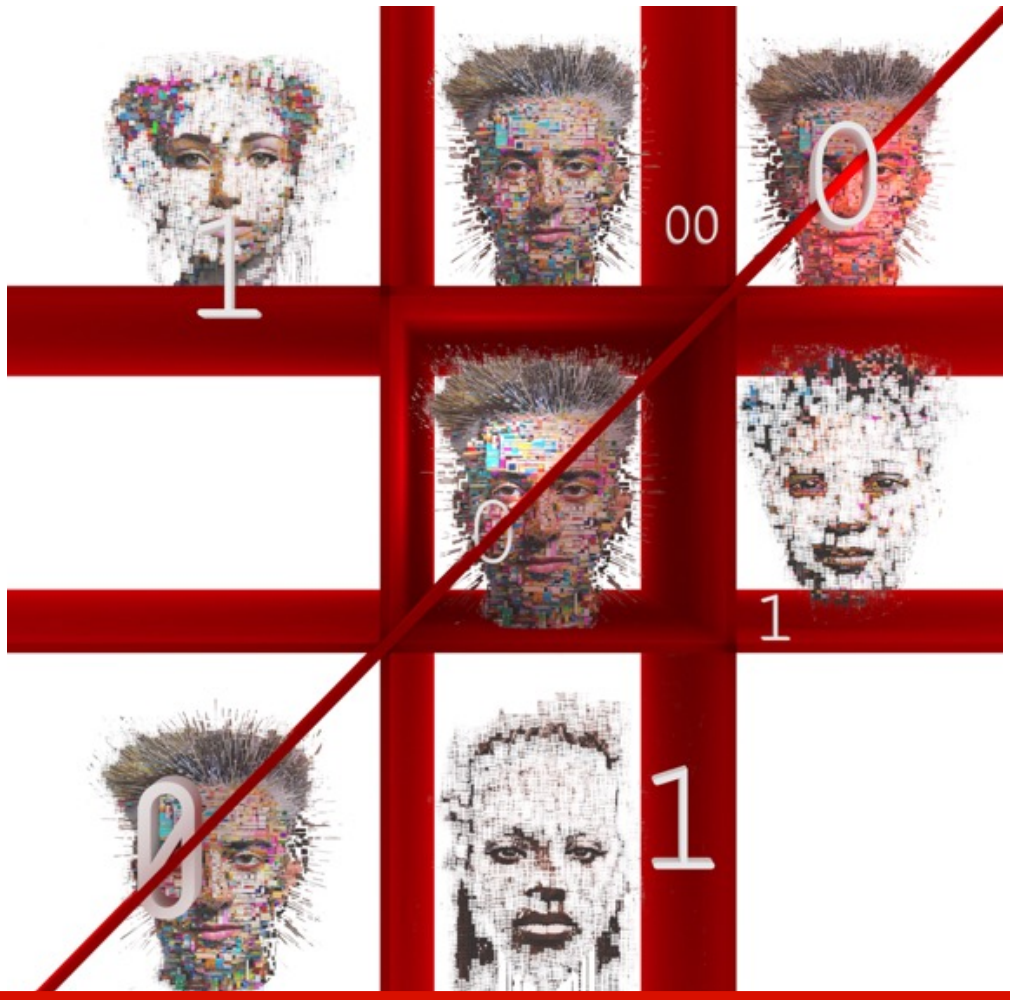
Impunity Watch

RAOUL WALLENBERG INSTITUTE



Overview Paper

Artificial Intelligence and Transitional Justice: Framing the Connections

January 2025

Co-funded by the European Union

Global Initiative Against Impunity for International Crimes and Serious Human Rights Violations

Ministry of Foreign Affairs of the Netherlands

Cover Photo: A tic-tac-toe board with human faces as digital blocks, symbolizing how AI works on pre-existing, biased online data for information processing and decision-making. Amritha R Warrier & AI4Media / Better Images of AI/ tic tac toe/ CC-BY 4.0.

Disclaimer

This publication was co-funded by the European Union and the Ministry of Foreign Affairs of the Netherlands. Its contents are the sole responsibility of Impunity Watch and do not necessarily reflect the views of the European Union or the Ministry of Foreign Affairs of the Netherlands.

Layout and Design

Wezank

Author

Dr Sue Anne Teo

Edited by Impunity Watch Policy and Innovation Team (Michelle Bouchebel and Thomas Unger)

About the Author

Dr Sue Anne Teo is a researcher on the Future of Human Rights project at the Raoul Wallenberg Institute of Human Rights and Humanitarian Law in Sweden. She holds a PhD in Law from the University of Copenhagen (2023) where her research focused on artificial intelligence and human rights. She was also a PhD Europaeum Scholar during the same period. Sue Anne also holds a Master of Law from the University of Cambridge and an MSc in Human Rights from the London School of Economics (LSE). She has taught courses on human rights, digital technology and democracy, and international law. Her work has been published in AI and Ethics, Law, Innovation and Technology, Computer Law and Security Review and the Nordic Journal of Human Rights, among others. Sue Anne is also a long-time human rights practitioner. Prior to her PhD, she worked at the Raoul Wallenberg Institute of Human Rights and Humanitarian Law as a Senior Programme Officer, served in a UN peacekeeping mission and also worked with the United Nations High Commissioner for Refugees and the Malaysian National Human Rights Commission.

About

Impunity Watch

Impunity Watch is an international non-profit organisation working with victims of violence to uproot deeply ingrained structures of impunity, deliver redress for grave human rights violations and promote justice and peace. We gather and share knowledge on priority themes, build partnerships and coalitions, and conduct international advocacy work to overcome impunity and transform justice. Impunity Watch currently works in Central America, North Africa and the Great Lakes region of Africa, the Middle East and the Western Balkans. The organisation also has a presence in Guatemala and Burundi. Impunity Watch's headquarters are based in The Hague in The Netherlands. Our work takes place at local as well as national, regional and international levels.
www.impunitywatch.org
info@impunitywatch.org

Raoul Wallenberg Institute of Human Rights and Humanitarian Law

The Raoul Wallenberg Institute of Human Rights and Humanitarian Law (RWI) is an independent academic institution established at Lund University in Sweden in 1984. We combine multi-disciplinary human rights research with education, support, and outreach to contribute to a wider understanding of, and respect for, human rights and international humanitarian law. We work on four thematic areas within human rights: rule of law and access to justice, international humanitarian law, human rights and the environment, and business and human rights. The institute is named after Raoul Wallenberg, a Swedish diplomat who saved tens of thousands of Jewish and other people at risk in Hungary during World War II.
https://rwi.lu.se/
info@rwi.lu.se

Global Initiative Against Impunity

Impunity Watch is one of eleven member organisations of the "Global Initiative Against Impunity for International Crimes and Serious Human Rights Violations: Making Justice Work." This civil society-led initiative aims to combat the growing climate of impunity for core international crimes and serious human rights violations by promoting comprehensive justice and accountability. This report is one of several key activities of Impunity Watch within the Global Initiative and is co-funded by the European Union, and the Ministry of Foreign Affairs of the Netherlands.
makingjusticework@fidh.org

ACKNOWLEDGEMENTS

Impunity Watch extends its deepest gratitude to Dr Sue Anne Teo for her insightful and comprehensive work in authoring the overview paper "Artificial Intelligence and Transitional Justice: Framing the Connections." This paper represents a significant step in exploring the intersections of AI and transitional justice.

We are thankful to the Raoul Wallenberg Institute of Human Rights and Humanitarian Law for their invaluable partnership and support throughout this initiative. Our appreciation goes to the participants of the AI and Transitional Justice expert meeting held at The Delft University of Technology (TU Delft) on November 6, 2024, organised by Impunity Watch in collaboration with the AI Futures Lab at TU Delft. Experts came from multi-disciplinary backgrounds, representing a range of transitional justice and human rights practitioners, legal scholars, AI designers, and technology experts, and brought together a wealth of diverse expertise, contributing to the paper's development.

Experts were affiliated with key institutions that are relevant to making the connection between AI and TJ , including the United Nations Office of the High Commissioner for Human Rights, the Harvard Humanitarian Initiative, TU Delft, Utrecht University, Columbia University, the University of Cambridge, the Council of Europe, the University of Quebec in Montreal, the International Criminal Court Trust Fund for Victims, the UN Independent Institution for Missing Persons in Syria, SAP, and the NGO Mnemonic.

A special thank you is extended to Dr Bernard Duhaime, Special Rapporteur on truth, justice, reparation, and guarantees of non-recurrence, for his active participation and invaluable insights as he is in the process of preparing a thematic report on the topic of AI and transitional justice.

Finally, Impunity Watch is grateful for the support of the European Union and the Ministry of Foreign Affairs of the Netherlands, whose support made this publication possible. We would also like to extend our heartfelt thanks to Anita Rice for her meticulous copy editing of the paper.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Artificial intelligence (AI) is ubiquitous in society today. It is said to hold the promise of revolutionising scientific discovery, addressing climate change, and improving public services and human welfare in general. However, harms associated with AI have also been highlighted – including its impacts on the right to privacy, non-discrimination, data protection, freedom of expression and human dignity.

In transitional justice settings, societies are confronted with a legacy of serious human rights violations and abuses, and are now also increasingly confronted with the impact of new AI technologies. Given that transitional justice is a morally and ethically contested field, considering the implications of AI is a timely exercise.

This overview paper maps and frames the connections between AI and transitional justice – identifying the potential harms and emerging concerns in relation to how AI can impact the field of transitional justice and its victim-centred approach. At the same time, the paper examines AI's positive role in transitional justice practices such as documentation and archiving as a tool for victim agency and empowerment in general. The overarching guiding perspective adopted is the victim-centred approach. In the context of AI, this entails ensuring the ideas, actions and implementation of AI in the transitional justice setting are guided and informed by the needs, choices and meaningful participation of victims.

The paper finds that existing harms posed by the design and use of AI can be amplified and exacerbated in the transitional justice context, noting especially its operations in post-conflict or post-authoritarian environments. These societies face a legacy of serious human rights violations and abuses, and the paper looks at the sensitive datasets involved and how AI can be used to destabilise the information sphere, potentially impacting truth-telling and peacebuilding efforts. At the same time, AI can be used for good in the transitional justice context. The potential for AI to improve workflow by optimising efficiencies around documentation, data analysis, outreach and awareness-raising, and victim and public engagement remain under examined and under-utilised. In this sense, actors working in the field of transitional justice can learn from the experiences of how AI has been used in adjacent contexts of human rights and conflict monitoring and adopt best practices, lessons learned and explore partnerships where feasible.

AI policy and regulatory processes can, however, also learn from transitional justice experiences. Much regulation in the area of AI and human rights centres around protecting against potential harms and using the law to punish acts that lead or contribute to harm. However, a key transitional justice perspective can facilitate the promotion of human rights and non-recurrence. Instead of merely avoiding harms, AI governance can also aim to promote agency by having in place

mechanisms, procedures and policies that enable people to exercise and enjoy their human rights, rather than merely avoiding the worst instances of harm.

This paper proposes thinking about AI and transitional justice from three different facets – focusing on *People*, *Practice* and *Policy*. This means putting people (i.e. victims) at the centre, considering the benefits and risks of AI in practical transitional justice contexts and, finally, policy recommendations in this area. Taking a multipronged approach also builds on a key finding of the paper that relying upon a regulatory response alone is insufficient.

The key takeaways and recommendations are as follows:

1. Key takeaways

- Despite AI having negative human rights and ethical impacts, there is increasing adoption of "AI for Good", including through high-level efforts at the UN.
- Decision-making around the use and adoption of AI in the transitional justice context should always be guided by foundational values such as victim-centredness. Efficiency, cost, time savings and other gains typically associated with the use of AI should be guided by the objective of being in service of the needs and aspirations of victims, respect for human dignity and be guided by the goals of transitional justice.
- The paper proposes three ways to consider where and how AI can play a role in the transitional justice context – the three Ps: centring the *persons* (victims), examining AI in the practical setting (*practice*) of transitional justice and, finally, *policy*. Key considerations include the need to adopt a human rights-based approach and a careful analysis of the benefits and burdens before adoption of AI applications.
- Data security concerns should be a primary consideration, especially when sensitive data of vulnerable populations are used or stored in the cloud or used in training AI models. Actors working with data processing need to put in place adequate data protection and cybersecurity measures.
- Multidisciplinary approaches are vital for fostering collaboration between transitional justice and AI expertise. Platforms like the Delft initiative organised by Impunity Watch and TU- Delft in November 2024 demonstrate the added value of such collaboration and could serve as a strong foundation for future efforts.

2. Recommendations

A. Centring People

- When considering ideas, feasibility or viability of using AI for transitional justice, victims' needs and wishes should be the primary factor to be taken into account.
- AI technologies should never be tested on vulnerable populations. The do no harm principle and informed consent are minimum measures needed in thinking about whether or not AI is suitable or should be used for a given purpose.
- Tech literacy in general, but AI literacy in particular, is necessary to ensure that victims or victims' groups have sufficient knowledge about technologies and know how to seek explanations and accountability should it result in harmful impacts. This includes understanding the limits of what AI can do and the impact of AI in the information ecosystem of the post-conflict or democratic transition context.
- Technological solutions should never be parachuted in without an understanding of local socio-economic and socio-political contexts, including through consultations with local victims' groups.
- At the same time, due to the sociotechnical nature of AI and its potential to be (mis)used for political ends, especially in post-conflict settings, a wider mapping of potentially affected persons and the societal impacts of AI should be undertaken.

B. AI and transitional justice in practice

- Organisations working in transitional justice could consider partnering up with organisations or joining coalitions of actors already deploying "AI for Good". See Annex 1 for a list of organisations already working in this space.
- Running a workshop or series of workshops in a practical creative lab setting to explore the possibilities of using AI in the transitional justice context.
- In identifying potentially new areas of focus or collaboration, a human rights-based approach to AI for transitional justice that foregrounds victim-centred approaches should be taken.
- Data protection policies should be in place or updated as necessary prior to and in conjunction with the use of AI in transitional justice contexts.
- Taking stock of and examine existing operational gaps (e.g. victim participation, gender gaps) as these could be exacerbated by the use and deployment of AI.
- If AI solutions are procured externally or in collaboration with others, there is still a need to raise awareness among staff about the capacities and limits of the AI system.

- In raising awareness, organisation should start with an in-depth series of studies exploring the potential use, use-cases and drawbacks of AI in different transitional justice mechanisms and processes. In doing so, a lifecycle approach towards AI (i.e. ideation, design, testing, implementation, follow-up and de-commissioning) should be undertaken.

