) Top 10 Strategic Technology Trends for 2019: AI-Driven Development. Gartner Research. ID G00377677

Cebotarean E (2011) Business intelligence. Journal of Knowledge Management, Economics and Information Technology. http://www.scientificpapers.org/wp-content/files/1102_Business_intelligence.pdf. Accessed on 23 Jan 2021.

Choudhury A (2019) Explainability vs. interpretability in artificial intelligence and machine learning. Analytics India Magazine. https://analyticsindiamag.com/explainability-vs-interpretability-in-artificial-intelligence-and-machine-learning/. Accessed on 10 Feb 2020.

Chui M, et al. (2018) Notes from the AI frontier: applications and value of deep learning. McKinsey Global Institute. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning. Accessed on 10 Feb 2020.

Cilliers P (2002) Complexity and postmodernism understanding complex systems. Routledge. https://uberty.org/wp-content/uploads/2015/04/Paul-Cilliers-Complexity-and-Postmodernism-Understanding-Complex-Systems-1998.pdf. Accessed 13 Mar 2021

Comes T (2016) Cognitive biases in humanitarian sensemaking and decision-making lessons from field research, pp 56–62. https://doi.org/10.1109/COGSIMA.2016.7497786

Committee on Standards in Public Life (2020) Artificial intelligence and public standards. Review. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/868284/Web_Version_AI_and_Public_Standards.PDF. Accessed on 01 May 2020

Complete guide to GDPR compliance. https://gdpr.eu/. Accessed on 30 Dec 2020.

Copeland B J (2019) Artificial intelligence. Encyclopædia Britannica. https://www.britannica.com/technology/artificial-intelligence. Accessed on 20 Feb 2020.

Cummings ML (2006) Automation and accountability in decision support system interface design. J Technol Stud http://scholar.lib.vt.edu/ejournals/JOTS/v32/v32n1/cummings.html. Accessed on 01\05\2020

Das A, Rad P (2020) Opportunities and challenges in explainable artificial intelligence (XAI): a survey, arXiv:2006.11371v2 [cs.CV] https://arxiv.org/pdf/2006.11371.pdf. Accessed on 30 Dec 2020

Dediu H, Comin D, Hobijn B (2004) Technology diffusion dataset. https://github.com/owid/owid-datasets/tree/master/datasets/Technology%20Diffusion%20-%20Comin%20and%20Hobijn%20(2004)%20and%20others. Accessed 06 Feb 2020

Development Pathways (2018) Targeting humanitarian aid: something to be left to opaque algorithms? https://www.developmentpathways.co.uk/blog/targeting-humanitarian-aid-something-to-be-left-to-opaque-alogorithms/ Accessed on 26 Dec 2020

Diakopoulos N (2017) Algorithmic accountability reporting: on the investigation of black boxes. Tow Center for Digital Journalism, Columbia University. https://doi.org/10.7916/D8ZK5TW2

Dodgson K, et al. (2020) A framework for the ethical use of advanced data science methods in the humanitarian sector. The Humanitarian Data Science and Ethics Group. https://www.hum-dseg.org/dseg-ethical-framework. Accessed on 01 May 2020

Dreyfus HL, Hubert L (1992) What computers still can't do: A critique of artificial reason. MIT press, Cambridge

Engler A C (2021) Independent auditors are struggling to hold AI companies accountable. FastCompany. https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue. Accessed on 02 Feb 2021.

EU General Data Protection Regulation (2016) https://gdprinfo.eu/. Accessed on 30 Dec 2020.

EU High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy artificial intelligence. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Accessed on 30 Dec 2020.

Eubanks V (2018) Automating inequality: how high-tech tools profile, police, and punish the poor. St. Martin's Press ISBN: 9781250074317

Floridi L, Cowls J (2019) A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review 1(1). https://doi.org/10.1162/99608f92.8cd550d1. Accessed 23 Mar 2021

Floridi L, Cowls J, King TC, Taddeo M (2020) How to Design AI for Social Good: Seven Essential Factors. Science and Engineering Ethics 26(3):1771-1796. https://doi.org/10.1007/s11948-020-00213-5

Friedman M (1970) The social responsibility of business is to increase its profits. The New York Times Magazine. https://web.archive.org/web/20060207060807/https://www.colorado.edu/studentgroups/libertarians/issues/friedman-soc-resp-business.html. Accessed on 23 Feb 2020.