- Building tech and AI literacy among victims, victims' groups and advocacy organisations.

- Using AI to highlight gaps or challenges as well as successes of transitional justice mechanisms could be explored.

C. Policy recommendations

- Developing a standard-setting document on the use of AI for transitional justice. This could take the form of a code of conduct, statement of principles, ethical principles or similar. Such a document could be developed together or in consultation with other organisations, victim groups, policymakers, the UN or other relevant stakeholders.

- Conducting a study exploring accountability frameworks or mechanisms clarifying the roles and responsibilities of private actors in conflict or periods of mass violence.

- Conducting human rights impact assessments before deploying AI, including AI used ostensibly for security and safety reasons (e.g. the use AI facial recognition in public spaces).

- Having a multi-pronged approach in ensuring the resilience of the information sphere post-conflict or during the transition period – including strengthening local journalism and promoting AI literacy alongside encouraging the use of content and provenance tracing methods such as C2PA.

- Supporting efforts to establish archives for user-generated content from social media or other platforms that can assist in establishing accountability for serious human rights and humanitarian law violations.

- Taking a victim-centred and a human rights-based approach in considering, designing and deploying AI for transitional justice. This can include eventually deciding not to deploy AI where human rights impacts are judged to be disproportionate.

1. INTRODUCTION

Artificial intelligence (AI) is now ubiquitous. AI technologies can help reform healthcare, education, transportation, scientific research and public services for the better - many of which contribute towards the enjoyment of human rights. At the same time, AI can also negatively impact human rights, including breaches of privacy, anti-discrimination measures, and freedom of expression. AI also increasingly features in violent conflict, both as part of military operations but also in the context of monitoring and addressing human rights and humanitarian law violations during conflict.

This paper examines how artificial intelligence could – and indeed already does - play a role in transitional justice. Transitional justice can be broadly understood as "the full range of processes and mechanisms associated with a society's attempts to come to terms with a legacy of large-scale past abuses, in order to ensure accountability, serve justice and achieve reconciliation."[1] While historically focused on the pursuit of justice through prosecutions, current transitional justice mechanisms also serve the wider goals of nation-building, rebuilding political identity and peace-building in general.[2] As such, transitional justice mechanisms can take a judicial or non-judicial form, including criminal trials, amnesty provisions, truth commissions, lustration, reparations or a combination thereof. While processes and mechanisms chosen to address past injustices and violence can vary depending on context,[3] it is part of a comprehensive policy that serves the purpose of recognition, re-establishing trust, strengthening the rule of law and promoting democracy and reconciliation.

This paper offers a broad overview of the connections between AI and transitional justice in light of many identified and emerging concerns in relation to how AI can impact the practice and field of transitional justice and its victim-centred approach. Disinformation, bias, privacy, lack of transparency and accountability, alongside a narrow focus on cost-cutting and efficiency afforded through AI, can negatively impact victims and transitional justice mechanisms. That said, AI can also play a positive role in transitional justice practices in terms of documentation and archiving and be an enabling tool for victim agency and empowerment. The paper aims to map and frame these connections. It will not, however, go into each element in great detail but instead lays out the issues and considerations that should feature in the conversation.

The overarching guiding perspective of this paper is the victim-centred approach. The victim-centred approach necessitates that actions, policies and practices are guided and informed by the needs and choices of victims.[4] It requires asking whether a given action, policy or practice is meaningful for victims in terms of making a tangible difference for them. The victim-centred

1. "Transitional Justice: A Strategic Tool for People, Prevention and Peace." Office of the United Nations High Commissioner for Human Rights (OHCHR) Secretary-General guidance note, October 2023, https://www.ohchr.org/en/documents/tools-and-resources/guidance-note-secretary-general-transitional-justice-strategic-tool.
2. Ruti G. Teitel, "Transitional Justice Genealogy (Symposium: Human Rights in Transition)," *Harvard Human Rights Journal* 16 (2003).
3. Pablo De Greiff, "Theorizing Transitional Justice," *Nomos* 51 (2012): 31–77
4. OHCHR Secretary-General guidance note, "Transitional Justice: A Strategic Tool for People, Prevention and Peace."

approach is thus central in this mapping exercise. It asks the question – what does the adoption (or otherwise) of AI in the transitional justice context mean for victims? The paper also invites further deliberation on victim-centred approaches, AI and transitional justice through a series of open-ended guiding questions in Annex 2.

The paper starts out with an overview of the definition of artificial intelligence. This is necessary as the definition of AI itself is contested and lacks clarity. It then examines the characteristics and features commonly associated with the use of artificial intelligence, noting that a sociotechnical approach towards AI is necessary when assessing its benefits and drawbacks. Having provided clarity on what AI is and can do, the paper then looks into how AI relates to transitional justice, including by examining the relationship between AI and the values and normative aims of transitional justice.

Section 2 maps the ethical, human rights and societal impacts of AI. It also examines examples of how AI has been or is currently being used in transitional justice efforts as well as within adjacent fields, such as human rights and humanitarian law. Widening the aperture beyond transitional justice is necessary as AI is still new within the transitional justice context, and there are many benefits to reviewing the lessons learned about how AI has been used in these other fields. However, while we can learn from human rights and humanitarian contexts, transitions are exceptional episodes in a society's history. Large-scale atrocities that occur during violent conflict or authoritarian rule cannot be addressed with conventional judicial means but instead call for extraordinary measures. Thus, due to the distinct normative aims of transitional justice, its approach to AI should be informed by these aims even as transitional justice can draw from lessons learned from the fields of human rights and humanitarian law in relation to AI. This section brings these elements together, looking at the potential ethical and human rights considerations the use of AI may pose in the transitional justice context. Lastly, it also maps the existing and emerging governance mechanisms for AI, tracking legislative developments in the multilateral and global fora, regionally and domestically.

Section 3 sets out a framework for a victim-centred approach on AI for transitional justice. In thinking about whether, why, how and when AI can be used in the transitional justice context, we offer a three-pronged analysis, examining it from the perspective of (i) the individual (victim); (ii) processes and procedures; and, (iii) the normative aim(s) of transitional justice. With this, we ask whether AI can make a difference for victim participation and empowerment, how AI can assist in the implementation of transitional justice mechanisms and how AI contributes towards the aims of transitional justice.

Finally, in Section 4, we provide a list of proposals - addressed to different stakeholders such as victims, victims' groups and other transitional justice grassroots organisations - for programme planning in general and also for policymakers, donors and civil society groups. However, we envisage this paper as the beginning of an ongoing conversation which serves to inform and enhance the capacities of these different stakeholders in navigating this domain, rather than to provide conclusive recommendations.

a. Defining artificial intelligence

Despite how frequently the term AI appears in media, policy-making and business circles, there is still much confusion as to what AI actually is. The term AI was first coined in 1956, and the initial idea was to create machines able to mimic human intelligence. However, human intelligence is itself a contested idea – variously linked to ideas of consciousness and intelligence as a social endeavour (how we relate to others). Stuart J. Russell, a key figure in the field of machine learning and AI, instead defines AI according to the capacity of machines to act as rational agents – taking actions based upon receiving input from its surroundings.[5]

Thus, instead of trying to clarify the concept of intelligence based upon the subjective quality of human intelligence, it may be more useful to define AI according to what it does. This is the approach taken by the European Union's Artificial Intelligence Act, the world's first comprehensive legislation regulating AI, which came into force in 2024.[6] Instead of defining AI as an abstract concept, the EU legislation defines what an AI system does as a system that operates with "varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers from the input it receives how to generate outputs such as predictions, content, recommendations, or decisions".[7] Three main features distinguish AI from other data-driven or big data systems, namely: its autonomy, its capacity to infer, and output generation. Put plainly, this means that an AI system can take actions without human intervention, can supply insights based on data and take various actions based on data. While the term AI is still defined and used differently in various jurisdictions, the EU approach is helpful as it helps to ground the discussion at a practical, rather than philosophical, level.

Seen this way, AI consists of various sets of technologies and techniques. This can be best illustrated by way of examples. Facial recognition technologies using computer vision is a type of AI system to detect facial biometrics, and has been increasingly deployed in different jurisdictions on grounds of national security and to assist with law enforcement. Emotion recognition technologies have been deployed in the context of borders and migration but also in classrooms and workplaces to monitor the emotional responses of migrants, students and employees. Public services such as social security, healthcare and unemployment insurance also increasingly rely on AI-driven decision-making or recommendations. This is where AI systems that have trained on various past datasets can then make recommendations and decisions on access to healthcare, social or unemployment benefits. AI has also been used to further the cause of human rights and humanitarian law, including by using image recognition technology to monitor the movement of weapons or by providing data analysis on trends and early warnings of potential conflicts.

5. Stuart J Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach.* (Third edition, Pearson, 2016).

6. Regulation (EU) 2024/1689 of the European Parliament and of the Council of June 13, 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (hereafter the EU Artificial Intelligence Act).

7. Article 3 (1) EU Artificial Intelligence Act.

While AI systems can consist of many different approaches, the term AI these days is mainly synonymous with machine learning. This consists of computational systems that learn from data in order to improve its performance. There are three main types of machine learning – supervised, unsupervised and reinforcement learning. Supervised learning entails learning based on data that has been labelled. Unsupervised learning clusters and finds patterns in unlabelled data. Reinforcement learning is an approach where the AI models improve based upon its interaction with the physical or computational environment and where desired actions are rewarded in order to maximise optimal outcomes. Machine learning is in turn operationalised through algorithms, which are a set of instructions, akin to a recipe, for the AI model to carry out its tasks.

The development and deployment of generative AI, namely AI models that can generate text, images and audio, signals the fast pace of change in the field. This development also introduces new vectors of threats such as making it difficult to distinguish truth or falsity and generally distorting the information environment. This is a pertinent concern for the transitional justice context, where establishing truth and building trust is fundamental. Increasingly powerful systems can also pose threats on a global scale, including its possible misuse for cybersecurity exploitation, bio and chemical weapons or manipulation. The rapid development of AI has brought forth fears that humans might lose control over this technology[8] and that it might widen systemic inequalities and introduce new harms.

b. The features and affordances of AI[9]

AI can be considered as **general-purpose technology** such as electricity or the steam engine. This means it can be used in different domains and for a variety of purposes. AI has also been termed as a **dual-use technology**, namely it can be used for either benevolent or malevolent ends (e.g. in war). As Lindsay Freeman, director of Technology, Law and Policy at the Human Rights Centre at UC Berkeley School of Law, puts it, AI can be used as weapons of war or tools of justice.[10] AI is neither inherently good nor bad and its use has to be judged according to the context. Correspondingly, in order to assess how, if or when AI should be designed or used, a **sociotechnical approach** is needed. This means that the design and implementation of technologies is not only a technical concern but is intimately linked to the social context in which technologies are deployed. Thus, in designing an algorithm to determine fair allocation, the process of designing such a system needs to take into account existing biases and forms of discrimination that are already present in society and ask how an algorithm (along with other measures) can address this inequality. In turn, the fast development of AI and the enormous potential which it can open up for economic and military competitiveness has been seen in nations such as China, so much so it has been branded

8. "Pause Giant AI Experiments: An Open Letter." Future of Life Institute, accessed April 5, 2023, https://futureoflife.org/open-letter/pause-giant-ai-experiments.

9. In this context, the term "affordances" refers to how a technology opens up various conditions of possibilities, including going beyond its initially envisioned use cases. Hutchby argues that affordances can be relational (wherein a particular tool can have differentiated impacts upon different groups) and interpretable (where usage of a given object goes beyond what it was initially designed for). Ian Hutchby, "Technologies, Texts and Affordances," *Sociology* 35, no. 2 (May 2001): 441–56, https://doi.org/10.1177/S0038038501000219.

10. Lindsay Freeman, "Weapons of War, Tools of Justice: Using Artificial Intelligence to Investigate International Crimes," *Journal of International Criminal Justice* 19, no. 1 (March 2021): 35–53, https://doi.org/10.1093/jicj/mqab013. Artificial intelligence is shaping warfare in the Digital Age. Fuelled by data rather than gasoline, artificial intelligence (AI).

a "leapfrog" technology.[11] The many national AI policies in place consider that AI offers a **distinct advantage** of speed, scale and efficiency. Some policymakers go as far as saying that AI is not just a tool, but in effect unlocks a whole new world.[12]

Collage with mirrors reflecting diverse human figures, symbolising AI data's human origin and the human-in-the-loop concept. Anne Fehres and Luke Conroy & AI4Media / Better Images of AI / Data is a Mirror of Us / CC-BY 4.0

AI is nothing without data. And data is plentiful given the growth of the internet and our existence online, which has enabled AI to take-off in earnest. However, individuals might not always be aware of when their data is being used and how it is being used to power AI-driven decisions or recommendations. The generative AI boom, seen through the popularity of ChatGPT for example, requires enormous amounts of data to train such systems, posing questions over the legality of widescale data scraping as well as the possible environmental footprint of training and running generative AI models.

As outlined earlier, AI is different from prior technologies because of the possibility of autonomous action. AI can make decisions, take actions and draw insights without human intervention. In addition, we might not necessarily know how an AI system comes to its eventual output or decision. AI systems, especially deep-learning neural networks in machine learning systems, can be complex and its workings opaque, both to the end user as well as the programmer. Furthermore, when deployed, AI can be used in ways that were not intended when designed, including by malicious actors for malevolent purposes.

11. "Full Translation: China's 'New Generation Artificial Intelligence Development Plan' (2017)," *DigiChina* blog, Stanford University, accessed November 13, 2024, https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/

12. European Commission Executive Vice-President Margrethe Vestager, speech on technology and politics at the Institute for Advanced Study, April 29, 2024, https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_24_1927

c. Transitional justice and AI: Tracking values and normative principles

Having defined and discussed the key features of AI, we can now examine how AI relates to the underlying values and normative principles of transitional justice. Transitional justice processes are attempts by society to come to terms with prior incidents of mass violence, systemic human rights abuses or conflict with the goal of seeking recognition, trust-building, strengthening the rule of law, and reconciliation. Various transitional justice mechanisms are available, depending on context, to seek accountability and justice for past harms.

The key element in determining the suitable measures and ends to be pursued through transitional justice is by centring and addressing the needs of victims. AI, through its ability to manage, analyse and distil insights from large datasets and find correlations between various data points, is able to assist in visualising the needs of victims expressed through surveys or interviews conducted in the field. Similarly, AI can also help analyse information in databases and archives which may facilitate justice-seeking mechanisms such as prosecutions, but also provide a clearer picture, pertinent to truth-seeking and reconciliation.

In turn, using AI to gather data from various online sources, for example through user-generated content on YouTube and other social media platforms, can help to provide a detailed picture of how a conflict unfolded, which people were targeted, and who the perpetrators might be. As conflict increasingly unfolds on social media, triangulating these forms of user-generated data from different platforms can provide a more comprehensive picture of possible accountability and help to promote justice and peace.

Content that we see online is increasingly selected for us via the use of algorithms and this curation can ensure that content reaches the most relevant audiences. Content curation and promotion can help transitional justice mechanisms to gain increased visibility for outreach, reflecting in increased participation of possible stakeholders and awareness-raising of specific events, trials or other transitional justice measures.[13]

AI systems have also been trialled in novel ways, such as through the use of AI for psychological and mental health advice and the use of AI in courtrooms to assess evidence or even determine cases.[14] However, some of these new uses pose challenges to the rule of law, raise ethical questions and require assessment in terms of impacts on human rights and human dignity before its rollout. The high-level takeaway at this point is that there is no inherent incompatibility between AI and the values and normative aims of transitional justice. However, a victim-centred and human rights-driven approach is needed in determining how and where AI can contribute to transitional justice.

13. Jasmin Haunschild, Laura Guntrum, Sofia Cerrillo, Franziska Bujara, and Christian Reuter, "Towards a Digitally Mediated Transitional Justice Process? An Analysis of Colombian Transitional Justice Organisations' Posting Behaviour on Facebook," *Peace and Conflict Studies* 30, no. 2 (May 2024), https://nsuworks.nova.edu/pcs/vol30/iss2/4

14. Tara Vasdani, "Robot Justice: China's Use of Internet Courts," *LexisNexis Canada*, accessed on September 6, 2024, https://www.lexisnexis.ca/en-ca/ihc/2020-02/robot-justice-chinas-use-of-internet-courts.page.

2. MAPPING THE FIELD

a. The human rights, ethical and societal impacts of AI

i. Human rights

The deployment of AI in various sectors and domains has, inevitably, led to concerns over negative impacts on human rights. One of the main issues is in relation to **AI bias**. Research has shown that facial recognition, decision-making and recommendation systems and natural language processing systems, all of which are powered by AI, often disadvantage racial, ethnic and linguistic minorities.[15] This is in part due to the fact that data used to train AI systems is not representative but also because societies are still biased and this can be reflected in the data that is gathered. This can mean inequalities that already exist in society can be amplified through the scaled use of AI, potentially violating the **right to equality and non-discrimination**.[16]

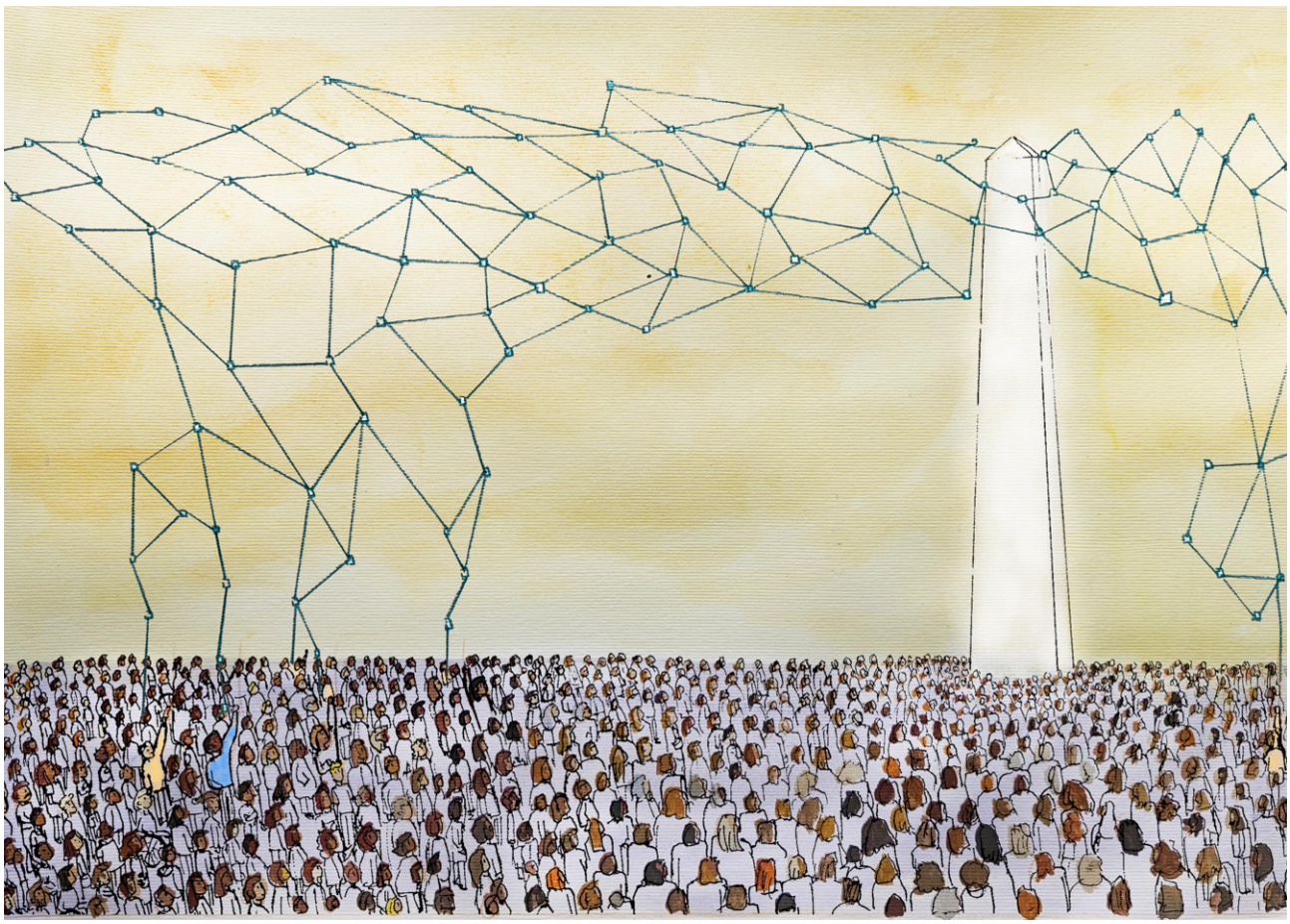
In turn, as AI is trained on data, this entails increasingly large portions of our online - and even offline - interactions being monitored and datafied. This potentially imperils the **right to privacy** as people might not be clear as to when they are being monitored, how information is used and what the impacts may be. For example, when data is used in the future or in contexts people did not initially consent to. In the context of unstable post-conflict scenarios, this is an even more pertinent concern as data can end up in the wrong hands and be misused. Facial recognition systems that are increasingly rolled out across cities in the world, ostensibly for reasons of security and terrorism prevention, can negatively impact on privacy as people are being watched for non-transparent reasons and can result in chilling effects on the **freedom of expression and association.**

The increasingly important role of social media, including in conflict situations, can also negatively impact **freedom of expression, information and freedom of thought**. Where social media algorithms censor and remove content relevant to establishing truth and accountability for crimes or acts of violence, this can hamper attempts to seek justice for past wrongs.[17]

15. Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 1-15, https://www.media.mit.edu/publications/gender-shades-intersectional-accuracy-disparities-in-commercial-gender-classification/.
Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, "Machine Bias" *ProPublica* (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

16. Angwin et al, "Machine Bias".

17. OHCHR Independent international fact-finding mission report on Myanmar, 12 September 2018, para 74, https://www.ohchr.org/sites/default/files/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf



A neural network emerges from the top of an ivory tower, above a crowd of people's heads. Some are reaching up to try and take control and pull the net down to them. Watercolour illustration. Jamillah Knowles & We and AI / Better Images of AI / People and Ivory Tower AI 2 / CC-BY 4.0

However, despite existing human rights and AI scholarship that focus on these impacts, AI can **impact a wide variety of human rights** as human rights are also **interrelated, interdependent and interconnected.** Biased or inaccurate AI decision-making within public services can mean that the **right to education, healthcare and work** can be negatively impacted. In turn, when it is difficult for individuals to gain insight into how an AI system made a decision or functions, accountability for wrongs or harms that occur cannot be achieved. This can jeopardise the **right to a meaningful remedy** under human rights law.

While the majority of these concerns have been raised and researched in the Global North context, this paper argues that these concerns are also relevant in the Global South context. As we shall see later, such human rights impacts can be exacerbated in the transitional justice context in ways that can negatively impact victims.

ii. Ethical impacts

At the same time, AI also raises ethical issues. As mentioned, data powers AI. Big tech companies possess the datapoints of billions of individuals who might not understand how their data is used or how it can affect them. In other words, **data is power**. LSE Professor of Media, Communications and Social Theory Nick Couldry and SUNY Oswego Professor of Communication Studies Ulises

Ali Mejias argue that this is a new form of colonialism – data colonialism.[18] Increasingly, this takes the form of having the power to shape narratives and impact democracies. The interests of large private companies and that of peace and democracy do not necessarily align due to commercial incentives – as evidenced through the monetisation of the spread of misinformation and disinformation across social media platforms.

However, this concern is not limited to large private companies. Nation states, whether deploying AI in the context of conflict or within domestic public services, have granular knowledge of individuals through their data. An AI system can use individuals' data and that of others to make inferences on whether or not people are likely to need healthcare, remain unemployed or have committed tax fraud. The lack of transparency both in terms of what data others hold and how that data is used to reach decisions can mean that the **exercise of power by the state lacks transparency and accountability.** This is especially the case where private technologies are used or where private actors partner with the public sector because algorithms that are used to assist in or reach decisions can be protected as **commercial trade secrets**. This raises ethical issues on how much influence the private sector can have on the public sector as the latter is driven by public values such as access to goods and services, solidarity and community rather than solely based upon efficiency and for-profit motives.

Similarly, **in the transitional justice and human rights context, power over data must be exercised with the utmost care and consideration for victims and rights holders.** Data leaks, lack of cybersecurity and data protection measures can mean the difference between life and death. The example of the Kivu Security Tracker database - a "data-centric crisis map" of atrocities in eastern Congo that publicly exposed the data of 8,000 people, including victims - should be a clarion call to take power over data seriously, even if one is using technology for good. These measures should be in place even in the absence of data protection legislation in the country of operations. This reminder from Adrien Ogée, the chief operations officer at the CyberPeace Institute, a Geneva-based cybersecurity institute, is both timely and relevant:

> *'If you're an NGO working in conflict zones with high-risk individuals and you're not managing their data right, you're putting the very people that you are trying to protect at risk of death.'*[19]

iii. Human dignity and humanity

AI raises not only human rights and ethical considerations. Some have raised concerns that AI systems that could eventually become more intelligent than human beings could pose an **existential threat** to humanity. Humanity could potentially lose control to a more intelligent entity. A fundamental question to ask from the perspective of human dignity is whether or not these

18. Nick Couldry and Ulises Ali Mejias, *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism* (Stanford University Press, 2019).

19. This quote from Adrien Ogée, the chief operations officer at CyberPeace Institute, is cited in Robert Flummerfelt and Nick Turse, "Online Atrocity Database Exposed Thousands of Vulnerable People in Congo," *The Intercept*, November 17, 2023, https://theintercept.com/2023/11/17/congo-hrw-nyu-security-data.

advanced AI initiatives should even be pursued. There is disagreement on how long it will take to reach this stage and whether or not superintelligence could even be achieved. Nonetheless, even if contentious, the pursuit of Artificial General Intelligence (AGI), where an AI is more capable than humans across different domains, has gained a lot of attention from policymakers due to the potential (existential) risks it could entail. Even as large language models and other generative AI systems such as image generators are more widely adopted today, the ultimate ambition of companies such as OpenAI is to pursue AGI.

However, even without AGI, existing generative AI such as large language models can already pose a risk to humanity, including through potential misuse by nefarious actors to carry out cyberattacks, or biochemical or genetically-engineered warfare – all of which can bring serious risks to human safety and security. Experimenting such uncertain (and risky) technological paths raises questions of human dignity as it treats adopters of such technology as a means to an end, especially where societies and individuals would bear the risk of harm.

An illustration showing the silhouettes of four people, two of whom have dogs on leads. They all cast shadows, and vary between realistic representations and those formed by representations of algorithms, data points or networks. The people and their data become indistinguishable from each other. Jamillah Knowles / Better Images of AI / Data People / CC-BY 4.0

Human dignity can also be potentially impacted in other ways. Where reliance upon data is increasingly the primary way for states and governments to gain a better insight into its population as well as social phenomena, this can reduce the lived experiences of persons - with its multifaceted cultural, social and economic influence that inform our existence - to a mere datapoint. While data has certainly helped in gaining a better understanding of society, for

example during the Covid-19 pandemic, taking a perspective that respects human dignity and requires us to take individual circumstances into account is vital, especially with vulnerable and marginalised populations.

Philip Alston, the former UN Special Rapporteur on Extreme Poverty and Human Rights, cautioned against "stumbling zombie like into a digital welfare dystopia".[20] He criticised the fact that social services are "increasingly driven by digital data and technologies that are used for diverse purposes, including to automate, predict, identify, surveil, detect, target and punish." Reliance on data as an indicator of truth against a neoliberal public management mindset can mean that we disregard individual extenuating circumstances and end up punishing recipients of social services who most need help. Respecting the lived experiences of people as a matter of human dignity means no less than having viable channels for populations to interact with the state – including by being able to engage with human service staff as a meaningful alternative. The ability to give an account of oneself respects the autonomy and dignity of persons.

iv. Democracy and the rule of law

The lack of transparency about how AI systems work and how we can potentially be affected also raises questions over **accountability deficits**. In the absence of sufficient knowledge of how AI was used and how decisions or recommendations were reached, we then lack the ability to challenge such decisions when something goes wrong. This goes against the rule of law and good governance principles where individuals should be able to understand how decisions about them are made and challenge unjust decisions. We have seen this resulting in harms such as large sums of debt being demanded from individuals by the state without the former knowing why, such as the "Robodebt" scheme, which was an automated government scheme by the Australian Government that wrongly demanded welfare recipients repay benefits. Many individuals received letters claiming they owed thousands of dollars in debt, based on an inaccurate algorithm. The policy impacted over half a million Australians.[21] Additionally, residents have been denied social security benefits based on algorithmic recommendations that lack transparency, as was the case in the Rotterdam welfare fraud algorithm that discriminated against residents based on gender and ethnicity. [22] Also, opaque algorithms have impacted access to education such as the case of a Dutch student who experienced issues because her university's surveillance software struggled to recognise her as a human being due to her skin colour.[23] The effects are even more pronounced for vulnerable and minority groups when AI systems are reliant on datasets that are not free of bias, as already noted in this paper.