Frosst N, Hinton G (2017) Distilling a neural network into a soft decision tree. arXiv preprint arXiv:1711.09784.

Galison P, Roland A (Eds.) (2000) Atmospheric Flight in the Twentieth Century. Springer, Netherlands. https://www.springer.com/gp/book/9780792360377

Ganesh MI (2018) A-words: Accountability, Automation, Agency, AI. https://medium.com/the-state-of-responsible-iot-2018/a-words-accountability-automation-agency-ai-3fb5beb93739. Accessed 23 Mar 2021

Gent E, (2019) Where should we draw the line between rejecting and embracing black box AI? Interview to Elizabeth Holm, in Singularity Hub. https://singularityhub.com/2019/04/17/in-defense-of-black-box-ai/. Accessed on 10 Feb 2020.

Gisel L (2016) The principle of proportionality in the rules governing the conduct of hostilities under international humanitarian law. ICRC. https://www.icrc.org/en/download/file/79184/4358_002_expert_meeting_report_web_1.pdf . Accessed on 30 Dec 2020.

Goddard K, Roudsari A, Wyatt JC (2012) Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 19(1): 121-127

Goodman B, Flaxman S (2019) European Union regulations on algorithmic decision-making and a "right to explanation". https://arxiv.org/abs/1606.08813

Greenwood F, Raymond N, Scarnecchia D, Poole D, and Howarth C (2017) The Signal Code: a human rights approach to information during crisis. Signal Program at Harvard Humanitarian Initiative. https://hhi.harvard.edu/publications/signal-code-human-rights-approach-information-during-crisis. Accessed 8 Feb 2020.

Gruskin S, Dickens B (2006) Human rights and ethics in public health. Am J Public Health 96(11):1903–1905. https://doi.org/10.2105/AJPH.2006.099606

Handelman GS et al (2019) Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. Am J Roentgenol 212(1):1–6. https://doi.org/10.2214/AJR.18.20224

Heidegger M, (1954 English Translation 1977) The question concerning technology," pp. 3-35. https://www.futurelearn.com/courses/philosophy-of-technology/0/steps/26315. Accessed 06 Feb 2020,

Humanitarian Congress Berlin (2018) Video of day 1. https://humanitarian-congress-berlin.org/2018/. Accessed on 08 Oct 2019.

IBM (2020) Deep learning. IBM Cloud Learning Hub. https://www.ibm.com/cloud/learn/deep-learning Accessed on 28 Dec 2020

ICO (2020) What is automated individual decision-making and profiling? Information Commissioners Office. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling/what-is-automated-individual-decision-making-and-profiling/ Accessed on 27 Dec 2020

ICRC (2006) Business and international humanitarian law: an introduction to the rights and obligations of business enterprises under international humanitarian law. https://www.icrc.org/en/publication/0882-business-and-international-humanitarian-law-introduction-rights-and-obligations. Accessed on 23 Feb 2020.

ICRC (2019) Artificial intelligence and machine learning in armed conflict: a human-centred approach. https://www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach. Accessed on 21 Mar 2020

ICRC (2020) Study on the use of the emblems: operational and commercial and other non-operational issues. https://www.icrc.org/en/publication/4057-study-use-emblems-operational-and-commercial-and-other-non-operational-issues. Accessed on 30 Dec 2020.

ICRC and Privacy International (2018) The humanitarian metadata problem: "doing no harm" in the digital era. https://privacyinternational.org/report/2509/humanitarian-metadata-problem-doing-no-harm-digital-era. Accessed on 23 Feb 2020.

ICRC, The Engine Room and Block Party (2017) Humanitarian futures for messaging apps. https://shop.icrc.org/humanitarian-futures-for-messaging-apps-print-en. Accessed on 30 Dec 2020.

ICT4D (2019) Highlights from the 2019 ICT4D Conference. https://www.ict4dconference.org/about/highlights-2019-ict4d-conference/. Accessed on 08 Feb 2020

Ingwin, J., Larson, J., Mattu, S., Kirchner, L. (2016) Machine bias. ProPublica https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed on 30 Dec 2020.