20. "World Stumbling Zombie-like into a Digital Welfare Dystopia, Warns UN Human Rights Expert," OHCHR, accessed August 23, 2024, https://www.ohchr.org/en/press-releases/2019/10/world-stumbling-zombie-digital-welfare-dystopia-warns-un-human-rights-expert.

21. Dorothy Roberts and nia t. evans, "The 'Benevolent Terror' of the Child Welfare System," *Boston Review*, March 31, 2022, https://bostonreview.net/articles/the-benevolent-terror-of-the-child-welfare-system/.
Luke Henriques-Gomes, "Robodebt Class Action: Coalition Agrees to Pay $1.2bn to Settle Lawsuit," *The Guardian*, November 16, 2020, https://www.theguardian.com/australia-news/2020/nov/16/robodebt-class-action-coalition-agrees-to-pay-12bn-to-settle-lawsuit.

22. Eva Constantaras, Gabriel Geiger, Justin Casimir-Braun, Dhruv Mehrotra and Htet Aung, "Inside the Suspicion Machine," *Wired*, March 6, 2023, https://www.wired.com/story/welfare-state-algorithm.

23. de Zwart, Hans, "Dutch Institute for Human Rights: Use of Anti-Cheating Software can be Algorithmic Discrimination (i.e. Racist)," *Racism and Technology Centre blog*, December 24, 2022, https://racismandtechnology.center/2022/12/24/dutch-institute-for-human-rights-use-of-anti-cheating-software-can-be-algorithmic-discrimination-i-e-racist.

In addition to the necessity of having transparency over the decision-making processes of AI systems, it has also been recommended that there should be a human-in-the-loop, meaning that someone is responsible for checking recommendations rendered by the AI system. However, the human tendency to (blindly) trust ostensibly objective and neutral findings of the algorithmic system might mean that having a human-in-the-loop is not a foolproof solution.

Additionally, misinformation and disinformation that increasingly pervade the information environment also threaten democracy. Where the truth or falsity of information is in perpetual doubt or where one is unable or unwilling to fact-check the veracity of content encountered, this can reduce civic trust and social solidarity. It is especially pertinent in the context of elections where misinformation and disinformation can be weaponised for political ends, including by external parties for malicious purposes.

Maria Ressa, an investigative journalist awarded the 2021 Nobel Prize for Peace, argued that we cannot have integrity of elections if we do not have integrity of facts.[24] While various measures are under way to enable us to check the veracity of the content and information sources we encounter, research on human psychology also reveals that these measures alone might not be sufficient. It has been shown that the tendency to believe mis- and disinformation is affected by whether the content comes from an in-group (e.g. friends), existing biases (e.g. political partisanship) and whether the content is repeatedly encountered.[25] Thus, technical measures, including watermarking and provenance-tracing might only partially address the tendency to believe "fake news". Further measures such as labelling, fact-checking, AI literacy and awareness raising and stronger investment in local journalism might be necessary complementary measures.

Finally, misinformation and disinformation can also contribute towards or exacerbate conflict, increase attacks and risks towards vulnerable populations and lead to increased hate speech inciting violence or the targeting of minority groups.

b. Current AI use in transitional justice and related fields

As AI use permeates across various domains and sectors in society, the same can be said for the field of transitional justice.[26] AI is primarily used in the transitional justice and justice-seeking context in general to process, organise, map and analyse various types of data (e.g. text, images and videos) from various sources (e.g. user-generated content from social media, YouTube and digitalised archives) to discern types of objects (e.g. types of weapons), places, persons and other content in order to help analyse the patterns of human rights violations or systemic violence.

24. Maria Ressa, "We're All Being Manipulated the Same Way," *The Atlantic*, 7 April 2022, https://www.theatlantic.com/ideas/archive/2022/04/maria-ressa-disinformation-manipulation/629483/.

25. Catherine Beauvais, "Fake News: Why Do We Believe It?" *Joint Bone Spine* 89, no. 4 (July 2022), https://doi.org/10.1016/j.jbspin.2022.105371. Marshall Shepherd, "Repeating Misinformation Doesn't Make It True, But Does Make It More Likely To Be Believed," *Forbes*, August 17, 2020, https://www.forbes.com/sites/marshallshepherd/2020/08/17/why-repeating-false-science-information-doesnt-make-it-true/. "What Psychological Factors Make People Susceptible to Believe and Act on Misinformation?" American Psychological Association, November 29, 2023, updated on March 1, 2024, https://www.apa.org/topics/journalism-facts/misinformation-belief-action.

26. Daniela Gavshon, "How New Technology Can Help Advocates Pursue Transitional Justice," *Oxford University Press*, July 1, 2019, https://blog.oup.com/2019/07/how-new-technology-help-advocates-pursue-transitional-justice.

Key accountability fora, such as the International Criminal Court (ICC), the International, Impartial and Independent UN monitoring mechanism in Syria (IIIM) and the civil society-backed Syrian Archives are just a few of the organisations using artificial intelligence for the abovementioned purposes. At the ICC, Project Harmony seeks to modernise the evidence management platform used by the Office of the Prosecutor by harnessing modern technology to enable rapid pattern identification, automated translation, facial recognition, targeted searches of source material and automated transcription and other tasks. Even with highly documented conflicts such as in Syria, AI complements human resources by reducing the time needed for manual checks and verification and also complements existing technologies by sifting through all data sources and removing duplicates of large bodies of documentary and evidentiary sources. AI can also help to show a pattern of widespread and systemic violence from available datasets. In other words, AI can help to make sense of the data at hand, doing so in a more efficient, less labour intensive and more cost-effective manner.

The IIIM, established by the UN to investigate and prosecute serious crimes in Syria since 2011, is one of the first international criminal investigation organisations to use computer vision. AI-driven computer vision helps in characterising and grouping images based on similar characteristics but it can also recognise specific signs and other markers (e.g. signature, insignia etc) pertinent towards building a legal case. The task of finding and clustering relevant images would typically involve painstaking and time-consuming human labour. Used in this manner, computer vision can also expand the reliability of evidentiary bases beyond text-based sources.[27] In the case of the IIIM, "hundreds of thousands of Arabic-language documents—much of them low-quality images—can be extracted and the chain of command and criminal accountability can be established through finding patterns such as official stamps, letterheads or signatures".[28] AI is also being considered for use in machine translation from the Burmese language to English as part of the work of the Independent Investigative Mechanism for Myanmar (IIMM).

Thus, AI is used to analyse and synthesise patterns from different sources of information, such as geospatial imagery, documents, videos, forensic evidence and social media posts to triangulate[29] and provide a better understanding of how a conflict unfolded and to meaningfully organise evidence for future justice-seeking and accountability measures.

27. Elena Radeva, "The Potential for Computer Vision to Advance Accountability in the Syrian Crisis," *Journal of International Criminal Justice* 19, no. 1 (March 2021): 131–46, https://doi.org/10.1093/jicj/mqab015.

28. Raja Abdulrahim, "AI Emerges as Crucial Tool for Groups Seeking Justice for Syria War Crimes," *Wall Street Journal*, February 13, 2021, https://www.wsj.com/articles/ai-emerges-as-crucial-tool-for-groups-seeking-justice-for-syria-war-crimes-11613228401.

29. In this context, **"triangulate"** refers to the process of cross-referencing and corroborating information from multiple sources to ensure its accuracy and reliability.

A laptopogram populated by scattered, monochromatic portraits and grouped according to similarity. The groupings vary in size, ranging from single to overlapping collections of several faces. The facial expressions are neutral, representing a mix of ages and genders. Philipp Schmitt & AT&T Laboratories Cambridge / Better Images of AI / Data flock (faces) / CC-BY 4.0

AI can also assist in the search for missing persons through the use of facial and object recognition, contributing towards strengthening accountability, truth-seeking and reconciliation. As AI enables the analysis of various types of data at scale, it can thus also assist in the reparations process – enabling harms to be properly assessed and compensation to be allocated and accounted for. By the same token, the ability of AI to sift through large datasets and discern key patterns and emerging trends can also mean that it can identify risk factors and act as an early warning mechanism, contributing towards non-recurrence as well as the goal of sustainable peace pursued by transitional justice.

However, while this is the most direct way AI currently plays a role in justice and truth-seeking, the ubiquity of AI, including when embedded or used in tandem with other tools, services and mechanisms, can mean that its use might not be immediately obvious but nonetheless beneficial for transitional justice. For example, social media platforms such as Facebook have been used by the Colombian Truth Commission, set up as part of the 2016 Peace Accord after almost five decades of human rights violation and armed conflict, in order to engage participation and solicit feedback from the public and affected communities.[30] These platforms are powered by algorithms that enable like-minded communities and those sharing common interests to connect. In this way, AI can serve to raise awareness and seek to further expand the participation of victims, affected communities, civil society and even the international community.

30. Jasmin Haunschild and others, "Towards a Digitally Mediated Transitional Justice Process? An Analysis of Colombian Transitional Justice Organisations' Posting Behaviour on Facebook," *Peace and Conflict Studies* 30, no. 2 (May 2024), https://nsuworks.nova.edu/pcs/vol30/iss2/4.

AI can also be used to establish truth and to engage the public in conversations on past atrocities, including by facilitating or complementing digital storytelling. Upon finalising its work, the Colombian Truth Commission presented it work and engaged with the public using an interactive platform consisting of installations, videos, audio recordings, and data visualisations and an installation powered by artificial intelligence. The latter answered questions from the public about the Colombian conflict based on the commission's Final Report.[31] As the Colombian archive was almost 100% digital, this enabled the voices of the victims of the conflict to be preserved, presented and understood and supports a community-centred approach to transitional justice. In terms of enhancing digitalised archives, AI-enabled searching and clustering of how information is presented can help to better present victim experiences and enable a more comprehensive understanding of past conflicts.

AI can also play a role in ensuring the effectiveness of transitional justice mechanisms and the pursuit of justice and peace. The Transitional Justice Evaluation Tool (TJET) is an initiative by Harvard University to comprehensively document transitional justice measures around the world.[32] The open-source platform can enable researchers to examine and analyse not only singular transitional justice efforts (within a country or mechanism) but also compare across mechanisms and jurisdictions, hence contributing towards evidence-based research on the efficacy of transitional justice measures.

Besides these examples, novel uses of AI are being proposed and considered. One area is the use of virtual reality (VR) to better understand the conditions in conflict by simulating the perspective of someone in a conflict zone. The UN has adopted VR technology in reaching out to policymakers, enabling them to experience life in a conflict zone "through the eyes of those who experienced it."[33] The use of VR could potentially generate empathy towards causes but also poses privacy and ethical issues.

In short, AI currently plays a role – and could potentially play a greater role - in enabling a victim-centred approach towards accountability, justice and peace-building. However, enthusiasm in embracing AI is not typically accompanied by human rights, ethical or societal assessments in relation to its adverse impacts. Sections c and d below will examine some potential issues the use of AI could raise in these adjacent contexts and critically evaluate if AI governance can be informed by transitional justice principles.

31. "At an Interactive Exhibit, Colombians Reflect on their Country's Painful Past and New Possibilities for its Future," International Center for Transitional Justice (ICTJ), May 4, 2024, https://www.ictj.org/latest-news/interactive-exhibit-colombians-reflect-their-country%E2%80%99s-painful-past-and-new.

32. "Transitional Justice Evaluation Tools homepage," Transitional Justice Evaluation Tools, accessed August 26, 2024, https://transitionaljusticedata.org/en.

33. Kate E. Bloch, "Virtual Reality: Prospective Catalyst for Restorative Justice," *American Criminal Law Review* 58, no. 2 (2021): 285, https://www.law.georgetown.edu/american-criminal-law-review/in-print/volume-58-number-2-spring-2021/virtual-reality-prospective-catalyst-for-restorative-justice/.

c. AI in the context of human rights and humanitarian law

Similar to the use of AI in the transitional justice and justice-seeking contexts, AI is primarily used in the human rights and humanitarian law contexts for the purposes of documenting, monitoring, mapping and analysing human rights and humanitarian law violations from a variety of data sources – text, audio, images and video. The transitional justice context overlaps with human rights and humanitarian law violations and thus, the means of accountability are also applicable here.

This overlap is also seen in the ways AI has been used in the human rights and humanitarian law contexts. We have seen AI being used to monitor and assess satellite imagery and thermal data, predict the possible scale and rate of displacement from conflict or natural disasters, and the use of computer vision to document, map and analyse human rights and humanitarian law violations such as the use of illegal weapons, invasion and occupation and unlawful detention.[34] AI tools can also both build upon and be used to strengthen existing human rights and humanitarian law monitoring datasets,[35] contributing towards a more enabling environment for human rights protection and promotion. AI can also protect human rights defenders, whether it be through the use of encrypted messaging apps or through the use of generative AI to convey human rights messaging without disclosing the identity of actual persons.

These and other efforts are part of the growing AI for Good movement, wherein AI is used to promote and protect human rights as well as to assist in reaching the UN Sustainable Development Goals.[36] For example, the use of emerging technologies, including AI, have been embraced by the UN Department of Political and Peacebuilding Affairs (DPPA), who have experimented with using VR for conflict prevention, mediation and peacebuilding. Some examples include the VR experience Iraq 360 and Sudan Now, both of which allowed viewers to experience being in the UN missions in both countries.[37] VR can also be a training simulator, helping mediators and peacekeeping forces to gain experience in managing challenging scenarios. As the means and methods of conflict change over time, the DPPA is currently looking into risks and benefits of using large language models in the peacebuilding context.[38] More generally, AI also has the potential to realise the UN SDG (Sustainable Development Goals) to transform the world for the better, including through poverty reduction, addressing environmental degradation and ensuring food security.[39] AI can be used to measure goal indicators through various datasets

34. Anne Dulka, "The Use of Artificial Intelligence in International Human Rights Law," *Stanford Technology Law Review* 26, no. 2 (August 2023), https://law.stanford.edu/publications/the-use-of-artificial-intelligence-in-international-human-rights-law/.

35. Examples include: Human Rights Data Analysis Group (HRDAG) https://hrdag.org/, Uppsala Conflict Data Program (UCPD) https://ucdp.uu.se/, Rule of Law in Armed Conflicts (RULAC) https://www.rulac.org/ and Armed Conflict and Location Event Data (ACLED) https://acleddata.com/.