Jacovi A, Goldberg Y (2020) Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness? In: Proceedings of the 58th annual meeting of the Association for Computational Linguistics, pp 4198–4205 https://www.aclweb.org/anthology/2020.acl-main.386.pdf. Accessed on 30 Dec 2020

Kaspersen A, Lindsey-Curtet C (2016) The digital transformation of the humanitarian sector. ICRC Humanitarian Law & Policy. https://blogs.icrc.org/la w-and-policy/2016/12/05/digital-transformation-humanitarian-sector/. Accessed on 20 Feb 2020

Keats Citron D. (2007) Technological Due Process. U of Maryland Legal Studies Research Paper No. 2007-26. Washington University Law Review 85:1249-1313. Available at SSRN: https://ssrn.com/abstract=1012360

Keller P, Duguay F, Precup D (2004) Redagent: winner of TAC SCM 2003. ACM SIGecom Exchanges 4(3):1–8. https://doi.org/10.1145/1120701.1120703

Kelliher A et al. (2018) Beyond black boxes: tackling artificial intelligence as a design material. Conference paper. Design Research Society Conference 2018. Shared by the author

Kelly J (2020) Coinbase won't allow discussions of politics and social causes at work—if employees don't like it, they're free to leave. Forbes. https://www.forbes.com/sites/jackkelly/2020/10/01/coinbase-wont-allow-discussions-of-politics-and-social-causes-at-work-if-employees-dont-like-it-theyre-free-to-leave/?sh=56e674a07459 . Accessed on 30 Dec 2020.

Keynon, M (2018). Bots at the Gate a human rights analysis of automated decision making in Canada's immigration and refugee system. Citizen Lab. https://citizenlab.ca/2018/09/bots-at-the-gate-human-rights-analysis-automated-decision-making-in-canadas-immigration-refugee-system/ Accessed on 26 Dec 2020

Knight W (2017) The dark secret at the heart of AI. MIT Technology Review, May\June 2017. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/. Accessed on 09 Feb 2020

Knight W (2019) An AI pioneer wants his algorithms to understand the 'Why'. Wired. https://www.wired.com/story/ai-pioneer-algorithms-understand-why/. Accessed on 09 Feb 2020

Kyriazi S (2019) UNHCR's newest artificial intelligence engineer on bias, coding, and representation. UNHCR Innovation Service. https://medium.com/unhcr-innovation-service/unhcrs-newest-artificial-intelligence-engineer-on-bias-coding-and-representation-3363c432dd98. Accessed on 08 Feb 2020.

Labbé J, Daudin P (2015) Applying the humanitarian principles: reflecting on the experience of the international committee of the red cross. Int Rev Red Cross 97:1–28. https://doi.org/10.1017/S1816383115000715

Lapadula J (2019) Interview: data privacy, distributed denial of service attacks, and human rights: a conversation with Nathaniel Raymond. https://www.yalejournal.org/publications/interview-data-privacy-distributed-denial-of-service-attacks-and-human-rights-a-conversation-with-nathaniel-raymond. Accessed on 30 Dec 2020.

Lepri B et al (2017) Fair, transparent, and accountable algorithmic decision-making processes: the premise, the proposed solutions, and the open challenges. Philos Technol 31. https://doi.org/10.1007/s13347-017-0279-x https://link.springer.com/article/10.1007/s13347-017-0279-x. Accessed on 27/12/2020

Lerman R (2018) Microsoft to invest $40 million in AI technology for humanitarian issues. The Mercury News. https://phys.org/news/2018-09-microsoft-invest-million-ai-technology.html. Accessed on 20 Feb 2020

London AJ (2019) Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hast Cent Rep 49(1):15–21. https://doi.org/10.1002/hast.973

Luengo-Oroz M (2019) Solidarity should be a core ethical principle of AI. Nat Mach Intell 1(11):494-494. https://www.nature.com/articles/s42256-019-0115-3

Lum K, Isaac W (2016) To predict and serve? Significance 13(5):14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Madianou M (2019) Technocolonialism: digital innovation and data practices in the humanitarian response to refugee crises. Soc Media Soc. https://doi.org/10.1177/2056305119863146

Market Research.biz (2020) Global artificial intelligence in military market analysis, drivers, restraints, opportunities, threats, trends, applications, and growth forecast to 2028. Market Research.biz. https://marketresearch.biz/report/artificial-intelligence-in-military-market/. Accessed on 20 Feb 2020.