36. Another analysis of how AI can be used for peace rather than a tool for conflict: Branka Panic and Paige Arthur, *AI for Peace* (First edition, CRC Press, Taylor & Francis Group 2024).

37. "Virtual Reality Bites: Using Technology to Bring Post-Conflict Situations to Life," United Nations Department of Political and Peacebuilding Affairs (UNDPPA) Politically Speaking website, August 11, 2022, https://dppa.medium.com/virtual-reality-bites-using-technology-to-bring-post-conflict-situations-to-life-bd5cb98ce3f6.

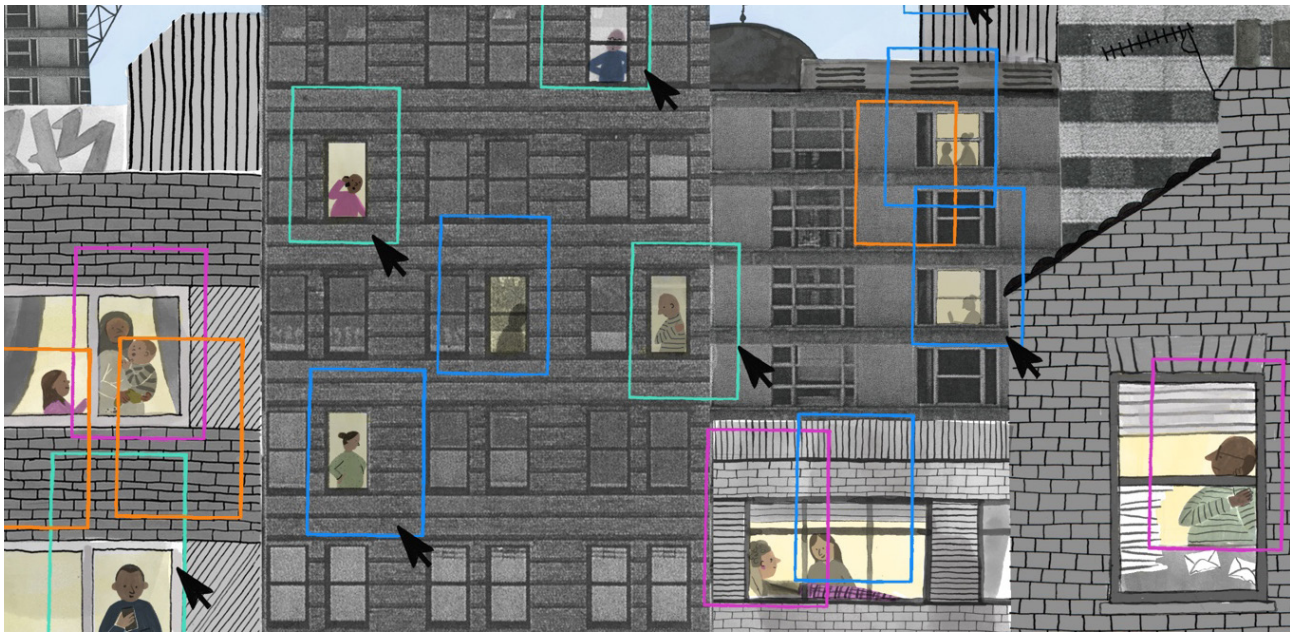
38. "Futuring Peace: Exploring the Power of Generative AI," UNDPPA, accessed August 26, 2024, https://www.futuringpeace.org.

39. Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini, "The Role of Artificial Intelligence in Achieving the Sustainable Development Goals," *Nature Communications* 11, no. 1 (January 13, 2020): 1–10, https://doi.org/10.1038/s41467-019-14108-y.

but also more directly contribute towards specific goals such as improving access to healthcare, optimising agricultural practices and reducing emissions.[40]

More generally, AI can be considered as an enabler of human rights. Generative AI and social media both empower freedom of expression; AI in education, including through personalised learning, can make the right to education more accessible and suitable; AI used in healthcare and scientific research can improve the right to an adequate standard of living and generally can improve human well-being. AI can also benefit specific rights holder groups. For example, the Be My Eyes application (app) helps to connect the sight-impaired with app users to enable the latter to narrate the surrounding environment. This app, now also powered by GPT-4,[41] empowers the sight-impaired to navigate the world around them in new ways.

Additionally, despite the existence of AI bias as examined earlier in this paper, AI can also be used to monitor and catch biases. For example, using data interactions from the social media platform X (formerly known as Twitter), researchers monitored how often women were subjected to harassment online, providing evidence-based research to underscore the scale of the problem.[42] The number of beneficial uses here is limitless: AI can enhance the right to health, address climate change challenges and empower children, persons with disabilities and other rightsholders.



Building blocks are overlayed with digital squares that highlight people living their day-to-day lives through windows. Some of the squares are accompanied by cursors. Emily Rand & LOTI / Better Images of AI / AI City / CC-BY 4.0

However, despite the promise of AI, the use of AI and other digital technologies can in itself raise human rights concerns. While research on AI in the transitional justice context remains

40. Dilek Fraisl, "The Potential of Artificial Intelligence for the SDGs and Official Statistics," *Paris21* Working Paper (April 2024): https://www.paris21.org/sites/default/files/related_documents/2024-04/the-potential-of-ai-for-the-sdgs-and-official-stats_working-paper_0.pdf.

41. "Be My Eyes" app, Open AI, https://openai.com/index/be-my-eyes/

42. Dulka, "The Use of Artificial Intelligence in International Human Rights Law."

nascent, there has been much more research carried out on the use of AI in the human rights and humanitarian law context, enabling a more nuanced picture of its drawbacks and benefits. The use of satellite imagery and thermal heat detection for monitoring spatial data and population movements have been said to pose privacy concerns. While there is a need to ensure that aid and other forms of assistance reach the designated beneficiaries, uncritical employment of facial recognition, iris scans and similar technologies has been criticised as a form of "surveillance humanitarianism".[43]

Similarly, the use of data in human rights and conflict monitoring raises data protection issues – including that of cybersecurity, data minimisation and consent. Even carelessness or lack of attention over data handling can result in data falling into the wrong hands, potentially resulting in severe consequences for rights holders – such as loss of life, imprisonment or torture.[44] This is an especially pertinent issue in the transitional justice context where datasets normally contain sensitive information documenting experiences from conflict, including details on ethnicity, family members and political affiliations.[45]

AI tools developed by private companies are also being increasingly deployed, raising concerns over the incompatible motivation of private companies (focused on commercial gain) versus human rights and humanitarian objectives. To be sure, there is nothing inherently problematic about the use of emerging technologies such as AI, even when it is developed or procured from private companies. However, a human rights impact assessment as well as smaller scale testing (in sandboxes or pilot phases) on its possible use should be required before it is deployed.

The use of technology for the benefit of humanity must be situated within a sociotechnical context, namely taking contextual appropriateness and socio, political and economic conditions into account.[46] Parachuting technological solutions without consultation and assessment is not only inappropriate but, in effect, treats marginalised and vulnerable populations as a testing ground, going against the very core humanitarian principle of doing no harm.[47] Further, the use of AI, especially generative AI models that consume vast amounts of energy, can bring negative environmental impacts, hindering one of the key UN SDG goals of addressing climate change and adversely impacting human rights.

43. Mark Latonero, "Stop Surveillance Humanitarianism," *The New York Times*, July 11, 2019, https://www.nytimes.com/2019/07/11/opinion/data-humanitarian-aid.html.

44. Kate Hodal, "UN Put Rohingya 'at Risk' by Sharing Data Without Consent, Says Rights Group," *The Guardian,* June 15, 2021, https://www.theguardian.com/global-development/2021/jun/15/un-put-rohingya-at-risk-by-sharing-data-without-consent-says-rights-group.

45. Robert Flummerfelt and Nick Turse, "Online Atrocity Database Exposed Thousands of Vulnerable People in Congo," *The Intercept*, November 17, 2023, https://theintercept.com/2023/11/17/congo-hrw-nyu-security-data.

46. Margie Cheesman, "Conjuring a Blockchain Pilot: Ignorance and Innovation in Humanitarian Aid," *Geopolitics (August 2024): 1-28, https://doi.org/10.1080/14650045.2024.2389284*

47. Kristin Bergtora Sandvik, Katja Lindskov Jacobsen, and Sean Martin McDonald, "Do No Harm: A Taxonomy of the Challenges of Humanitarian Experimentation," *International Review of the Red Cross* 99, no. 904 (April 2017): 319–44, https://doi.org/10.1017/S181638311700042X.

d. AI and transitional justice: Potential ethical and human rights considerations

The human rights impacts, both positive and negative, have been addressed in different parts of this paper. Many of the issues on the use of AI and the transitional justice context are similar to the concerns mentioned before, namely privacy, lack of transparency impacting the rule of law, bias and non-discrimination and impacts upon economic, social, cultural, civil and political rights. This section will zoom in on potential impacts more specifically addressed to the transitional justice context, the **focus on victim-centredness and the goals of transitional justice for reconciliation, recognition, institution and trust-building and sustainable peace.**

First, the human rights and ethical concerns already highlighted can be **amplified in the transitional justice context**. For example, while the lack of representation from minority and vulnerable groups in training datasets can result in AI systems that are inaccurate and perpetrate further discrimination, the stakes are higher in transitional justice contexts when non-representative data is used to train AI systems and where systems are designed without the input of victims and victims' groups. In this case, the needs of especially vulnerable and marginalised groups can be entirely sidelined or not taken into account in accountability mechanisms. Specifically, this can range from inadequately designed and administered reparations—for instance, biases in identifying and registering victims, assessing their needs, and determining the suitable type of reparation—to the neglect of specific victim needs and harms, such as those experienced by male victims of sexual violence.

Unclear or non-existent data policies can also mean that victims' data can be used for unintended purposes, or worse, end up in the wrong hands. In the worst-case scenario, this may lead to loss of life, torture and other human rights harms, including the re-traumatisation of victims – who now have to live with the uncertainty of not knowing who has their data and what it will be (mis)used for at some future point. **Transitional justice actors working with AI and data have to ensure that policies and best practices are in place in relation to data use – including getting informed consent, respecting data minimisation, access and amendment and other established data protection principles.**

Second, AI can be an enabler for transitional justice work beyond the focus on retributive mechanisms such as judicial means of achieving accountability. Despite the established practice within the field of human rights and humanitarian law in general in using AI for purposes of monitoring and documentation, which in turn facilitates evidence gathering and accountability through judicial means, AI can both directly and indirectly be used **to raise awareness of transitional justice mechanisms, engage with victims, affected communities and other stakeholder groups** and can even be used to itself address some harms associated with the use of AI such as bias. Thus, **while one can start with low-hanging fruit such as using AI for documentation and archiving, widening the aperture to other use cases means that AI can also benefit the wider goals of transitional justice**, including for restorative justice, truth-telling and peace-building.

Besides the abovementioned points, AI can also pose novel concerns for the transitional justice context. Transitional justice involves different types of mechanisms to enable communities to address the legacy of violence and systemic human rights violations that have occurred in the past. The post-conflict context of transitional justice is one defined both by fragility – in terms of the security situation, access to basic socio-economic entitlements, power struggles and control of narratives - as well as context specificity when it comes to suitable transition justice mechanisms.

First, the rise of disinformation and misinformation enabled by AI, including through the use of deep-faked imagery but also generative AI tools such as image, audio and text generation, can actively enable hostile counter-narratives and distort the fragile and contested post-conflict information environment. This can directly impact truth-telling efforts by grassroots and community groups and "undermine the work of human rights documenters and advocacy efforts for the justice of victims."[48] Addressing such a challenge might require taking **a multi-pronged response,** involving not only regulation but also promoting AI literacy amongst civil society, victims' groups and affected communities, addressing the role and responsibilities of platforms that host and curate such content and working with existing organisations or efforts to address content provenance and authenticity, such as through the Coalition for Content Provenance and Authenticity (C2PA).[49]

An illustration that can be viewed in any direction. It has several scenes within it: people in front of computers seeming stressed, a number of faces overlaid over each other, squashed emojis and other motifs. Clarote & AI4Media / Better Images of AI / User/Chimera / CC-BY 4.0

48. Concept Note, Impunity Watch Expert Group Meeting, October 2024

49. C2PA is a project founded by industry actors such as Microsoft and Adobe which aims to build technical standards to understand authenticity and provenance of different types of media: www.C2PA.org.

Relying upon a regulatory response alone is insufficient as it is hard to define with (legal) certainty what disinformation is and in seeking to criminalise misinformation and disinformation, one might end up trampling on freedom of expression in the process. In essence, efforts in this area should be guided by the fact that it **is not only about fighting what is fake but also protecting what is true.** The latter is core to transitional justice work in archiving and documenting atrocities and foregrounding the experiences of victims.

Second, the harms addressed by transitional justice mechanisms are largely shaped by the perspectives of key actors (e.g. states), identification of perpetrators and the availability and viability of regulatory responses. However, AI can also enable **new harms that do not necessarily fall within existing frameworks of accountability.** In other words, the old frameworks of accountability might not match or cover the new harms that AI can potentially pose. For example, the ease of engaging in spreading disinformation is enabled by AI. This might not harm individuals as such but can impact the overall environment of trust towards information sources and institutions. The danger lies not in people believing false information as such, but rather in causing populations to **doubt and distrust all information** sources as a new default. Further, some forms of AI, such as through the use of facial recognition systems in identifying possible perpetrators might serve the cause of security and justice, but at the same time, facial recognition deployed for reasons of safety and security can be grossly disproportionate when weighed against its potential human rights impacts. This is in addition to the fact that such systems can have high false positive rates – meaning that someone can be wrongly identified by the system. In relation to transitional justice, the deployment of facial recognition systems and the risks of false positives in identification, for example, of missing persons, can mean that survivors, family members or other next of kin are given false hopes.

It is essential to not view AI as a monolith technology but to assess its use cases according to the context and through a human rights lens, including **assessing its impacts towards wider groups of potentially affected persons and at the societal level.**

Third, it is not only the forms of harms engendered through the use of AI that can seem uncertain and should be considered. The landscape of **actors that can cause widespread harm are also arguably expanding.** The widespread availability of AI tools such as generative AI used for creating fake imagery and audio, as well as the scale of algorithmically enabled spreading of information, including through social media platforms, can mean that individual acts can result in widespread harm. The effects can be even more pronounced where such harms are intended by nefarious actors.

Another important consideration as a potentially new facilitator of harms are **private companies**. While transitional justice has directed its efforts towards addressing wrongdoing and accountability to state and non-state actors such as militia groups, the emergence of private companies as enablers of violence and human rights abuses in a conflict-related context is a relatively recent phenomenon. The 2018 UN fact-finding mission in Myanmar, tasked to investigate whether genocide and other crimes against humanity had taken place in Myanmar,

found that Facebook was a "useful instrument" for spreading hate against the minority Muslim Rohingya population. Facebook (now Meta) has also been accused of facilitating human rights abuses in Tigray, Ethiopia as recently as 2022.[50] The company has taken measures such as increasing content moderators, removing harmful content on incitement to violence and monitoring for harms but these have been criticised as slow, inadequate and lacking in transparency.[51] Others call for reform of its attention economy business model.[52] While a platform does not directly cause harm, its, at times, outsized role in conflict settings requires a more critical form of engagement and examination of its roles and responsibilities.

Fourth, establishing truth and accountability in post-conflict situations can be complicated by **AI-enabled means and methods of warfare.** The use of AI support systems such as Lavender[53] used in the context of the conflict in Israel is one high-profile example of how AI can escalate conflict, obfuscate chains of responsibility and accountability. AI-supported target generation is meant to minimise collateral damage in warfare but the speed in which it generates targets can mean that many more lives end up being lost when compared to conventional warfare with humans checking every potential target. The scale and speed of war can be greatly increased with AI. AI-enabled drones and robots deployed in conflict settings can further blur chains of responsibility, including accountability mechanisms when things go wrong. International humanitarian law is premised upon human control and judgement, including through concepts such as distinction and proportionality, both of which critically engage human judgement. If we cannot be sure as to how human judgement is clouded, influenced or informed by AI, nor where and how AI played a role in conflict, then establishing truth and accountability post-conflict can similarly be complicated.

e. AI and transitional justice: Practical challenges

In addition to the human rights and ethical aspects, there can also be practical challenges in the adoption of AI in transitional justice settings. First, AI in civilian settings, whether these are systems used to make decisions on welfare, fraud or healthcare access, can draw from a wealth of data collected from the general population. This is similarly the case for generative AI systems that are trained on data available on the internet. In other words, these settings are not affected by a lack of data. However, **it can be difficult to get adequate and high-quality data for human rights monitoring and investigation purposes in transitional justice settings.**

For example, in order to train an image recognition model on the use of illegal weapons, an AI model needs to have access to a vast number of images showing these (and other) weapons. It is only by being trained with many of these images that a model can with a high level of confidence identify illegal weapons in new images it encounters. However, this is not data that can easily

50. "Ethiopia: Meta's Failures Contributed to Abuses against Tigrayan Community during Conflict in Northern Ethiopia," Amnesty International, October 31, 2023, https://www.amnesty.org/en/latest/news/2023/10/meta-failure-contributed-to-abuses-against-tigray-ethiopia.

51. Caroline Crystal, "Facebook, Telegram, and the Ongoing Struggle Against Online Hate Speech," *Carnegie Endowment for International Peace*, 7 September 2023, https://carnegieendowment.org/2023/09/07/facebook-telegram-and-ongoing-struggle-against-online-hate-speech-pub-90468.

52. 'Ethiopia: Meta's Failures Contributed to Abuses against Tigrayan Community during Conflict in Northern Ethiopia' (n 50).

53. Yuval Abraham, "'Lavender': The AI Machine Directing Israel's Bombing Spree in Gaza," *+972 Magazine*, 3 April 2024, https://www.972mag.com/lavender-ai-israeli-army-gaza - accessed 29 October 2024.

be found on the internet. The same goes for data that pertains to human rights work. Thus, in order to properly train AI models so that they can assist in future human rights monitoring work, one would need to either conscientiously and painstakingly gather from existing sources or use synthetic data, that is to say, data that is not real but which mirrors characteristics that one is interested in. The latter was the approach taken by the Syrian Archives.[54] **In essence, the utility of AI depends heavily on the availability of reliable and good quality data.**

While synthetic data can be a possible solution, especially in its potential to address biased data, this can also be time-consuming and costly to produce and might also not be fit for purpose. More generally, this point underscores the fact that **AI systems for specific aspects of transitional justice work (e.g. documentation, truth-telling, gathering evidence for prosecution, etc.) would likely need to be tailor-made specific to those purposes or - where procured from external providers - be customised to a large extent in order to be fit for purpose.** While some AI tools, such as those used for translation, can theoretically be used off the shelf, it is not – in practice - inclusive as many minority languages are not included or are of low quality.[55] Instead, smaller models (as opposed to large language models) may be a more suitable option as these can be trained using the most directly relevant datasets and the model can be designed for specific use cases (e.g. addressing transitional justice-related queries from the public).

More importantly, victim representation and involvement during the ideation and design process is critical. Clarifying how, to what extent and which victims (in relation to representation and intersectionality) to involve is a vital question to be addressed.

Secondly, another practical consideration is the cost and access to AI technology. While there is a lot of media hype about AI, the **availability and access of AI technology is not equal across the board.** The **digital divide** is still a reality in many countries and regions, where access to basic forms of technology such as mobile phones or the internet remains low or cost prohibitive. The digital divide exists not only in infrastructure and access to technology but also in terms of AI literacy and understanding of its benefits and pitfalls. Thus, when thinking about AI adoption for various aspects of transitional justice work, the realities of the country of operation must to be taken into account. If local populations have unequal access to basic technologies and have little understanding of the benefits or burdens of AI, it is incumbent upon the organisation considering the possible adoption of AI in transitional justice contexts to take these factors into account and avoid parachuting in technology without a careful consideration of suitability and the participation of stakeholders.

At the same time, it is not only access to AI technologies in general that are unequal. Access to AI talent and skillsets are also unequal. Organisations intending to use or adopt AI in their work have to factor in the high cost associated in attracting technical talent and also consider coalitions or

54. Raja Abdulrahim, "AI Emerges as Crucial Tool for Groups Seeking Justice for Syria War Crimes."

55. Quote from the fifth report of the Independent Investigative Mechanism for Myanmar (IIMM), 2023, 14 "the translation of digital Burmese text into English using artificial intelligence technology is a challenge, as such technology is not available for Myanmar languages, or is not as developed as it is for other languages," https://iimm.un.org/wp-content/uploads/2023/08/G2312500.pdf.

cooperating with other organisations already involved in this space where viable. A list of existing organisations already involved in this space is provided in Annex 1.

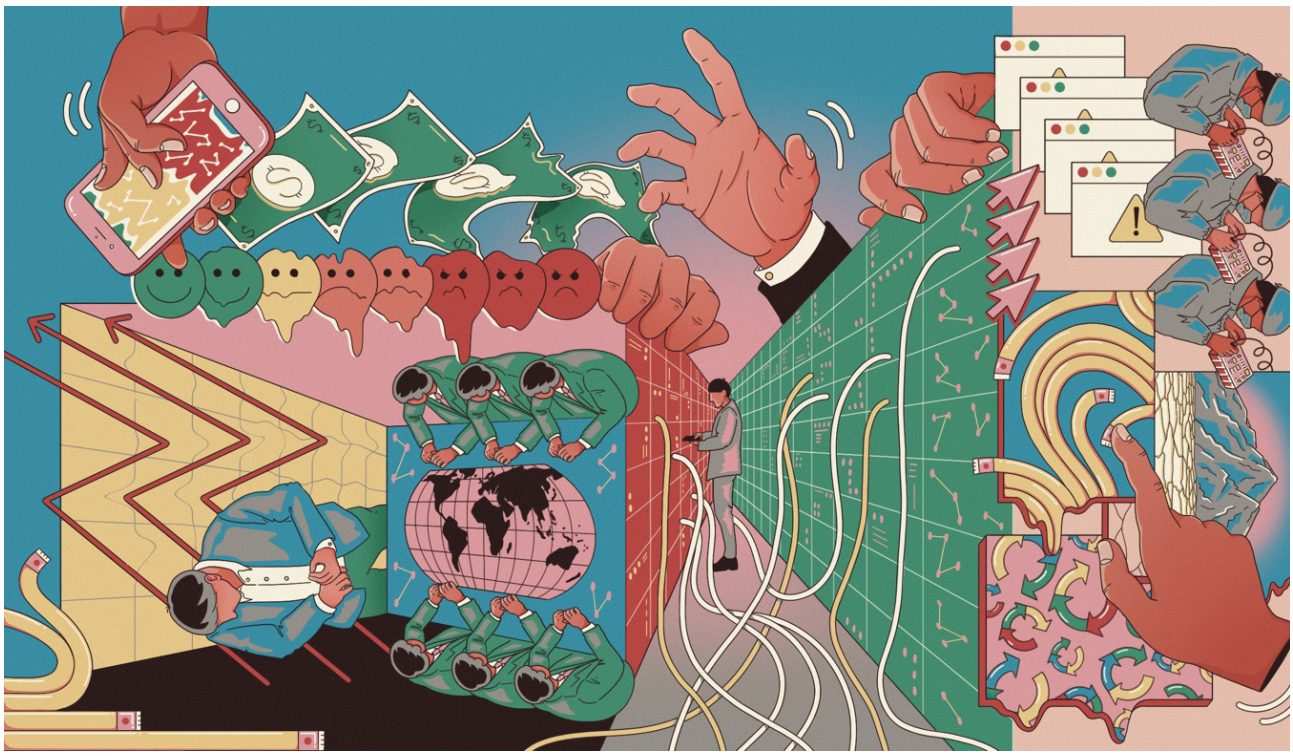
f. The AI regulatory and governance landscape

The concerns and implications of the increased deployment of AI across society has led to calls for it to be regulated. The first comprehensive legislation on AI came from the EU, as noted earlier in this report. The EU AI Act, which came into force in 2024, takes a risk-based approach, meaning that the regulation targets the most problematic use cases of AI. Use cases of AI deemed detrimental to health, safety and fundamental rights have been banned outright. This includes AI systems that are manipulative and those that exploit a person's vulnerability based on disabilities, age or socio-economic status that could lead to significant harm. Others such as social scoring, where each individual is given a score which determines access to goods and services, are also banned.[56] The legislation also designated certain uses of AI as high risk and imposes many obligations on providers of those systems. These include where AI is used in critical infrastructure, remote biometric identification systems, public services, education and employment settings amongst others. Where not deemed high risk, AI providers have the obligation to provide transparency to users. General purpose AI models such as large language models are also regulated – requiring transparency on how the model was trained and on copyright obligations. General purpose AI models that are deemed to pose systemic risks to society must meet stricter requirements in terms of rigorously testing and documenting their model before deployment.

While the EU has led the way in terms of AI regulation, other jurisdictions have also developed legislative and governance structures. However, different value systems underlie such regulation.[57] Broadly speaking, the EU approach is based upon respect for fundamental rights, the US approach is informed by a free-market and innovation-driven ideology while China takes a top-down state control approach. The US does not have a federal law regulating AI but has passed an executive order governing the use of AI in public services and AI systems where there are national security or public safety concerns. Different states in the US (e.g. New York and California) are also pursuing sector-specific legislation.

56. Article 5 Regulation (EU) 2024/1689 of the European Parliament and of the Council of June 13, 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

57. Anu Bradford, *Digital Empires: The Global Battle to Regulate Technology* (New York: Oxford University Press 2023).



An illustration that can be viewed in any direction. It has many elements to it working together: men in suits around a table, someone in a data centre, big hands controlling the scenes and holding a phone, and people in a production line. Motifs such as network diagrams and melting emojis are placed throughout the busy vignettes. Clarote & AI4Media / Better Images of AI / Power/Profit / CC-BY 4.0

AI regulation in China is largely based on specific AI use cases – with legislation focusing on data protection, content and information generated and disseminated online, and algorithmic decision-making on individuals. Other countries, such as Brazil and Canada, are in various stages of having comprehensive AI legislation in place while many are choosing to regulate by sector (e.g. data protection) or based on specific issues (e.g. disinformation). These divergent approaches may signal fragmentation and lack of agreement on risks as well as values.

Multilateral and international organisations are also shaping the AI governance space. The Organisation for Economic Co-operation and Development (OECD) AI Principles were adopted in 2019 to steer the development of trustworthy AI. This was followed by the UNESCO Recommendations on the Ethics of AI which was adopted by all member states in 2021. A notable development is the Council of Europe's Framework Convention on AI, Human Rights, Rule of Law and Democracy, the first international, binding treaty on AI.[58] This convention takes a similar risk-based approach to the EU's AI legislation and was opened for signature on September 5, 2024.

The United Nations is also entering the AI governance space. The UN High-Level Advisory Body on Artificial Intelligence was set up in October 2023 to create multilaterally-driven international governance of AI. The UN entering the AI governance space may also be a catalyst for core consensus to be achieved in regulating AI, hence tempering the divergent approaches outlined

58. Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, CETS No. 225, September 2024, https://rm.coe.int/1680afae3c.

above. The advisory body has since released their interim report on governing AI for humanity.[59] While not wholly focused on AI as such, the UN Global Digital Compact aims to address the digital, data and innovation divides that exist and outlines shared principles for an open, free, and secure digital future for all. This was part of the Pact for the Future that was adopted at the UN Summit for the Future which took place in September 2024, branded as a "once-in-a-generation opportunity to reinvigorate global action, recommit to fundamental principles, and further develop the frameworks of multilateralism so they are fit for the future."[60]

The rapid advancement of generative AI has also ignited governance momentum. The G7 adopted the Hiroshima AI Process Comprehensive Policy Framework in December 2023 to guide the development of safe, secure and trustworthy advanced AI systems. The UK convened the AI Safety Summit in Bletchley Park in November 2023, which resulted in the adoption of the Bletchley Declaration focusing on the development of safe frontier AI, namely advanced AI systems that can pose extreme risks to safety and humanity.

Several AI safety institutes have also been set up to govern advanced frontier AI.[61] This is by no means a comprehensive account of the regulatory and governance frameworks that exist to govern AI. Many more are in development or continual adoption. However, despite this governance and regulatory momentum, it remains to be seen how regulatory requirements are adhered to in practice and the potential practical challenges such regulation might encounter. One key implementation concern is how well general high-level legal texts and principles translate into potentially technical or operational measures (e.g. to counter bias).

g. Lessons for AI governance through the lens of transitional justice

Three key observations can be gleaned from these regulation efforts and how this relates (or does not relate) to transitional justice.