Metcalfe V, Martin E, Pantuliano P (2011) Risk in humanitarian action: towards a common approach? Humanitarian Policy Group. https://cdn.odi.org/media/documents/6764.pdf

Metcalf J et al (2021) Algorithmic impact assessments and accountability: the co-construction of impacts. In: ACM conference on fairness, accountability, and transparency (FAccT '21) March 3–10, 202. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3736261. Accessed on 11 Feb 2021

Mijente (2019) Palantir played key role in arresting families for deportation, document shows. Mijente. https://mijente.net/2019/05/palantir-arresting-families/. Accessed on 28 June 2020

Miller Devens R (1865). Business intelligence. In Cyclopaedia of commercial and business anecdotes; comprising interesting reminiscences and facts, remarkable traits and humors of merchants, traders, bankers Etc. in all ages and countries. D. Appleton and company. p. 210. https://archive.org/details/cyclopaediacomm00devegoog/page/n262 . Accessed on 23 Jan 2021.

Molnar C (2020a) Interpretable machine learning. A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/index.html. Accessed on 30 Dec 2020

Molnar P (2020b) Technological testing grounds. EDRI. https://edri.org/wp-content/uploads/2020/11/Technological-Testing-Grounds.pdf. Accessed on 30 Dec 2020.

Molnar P, Gill L (2018) Bots at the Gate. University of Toronto & The Citizen Lab. https://ihrp.law.utoronto.ca/sites/default/files/media/IHRP-Automated-Systems-Report-Web.pdf. Accessed on 09 Feb 2020

Naden C (2019) It's all about trust. Isofocus. https://www.iso.org/news/ref2452.html. Accessed on 20 Mar 2020

Noorman M, Zalta E N (2014) Computing and moral responsibility. In: Zalta EN (ed.) The Stanford Encyclopaedia of philosophy. http://plato.stanford.edu/archives/sum2014/entries/computing-responsibility. Accessed 06 Feb 2020.

Norwegian Ministry of Foreign Affairs (2016, updated in 2019) Ensuring respect for the humanitarian principles: guidance note for support provided from the Norwegian Ministry of Foreign Affairs to NGOs. https://www.regjeringen.no/en/dokumenter/note-humanitarian-principles/id2568659/. Accessed on 23 Feb 2020.

NYC AMPO (2020) Agency Compliance Report. NYC. https://www1.nyc.gov/assets/ampo/downloads/pdf/AMPO-CY-2020-Agency-Compliance-Reporting.pdf. Accessed on 03 Feb 2021.

OCHA (2019). Catalogue of predictive analytics models in the humanitarian sector. United Nations Office for the Coordination of Humanitarian Affairs. Centre for Humanitarian Data. https://centre.humdata.org/catalogue-for-predictive-models-in-the-humanitarian-sector/ Accessed on 27 Dec 2020

OCHA (2020). Anticipatory action in Bangladesh before peak monsoon flooding. United Nations Office for the Coordination of Humanitarian Affairs. Centre for Humanitarian Data. https://centre.humdata.org/anticipatory-action-in-bangladesh-before-peak-monsoon-flooding/ Accessed on 26 Dec 2020

OECD (2020) Review into bias in algorithmic decision-making. Center for Data Ethics and Innovation. https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making. Accessed on 30 Dec 2020.

Page J, Bain M, Mukhlish F (2018) The risks of low level narrow artificial intelligence. In: 2018 IEEE international conference on intelligence and safety for robotics (ISR). https://doi.org/10.1109/IISR.2018.8535903

Parker B (2019) New UN deal with data mining firm Palantir raises protection concerns. The New Humanitarian. https://www.thenewhumanitarian.org/news/2019/02/05/un-palantir-deal-data-mining-protection-concerns-wfp. Accessed on 30 Dec 2020.

Pasquale F (2016) The black box society: the secret algorithms behind money and information. Harvard University Press, Cambridge

Perdicoulis A (2016) Action-to-outcome maps in impact assessment. The Systems Thinker. https://thesystemsthinker.com/action-to-outcome-maps-in-impact-assessment/. Accessed on 01 May 2020

Pictet J (1979) The fundamental principles of the Red Cross: commentary. ICRC. https://www.icrc.org/en/doc/resources/documents/misc/fundamental-principles-commentary-010179.htm. Accessed on 23 Feb 2020.

Pizzi M, Romanoff M, Engelhardt T (2020) AI for humanitarian action: Human rights and ethics. Int Rev Red Cross 102(913):145-180

Polack P (2020) Beyond algorithmic reformism: forward engineering the designs of algorithmic systems. Big Data Soc 7(1):205395172091306. https://doi.org/10.1177/2053951720913064

Principles for Digital Development (2015) https://digitalprinciples.org/. Accessed on 30 Dec 2020.

Ramaraj P (2010) Information systems flexibility in organizations: conceptual models and research issues. Glob J Flex Syst Manag 11(1-2):1–12. https://doi.org/10.1007/BF03396574

Raymond N A, and Card B L (2015) Applying humanitarian principles to current uses of information communication technologies: gaps in doctrine and challenges to practice. Harvard Humanitarian Initiative. https://hhi.harvard.edu/sites/default/files/publications/signal_program_humanitarian_principles_white_paper.pdf. Accessed on 23 Feb 2020.

Regis E (2020) The enigma of aerodynamic lift. Sci Am 322. https://doi.org/10.1038/scientificamerican0220-44

Rizzi FT, Pera A (2020) Balancing tests as a tool to regulate artificial intelligence in the field of criminal law. In: Special collection on artificial intelligence UNICRI. http://www.unicri.it/node/3228. Accessed on 30 Dec 2020

Robbins S (2019) A Misdirected Principle with a Catch: Explicability for AI. Minds Mach 29(4):495-514. https://doi.org/10.1007/s11023-019-09509-3

Rolle B, Lafontaine E (2009) The emblem that cried wolf: ICRC study on the use of the emblems. Int Rev Red Cross 91(876):759-778. https://doi.org/10.1017/S1816383110000172

Russell S, Norvig P (2010) Artificial intelligence: a modern approach, 3rd edn. Prentice Hall ISBN 978-0-13-207148-2 https://www.cin.ufpe.br/~tfl2/artificial-intelligence-modern-approach.9780131038059.25368.pdf Accessed on 27 Dec 2020

Sandvik K, Jacobsen K, McDonald S (2017) Do no harm: a taxonomy of the challenges of humanitarian experimentation. Int Rev Red Cross 99(904):319–344. https://doi.org/10.1017/S181638311700042X

Schuppli S (2014) Deadly algorithms - can legal codes hold software accountable for code that kills? Radical philosophy 187. http://www.susanschuppli.com/wp-content/uploads/2014/11/Deadly-Algorithms.pdf. Accessed on 01 May 2020

Shankar R, et al. (2020) Failure modes in machine learning systems. Arxiv eprint. https://arxiv.org/abs/1911.11034v1. Accessed on 28 June 2020

Singh A (2019) Artificial Intelligence and International Security: The Long View. Cambridge University Press. Journal of Ethics & International Affairs 33(2)

Slim H (1998) Sharing a universal ethic: the principle of humanity in war. Int J Human Rights 2(4):4–48. https://doi.org/10.1080/13642989808406759

Slim H (2015) Humanitarian ethics: a guide to the morality of aid in war and disaster. Oxford University Press, Oxford

Springer A, Hollis V, and Whittaker S (2017) Dice in the black box: user experiences with an inscrutable algorithm. Technical report. AAAI 2017 Spring Symposium on Designing the User Experience of Machine Learning Systems. https://aaai.org/ocs/index.php/SSS/SSS17/paper/view/15372/14580. Accessed on 22 Mar 2020

Technology diffusion dataset. (2004) Github. https://github.com/owid/owid-datasets/tree/master/datasets/Technology%20Diffusion%20-%20Comin%20and%20Hobijn%20(2004)%20and%20others. Accessed on 06 Feb 2020,

Theodorou A, Wortham RH, Bryson J (2017) Designing and implementing transparency for real time inspection of autonomous robots. Connect Sci 29(3):230–241. https://doi.org/10.1080/09540091.2017.1310182

Tsukerman E. Sound logic and monotonic AI models. Toptal. https://www.toptal.com/machine-learning/monotonic-ai-models. Accessed on 30 Dec 2020.