First, diverse approaches on regulation and governance are motivated by different underlying values. Transitional justice goals of accountability and reconciliation do not directly feature in any governance efforts. However, victim-centredness indirectly informs the EU approach – where fundamental rights are a driving force of the legislation. Those who deploy AI systems in a public sector setting must carry out a fundamental rights impact assessment. Further, bias detection and mitigation measures have to be in place for providers of systems for high-risk AI use cases. While these are not victim-centred as such, it is informed by research of **how AI systems have impacted individuals and is driven by the need to centre the experiences of individuals who will encounter these AI systems**, including marginalised and vulnerable groups.

59. "Interim Report, Governing AI for Humanity," UN High Level Advisory Body, December 2023, https://www.un.org/sites/un2.un.org/files/un_ai_advisory_body_governing_ai_for_humanity_interim_report.pdf

60. "Summit of the Future," SDG Knowledge Hub blog, accessed October 29, 2024, https://sdg.iisd.org/events/summit-of-the-future/.

61. Prithvi Iyer, "From Safety to Innovation: How AI Safety Institutes Inform AI Governance,", *Tech Policy Press*, October 25, 2024, https://techpolicy.press/from-safety-to-innovation-how-ai-safety-institutes-inform-ai-governance.

Secondly, AI used for national security as well as for military and defence purposes is excluded from the majority of the regulatory and governance measures. Instead, governance in this area is nascent, despite the long-standing focus on military uses of AI (including the fear of "killer robots" on the battlefield).[62] The International Committee of the Red Cross (ICRC) formed the International Committee for Robot Arms Control in 2009 and the large-scale Campaign to Stop Killer Robots, launched in 2013, both focused on the issue. In 2013, a meeting of state parties to the convention on prohibitions or restrictions on the use of certain conventional weapons which may be deemed to be excessively injurious or have indiscriminate effects (Convention on Certain Conventional Weapons – CCW) mandated the examination of the legality and governance around lethal autonomous weapons systems (LAWS). A Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System (GGE LAWS) was in turn formed in 2016 for this purpose. Despite its dedicated mandate, progress on the issue of regulation has been slow. However, it has put forth a two-tiered approach – where certain systems unable to comply with international humanitarian law would be prohibited, while others would be restricted. Work in this area is ongoing.

However, the lack of specific clarity around AI in warfare does not mean that there is a legal vacuum. International humanitarian law continues to apply – including principles of distinction, proportionality, necessity and humanity. At the same time, these concepts depend on and engage with human judgement because what is considered necessary and proportionate may depend on the circumstances and cannot be easily generalised.

As it stands today, the existing lack of clarity on red lines in governing the use of military AI complicates not only the regime of accountability under international humanitarian law but also the goal of attaining justice and truth-seeking under transitional justice. Fundamentally, the lack of clarity surrounding the use of AI in military settings signals the unwillingness of states to give up their capacity to develop and deploy AI as a matter of military and political advantage. This is antithetical to a victim-centred and peace-driven perspective of transitional justice that is based on the resolution, rather than the potential escalation, of conflict.

Third, even with the emergence of regulatory and governance structures, the benefits and burdens of AI remain unequally distributed. In relation to benefits, the digital divide that exists means that many countries from the Global South do not get access and do not, therefore, equally benefit from AI. This includes within the field of human rights monitoring generally and transitional justice specifically. The task of establishing truth and accountability remain very much a manual exercise of interviewing, surveying and gathering information from victims on the ground. This is partly due to necessity – respecting the agency and human dignity of victims - but also partly due to the lack of equitable access to technology more broadly speaking. However, while the digital divide should be closed, this need not entail an uncritical wholesale embrace of AI. Instead, the use of AI within the transitional justice context should be informed by the perspective of respecting victims (see Section 3 below).

62. Recent efforts in this area include the Summit on Responsible Artificial Intelligence in the Military Domain (REAIM) hosted by the Netherlands in 2023 and US initiative the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy.

As for unequal burdens, AI governance has been dominated by the experiences and influence of companies that tend to be concentrated in the Global North. In contrast, the harms and experiences of those from the Global South do not feature as prominently, despite significant AI-related violations and harms occurring in the Global South and in autocratic nations with limited respect for human rights and the rule of law and where there is minimal accountability. This disparity highlights a critical gap in the global AI governance discourse that focuses on narrowing the digital divide in terms of access to technology. This assumes that AI can only ever be an unalloyed good and that its benefits should flow from the "haves" to the "have nots". Instead, both a more critical adoption of technology alongside the protection of victims and a meaningful accounting of how AI can disproportionately harm those in less regulated environments should feature more prominently in the governance and policy debates.

Finally, as AI is making inroads into more ethically contested domains, including in mental health, warfare and as personal companions, **the need to take a critical, informed and human rights-based approach towards its use is increasingly urgent.** As transitional justice measures are informed by centring the needs of victims, so too can AI governance learn by engaging in participatory mechanisms, including by proactively centring the lived experiences of individuals and of those belonging to vulnerable and minority groups in the initial ideation and design stage of AI systems. Much regulation centres around protecting against potential harms and using the law to punish acts that lead or contribute towards harm. However, a key transitional justice perspective can play a role in a related area – namely agency. **Instead of merely avoiding harms, AI governance can also aim to promote agency** – by having in place mechanisms, procedures and policies that enable people to exercise and enjoy their human rights, rather than merely avoiding the worst instances of harm.

3. DEVELOPING VICTIM-CENTRED AI FOR TRANSITIONAL JUSTICE

This paper has so far explored where AI is used (including in the transitional justice context), human rights and ethical concerns, and various regulatory and governance approaches. This section examines how best to design and implement AI in the transitional justice context.

The paper proposes that developing victim-centred AI for transitional justice should be informed by three key parameters, drawing from the best practices and principles derived from three different, but inter-connected, fields (transitional justice, international humanitarian law and human rights). These are: victim centredness, the principle of do no harm and a human rights-based approach.

In framing a victim-centred operationalisation framework of AI for transitional justice, it is helpful to approach it from three different levels (persons, practice, policy) – in order to capture the different facets of the promises and perils of AI in the transitional justice context. Examining it from these three different facets provides a more holistic perspective on the role and impacts of AI, from centring victims in design, using AI to advance transitional justice work and the larger policy implications, such as the impacts of AI on the normative premises of transitional justice. This avoids the tunnel vision of seeing how AI is used in specific areas (e.g. documentation and analysis) – while ignoring its institutional and wider contextual impacts.

a. Centring people: Participation and empowerment of victims through AI

In adopting a victim-centred approach, the participation of victims or victims' groups in the decision-making process on how, whether, when and where to use AI is essential. Victims are experts on their own lived experiences and including their voices and views is not only essential but also a key element of the human rights-based approach of including potentially affected stakeholders in the consultation process in order to assess potential human rights impacts. AI-enabled participation tools to gauge consensus on key concerns, needs and interests such as the open source pol.is platform,[63] can also be used where judged to be contextually suitable and appropriate.

However, participation taken in isolation is insufficient. To meaningfully participate and recover their agency, victims must be empowered – including through AI literacy efforts and capacity building, enabling informed decision-making on the suitability of the adoption of AI for any given purpose. Thus, victims' groups and representatives should have enough knowledge of the benefits

63. Polis is a real-time system for gathering, analysing and understanding what large groups of people think in their own words, enabled by advanced statistics and machine learning.

and negative impacts of using AI, both generally (e.g. in relation to its human rights impacts) and in relation to particular tools under consideration. At the same time, we should be cognisant in not over-burdening victims or victims' groups.

Empowerment also entails that adequate data policies are in place to protect individual data; this is a pre-requisite of whether or not AI is used. In turn, another way to view empowerment is that AI should be used to empower victims – whether it be furthering their agency through truth-telling, identifying gaps in needs assessment or improving efficiencies around documentation and archiving practices. This is to ensure that **AI adoption in the transitional justice context does not proceed from technological solutionism, where efficiency, cost and time savings and scaling come at the cost of victim participation and empowerment.**

Where AI is adopted, human rights and ethical concerns covered elsewhere in this paper should feature in its design and deployment, including through a human rights and ethical impact assessment. This includes ensuring that systems can accommodate minority languages (e.g. in content scanning or chatbots). AI systems, especially with predictive capabilities, should also be rigorously tested in sandboxes or pilot phases before being more widely used.

b. Improving practice: Enhancing transitional justice mechanisms using AI

Another facet to help us think about AI in the transitional justice context involves looking at how AI can assist in the procedures and processes involved in transitional justice mechanisms. It is prudent to start by looking at current practices. Transitional justice mechanisms, whether judicial or non-judicial, primarily involve capturing the experiences of victims using different means (e.g. community dialogues, interviews, surveys, etc.) and documenting or archiving testimonies, official records and other relevant past practices or evidentiary material. AI is particularly suited for tasks that involve data analysis and pattern detection using data. The exact methodology of deployment and technique used should be discussed and designed with the technical team from the very beginning. While it is also possible to use AI-enabled analysis on existing datasets, the data quality in question should both be suitable and labelled, where required, in order for the datasets to be actionable. Sandbox testing and **pilot projects** should be carried out before any large-scale deployment or use of AI tools.[64]

AI can be deployed to analyse text, images and audio sources in order to find relevance between distinct datasets and materials but also more concretely to gauge evidentiary weight and help build a legal case, as the high-profile example of the use of AI at the ICC shows.[65]

AI can also be used to factcheck or to gauge the authenticity and provenance of content. Efforts such as C2PA are backed by the industry and promoted by human rights organisations as a

64. For an example of pilot projects, see Elena Radeva, 'The Potential for Computer Vision to Advance Accountability in the Syrian Crisis' (2021) 19 Journal of International Criminal Justice 131.

65. Gabrielle McIntyre and Nicholas Vialle, "The Use of AI at the ICC: Should We Have Concerns? Part I," *Opinio Juris*, 11 October 2023, https://opiniojuris.org/2023/10/11/the-use-of-ai-at-the-icc-should-we-have-concerns-part-i - accessed 3 September 2024.

technical standard to trace the origin of different types of media. Where nefarious actors or individuals can now spread malicious content, falsehoods or even misinformation at scale with AI, it is necessary to ensure that the information sphere is not undermined, especially in fragile post-conflict and transitional environments. This is also necessary to preserve the integrity of the mechanisms such as truth-telling, to prevent the distortion of facts, including historical events. However, deepfakes, mis- and disinformation are sociotechnical phenomena, where technical affordances interact with social and political contexts. One should not expect technical solutions to fully address sociotechnical problems. For example, people may trust content due to their emotional resonance to it, regardless of whether or not it is objectively true or otherwise.

As mentioned, it is theoretically possible to utilise AI technologies for other transitional justice mechanism processes such as the calculation of individual financial reparations[66] or in gauging victim sentiment on the suitability of certain transitional justice mechanisms and to better assess and visualise victim needs. Similarly, a human rights-based approach to accountability measures, including the possibility for victims to understand and mount challenges, should also be afforded and be made accessible. In other words, there should be human accountability built into the process. AI-powered chatbots that are able to answer queries from the public on the specific transitional justice mechanism, its legal basis and other technical or procedural questions can similarly be deployed, but with caution and oversight. As covered in other examples, there are also many ways in which AI can be utilised to raise awareness and public engagement.

In considering using AI for transitional justice, **possible collaborations with organisations already working or having expertise within this space is to be encouraged as this can leverage existing skillsets and familiarity in navigating the ethical and human rights challenges of using AI in fields adjacent to transitional justice.** An initial mapping of these organisations and their focus areas is available in Annex 1. Where possible, these collaborations should also be interdisciplinary in nature – to leverage expertise from different fields.

Alternatively, smaller scale joint projects could also be developed with computer science departments at universities or other institutions of higher education. A needs and gap assessment should be undertaken, in addition to stakeholder consultations (with victims or representative victims' groups) before the comprehensive introduction on the use of AI. In order to guide organisations working with transitional justice on how to work with AI (and ensuring victim participation in the process), a set of guiding questions can be found in Annex 2.

66. Which should, however, live up to the international standard of providing an "adequate, effective and prompt reparation for harm suffered" that is "proportional to the gravity of the violations and the harm suffered" in accordance with the 2005 UN Basic Guidelines on the Right to a Remedy and Reparations for Victims of Gross Violations of Human Rights and Serious Violations of Humanitarian Law, General Assembly resolution 60/147, 15 December 2005.

c. Policy engagement: AI and the wider context of transitional justice

In addition to empowering victims and enhancing processes and procedures of transitional justice mechanisms, **AI can also inform and improve on transitional justice practices.** Comparisons across different transitional justice datasets, disaggregated according to country, region or mechanism are made possible with the existence of publicly available datasets, enabling a macro view in gauging how effective transitional justice policies are and how well they address victim expectations and pursuing justice and accountability. It can help to determine practical bottlenecks but also provide clarity on whether or not theoretical and policy goals have actually manifested in practice. Such comparisons can already be undertaken using existing open-source databases available on platforms such as TJET, either as standalone data or cross-referenced with datasets from other platforms.[67]

The work on gauging effectiveness and meeting victim expectations need not only be gleaned from existing datasets but also through interviews, surveys and other tools where AI can then be used to analyse the data, gauge sentiment or detect patterns on gaps and challenges. At the same time, **sources of data can differ in terms of quality, independence and coverage and these limitations should be assessed and recognised when using them to conduct research.**

As mentioned before, the information environment is increasingly both mediated by AI systems (e.g. through social media platforms) and enabled by AI (e.g. deep faked images, disinformation). However, from a macro perspective, due to the inundation of information available online, it is not feasible to rely on individuals or institutional efforts alone (e.g. journalists) to detect, verify and gatekeep individual pieces of content. It may be inevitable for AI to be used to either mark, warn, label or flag content. Actors working in the transitional justice space can collaborate with organisations already working in this space[68] and be continually informed about the changing landscape of content provenance and verification.

The increasing influence and role of social media platforms in playing a contributory role in facilitating harms and systemic violence during conflict highlights a gap – that of accountability. Private sector involvement in the conflict setting is not new as such, for example the deployment of private military contractors or the use of privately developed AI-enabled weapons or decision support systems. It can also take an indirect form, such as through social media platforms and its content moderation policies. The business model of engagement dependent upon an attention economy and virality – combined with a lack of attention to local contexts - can have devasting consequences, as we have seen in the situation of the Rohingya in Myanmar. The UN Fact-Finding mission concluded that: "[The] role of social media is significant. Facebook [now known as Meta]

67. Examples include the Uppsala Conflict Data Program, Armed Conflict Location and Event Data (ACLED), and the Rule of Law in Armed Conflict Project (RULAC).