Turek M (2016) Explainable Artificial Intelligence (XAI). Defense Advanced Research Projects Agency (DARPA). United States Government. https://www.darpa.mil/program/explainable-artificial-intelligence#:~:text=XAI%20is%20one%20of%20a,to%20characterize%20real%20world%20phenomena

Turing AM (1950) Computing machinery and intelligence. Mind 49:433–460 https://www.csee.umbc.edu/courses/471/papers/turing.pdf. Accessed on 20 Feb 2020

UNESCO (2019). Artificial intelligence: towards a humanistic approach. Artificial intelligence with human values for sustainable development. https://en.unesco.org/artificial-intelligence. Accessed on 05 Jan 2021

UNICRI- INTERPOL (2019) Artificial intelligence and robotics for law enforcement. Report at the High-Level Meeting: Artificial Intelligence and Robotics-Reshaping the Future of Crime, Terrorism and Security. https://www.europarl.europa.eu/cmsdata/196207/UNICRI%20-%20Artificial%20intelligence%20and%20robotics%20for%20law%20enforcement.pdf . Accessed on 30 Dec 2020

United Nations (2018) Secretary-General's strategy on new technologies. United Nations. https://www.un.org/en/newtechnologies/images/pdf/SGs-Strategy-on-New-Technologies.pdf. Accessed on 09 Sep 2019.

United Nations (2019) Report of the special rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance. https://www.ohchr.org/EN/newyork/Documents/A-75-590-AUV.docx . Accessed on 30 Dec 2020.

United Nations (2020) Secretary-General's roadmap for digital cooperation. United Nations. https://www.un.org/en/content/digital-cooperation-roadmap/. Accessed on 28 June 2020

Van de Walle B, Comes T (2015) On the nature of information management in complex and natural disasters. In: Vidan A, Shoag D (eds) (2015) humanitarian technology: science, systems and global impact 2015, vol 107. HumTech2015, pp 403–411. https://doi.org/10.1016/j.proeng.2015.06.098 Accessed on 06 Feb 2020

Van Den Meerssche D (2020) 'The time has come for international regulation on artificial intelligence' – an interview with Andrew Murray. OpinioJuris. http://opiniojuris.org/2020/11/25/the-time-has-come-for-international-regulation-on-artificial-intelligence-an-interview-with-andrew-murray/. Accessed on 02 Feb 2021.

Vaughan D (1997) The Challenger Launch Decision. University of Chicago Press. https://www.press.uchicago.edu/ucp/books/book/chicago/C/bo22781921.html. Accessed 23 Mar 2021

Venkatasubramanian V (2019) The promise of artificial intelligence in chemical engineering: Is it here, finally?. AIChE J 65(2):466-478. https://doi.org/10.1002/aic.16489

Vigdor N (2019) Apple card investigated after gender discrimination complaints. New York Times. https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html. Accessed on 30 Dec 2020.

Vonèche Cardia I et al (2017) Towards a principled approach to humanitarian information and communication technology. https://doi.org/10.1145/3136560.3136588

Warshaw J et al (2015) Proceedings of the 33rd annual ACM conference on human factors in computing systems. https://doi.org/10.1145/2702123.2702274

Watson D and Floridi L (2020) The explanation game: a formal framework for interpretable machine learning. Ssrn. https://www.academia.edu/41652207/The_Explanation_Game_A_Formal_Framework_for_Interpretable_Machine_Learning. Accessed on 03 Feb 2021.

Wright J and Verity A (2020) Artificial intelligence principles for vulnerable populations in humanitarian contexts. DHNetwork. https://www.academia.edu/41716578/Artificial_Intelligence_Principles_For_Vulnerable_Populations_in_Humanitarian_Contexts. Accessed 17 Feb 2020

Zomignani Barboza J, Diver L, Jasmontaite L (2020) Aid and AI: the challenge of reconciling humanitarian principles and data protection. In: Privacy and identity management. Data for better living: AI and privacy, pp 161–176. https://doi.org/10.1007/978-3-030-42504-3_11

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.