68. One example is human rights NGO Witness, see "Deepfakes, Synthetic Media and Generative AI", https://www.gen-ai.witness.org/?pk_vid=845ad2b23831425a1725955611f950bf

has been a useful instrument for those seeking to spread hate, in a context where, for most users, Facebook is the internet."[69] A subsequent independent report commissioned by the platform found that "the troubling legal context in Myanmar, when combined with the widespread use of Facebook and other social media platforms for character assassinations, rumor-spreading, and hate speech against minority individuals, creates an enabling environment for the ongoing endorsement and proliferation of human rights abuse"[70]. In response, Facebook agreed that they were not doing enough "to help prevent our platform from being used to foment division and incite offline violence".[71] Attempts to gain forms of accountability and clarity in relation to Facebook's role in particular, but also the role of social media and platforms in conflict generally, are ongoing,[72] including through the private rights of action, where Facebook has been sued in certain foreign jurisdictions.[73]

This demonstrates the need to clarify the role and forms of accountability for private actors in the context of conflict but also during post-conflict periods, especially as platforms are emerging as the default means of communication.

69. "Report of the independent international fact-finding mission on Myanmar" UN, para. 74, A/HRC/39/64, 12 September 2018, https://www.ohchr.org/sites/default/files/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf.

70. "Human Rights Impact Assessment: Facebook in Myanmar", Business and Human Rights Resource Centre (BSR) report commissioned by Facebook, October 2018, https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf.

71. Alex Warfoka, "An Independent Assessment of the Human Rights Impact of Facebook in Myanmar," Meta website, November 5, 2018, https://about.fb.com/news/2018/11/myanmar-hria/ . See also
"Facebook Admits It Was Used to 'incite Offline Violence' in Myanmar," *BBC News*, 6 November 2018, https://www.bbc.com/news/world-asia-46105934.

72. "Human Rights Impact Assessment: Facebook in Myanmar", BSR, October 2018, https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf. Others have criticised the adequacy of the assessment, including: Mark Latonero and Aaina Agrawal, "Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar", (2021) 13, https://www.hks.harvard.edu/sites/default/files/2023-11/2021_13_facebook-failure-in-myanmar_0.pdf.

73. The legal case of *The Gambia v. Myanmar* at the International Court of Justice is ongoing. The case focuses on state accountability.

4. RECOMMENDATIONS AND CONCLUSIONS

a. Key takeaways

- Despite AI having negative human rights and ethical impacts, there is increasing adoption of "AI for Good", including through high-level efforts at the UN.[74]
- Decision-making around the use and adoption of AI in the transitional justice context should always be guided by foundational values such as victim-centredness. Efficiency, cost, time savings and other gains typically associated with the use of AI should be guided by the objective of being in service of the needs of victims, respect for human dignity and be steered by the goals of transitional justice.
- The paper proposes three ways to consider where and how AI can play a role in the transitional justice context – the three Ps: centring the persons (victims), examining AI in the practical setting (practice) of transitional justice and finally policy. Key considerations include the need to adopt a human rights-based approach and a careful analysis of the benefits and burdens before adoption.
- Data security concerns should be a primary consideration, especially when sensitive data of vulnerable populations are used or stored in the cloud or used in training AI models. Actors working with data processing need to put in place adequate data protection and cybersecurity measures.
- Multidisciplinary approaches are vital for fostering collaboration between transitional justice and AI expertise. Platforms like the Delft initiative organised by Impunity Watch and TU- Delft in November 2024 demonstrate the added value of such collaboration and could serve as a strong foundation for future efforts.

b. Recommendations

The recommendations here mirror the three-part consideration in Section 4. These are not standalone recommendations but should be read as being complementary to each other. This Framework Paper should be also treated in the spirit of a conversation starter and a list of guiding questions are thus attached in Annex 2 on how best to navigate this area.

i. Centring people

- When considering ideas, feasibility or viability of using AI for transitional justice, victims' needs and wishes should be the primary factor to be taken into account.

74. See, for example, UN Global Pulse, https://www.unglobalpulse.org/ and AI for Good at ITU, https://aiforgood.itu.int/.

- AI technologies should never be tested on vulnerable populations. The do no harm principle and informed consent are minimum measures needed in thinking about whether or not AI is suitable or should be used for a given purpose.
- Tech literacy in general, but AI literacy in particular, is necessary to ensure that victims or victims' groups have sufficient knowledge about technologies and know how to seek explanations and accountability should it result in harmful impacts. This includes understanding the limits of what AI can do and the impact of AI in the information ecosystem of the post-conflict or democratic transition context.
- Technological solutions should never be parachuted in without an understanding of local socio-economic and socio-political context(s), including through consultations with local victims' groups.
- At the same time, due to the sociotechnical nature of AI and its potential to be (mis)used for political ends, especially in post-conflict settings, a wider mapping of potentially affected persons and the societal impacts of AI should be undertaken.

ii. AI and transitional justice in practice

- Organisations working with transitional justice can consider partnering up with organisations or joining coalitions of actors already deploying "AI for Good". See Annex 1 for a list of organisations already working in this space.
- Running a workshop or series of workshops in a practical creative lab setting to explore the possibilities of using AI in the transitional justice context.
- In identifying potentially new areas of focus or collaboration, a human rights-based approach to AI for transitional justice that foregrounds victim-centredness should be taken.
- Data protection policies should be in place or updated as necessary prior to and in conjunction with the use of AI in transitional justice contexts.
- Taking stock of and examine existing operational gaps (e.g. victim participation, gender gaps) as these could be exacerbated by the use and deployment of AI.
- If AI solutions are procured externally or in collaboration with others, there is still a need to raise awareness among staff about the capacities and limits of the AI system.
- In raising awareness, organisations should start with an in-depth series of studies exploring the potential use, use-cases and drawbacks of AI in different transitional justice mechanisms and processes. In doing so, a lifecycle approach towards AI (i.e. ideation, design, testing, implementation, follow-up and de-commissioning) should be undertaken.

- Building the tech and AI literacy capacity of victims, victims' groups or advocacy organisations
- Using AI to highlight gaps or challenges as well as successes of transitional justice mechanisms.

iii. Policy recommendations

- Developing a standard-setting document on the use of AI for transitional justice. This could take the form of a code of conduct, statement of principles, ethical principles or similar. Such a document could be developed together or in consultation with other organisations, victim groups, policymakers, the UN or other relevant stakeholders.
- Conducting a study exploring accountability frameworks or mechanisms clarifying the roles and responsibilities of private actors in conflict or periods of mass violence.
- Conducting human rights impact assessments before deploying AI, including AI used ostensibly for security and safety reasons (e.g. the use of AI facial recognition in public spaces).
- Having a multi-pronged approach in ensuring the resilience of the information sphere post-conflict or during the transition period – including strengthening local journalism and promoting AI literacy alongside encouraging the use of content and provenance tracing methods such as C2PA.
- Supporting efforts to establish archives for user-generated content from social media or other platforms that can assist in establishing accountability for serious human rights and humanitarian law violations.
- Taking a victim-centred and a human rights-based approach in considering, designing and deploying AI for transitional justice. This can include eventually deciding not to deploy AI where human rights impacts are judged to be disproportionate.

5. BIBLIOGRAPHY

Abdulrahim, Raja. "AI Emerges as Crucial Tool for Groups Seeking Justice for Syria War Crimes." *Wall Street Journal*, February 13, 2021. https://www.wsj.com/articles/ai-emerges-as-crucial-tool-for-groups-seeking-justice-for-syria-war-crimes-11613228401.

Abraham, Yuval. "'Lavender': The AI Machine Directing Israel's Bombing Spree in Gaza." *+972 Magazine*, April 3, 2024. https://www.972mag.com/lavender-ai-israeli-army-gaza/.

American Psychological Association. "What Psychological Factors Make People Susceptible to Believe and Act on Misinformation?" November 29, 2023, updated on March 1, 2024. https://www.apa.org/topics/journalism-facts/misinformation-belief-action.

Amnesty International. "Ethiopia: Meta's Failures Contributed to Abuses against Tigrayan Community during Conflict in Northern Ethiopia," October 31, 2023. https://www.amnesty.org/en/latest/news/2023/10/meta-failure-contributed-to-abuses-against-tigray-ethiopia/.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

BBC News website. "Facebook Admits It Was Used to 'Incite Offline Violence' in Myanmar." November 6, 2018. https://www.bbc.com/news/world-asia-46105934.

Beauvais, Catherine. "Fake News: Why Do We Believe It?" *Joint Bone Spine* 89, no. 4 (July 2022): https://doi.org/10.1016/j.jbspin.2022.105371.

Bloch, Kate E. "Virtual Reality: Prospective Catalyst for Restorative Justice." *American Criminal Law Review* 58, no. 2 (2021). https://www.law.georgetown.edu/american-criminal-law-review/in-print/volume-58-number-2-spring-2021/virtual-reality-prospective-catalyst-for-restorative-justice/.

Bradford, Anu. *Digital Empires: The Global Battle to Regulate Technology*. New York: Oxford University Press, 2023.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research* 81 (2018): 1-15. https://www.media.mit.edu/publications/gender-shades-intersectional-accuracy-disparities-in-commercial-gender-classification/

Cheesman, Margie. "Conjuring a Blockchain Pilot: Ignorance and Innovation in Humanitarian Aid." *Geopolitics* (August 2024): 1–28. https://doi.org/10.1080/14650045.2024.2389284.

Constantaras, Eva, Gabriel Geiger, Justin-Casimir Braun, Dhruv Mehrotra, and Htet Aung. "Inside the Suspicion Machine." *Wired*, March 6, 2023. https://www.wired.com/story/welfare-state-algorithms/.

Couldry, Nick, and Ulises Ali Mejias. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford, California: Stanford University Press, 2019.

De Greiff, Pablo. "Theorizing Transitional Justice." *Nomos* 51 (2012): 31–77.

de Zwart, Hans. "Dutch Institute for Human Rights: Use of Anti-Cheating Software Can Be Algorithmic Discrimination (i.e. Racist)." *Racism and Technology Center blog*, December 24, 2022.

Dulka, Anne. "The Use of Artificial Intelligence in International Human Rights Law." *Stanford Technology Law Review* 26, no. 2 (August 2023). https://law.stanford.edu/publications/the-use-of-artificial-intelligence-in-international-human-rights-law/.

Flummerfelt, Robert, and Nick Turse. "Online Atrocity Database Exposed Thousands of Vulnerable People in Congo." *The Intercept*, November 17, 2023. https://theintercept.com/2023/11/17/congo-hrw-nyu-security-data/.

Fraisl, Dilek. "The Potential of Artificial Intelligence for the SDGs and Official Statistics." *Paris21* Working Paper, April 2024. https://www.paris21.org/sites/default/files/related_documents/2024-04/the-potential-of-ai-for-the-sdgs-and-official-stats_working-paper_0.pdf.

Freeman, Lindsay. "Weapons of War, Tools of Justice: Using Artificial Intelligence to Investigate International Crimes." *Journal of International Criminal Justice* 19, no. 1 (March 2021): 35–53. https://doi.org/10.1093/jicj/mqab013.

Future of Life Institute. "Pause Giant AI Experiments: An Open Letter," March 22, 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Gavshon, Daniela. "How New Technology Can Help Advocates Pursue Transitional Justice." *OUPblog*, July 1, 2019. https://blog.oup.com/2019/07/how-new-technology-help-advocates-pursue-transitional-justice/.

Haunschild, Jasmin, Laura Guntrum, Sofía Cerrillo, Franziska Bujara, and Christian Reuter. "Towards a Digitally Mediated Transitional Justice Process? An Analysis of Colombian Transitional Justice Organisations' Posting Behaviour on Facebook." *Peace and Conflict Studies* 30, no. 2 (May 2024). https://nsuworks.nova.edu/pcs/vol30/iss2/4.

Henriques-Gomes, Luke. "Robodebt Class Action: Coalition Agrees to Pay $1.2bn to Settle Lawsuit." *The Guardian*, November 16, 2020. https://www.theguardian.com/australia-news/2020/nov/16/robodebt-class-action-coalition-agrees-to-pay-12bn-to-settle-lawsuit.

Hodal, Kate. "UN Put Rohingya 'at Risk' by Sharing Data without Consent, Says Rights Group." *The Guardian*, June 15, 2021. https://www.theguardian.com/global-development/2021/jun/15/un-put-rohingya-at-risk-by-sharing-data-without-consent-says-rights-group.

International Center for Transitional Justice. "At an Interactive Exhibit, Colombians Reflect on Their Country's Painful Past and New Possibilities for Its Future." Accessed August 26, 2024. https://www.ictj.org/latest-news/interactive-exhibit-colombians-reflect-their-country%E2%80%99s-painful-past-and-new.

Iyer, Prithvi. "From Safety to Innovation: How AI Safety Institutes Inform AI Governance." *Tech Policy Press*, October 25, 2024. https://techpolicy.press/from-safety-to-innovation-how-ai-safety-institutes-inform-ai-governance.

Latonero, Mark. "Stop Surveillance Humanitarianism." *The New York Times*, July 11, 2019. https://www.nytimes.com/2019/07/11/opinion/data-humanitarian-aid.html.

McIntyre, Gabrielle, and Nicholas Vialle. "The Use of AI at the ICC: Should We Have Concerns? Part I." *Opinio Juris*, October 11, 2023. https://opiniojuris.org/2023/10/11/the-use-of-ai-at-the-icc-should-we-have-concerns-part-i/.

Office of the United Nations High Commissioner for Human Rights (OHCHR). "Report of the independent international fact-finding mission on Myanmar, A/HRC/39/64." September 12, 2018. https://www.ohchr.org/sites/default/files/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf

OHCHR Secretary-General guidance note: "Transitional Justice: A Strategic Tool for People, Prevention and Peace." (October 2023): https://www.ohchr.org/en/documents/tools-and-resources/guidance-note-secretary-general-transitional-justice-strategic-tool.

OHCHR. "World Stumbling Zombie-like into a Digital Welfare Dystopia, Warns UN Human Rights Expert." Accessed August 23, 2024. https://www.ohchr.org/en/press-releases/2019/10/world-stumbling-zombie-digital-welfare-dystopia-warns-un-human-rights-expert.

Panic, Branka, and Paige Arthur. *AI for Peace.* Boca Raton: CRC Press, Taylor & Francis Group, 2024.

Radeva, Elena. "The Potential for Computer Vision to Advance Accountability in the Syrian Crisis." *Journal of International Criminal Justice* 19, no. 1 (March 2021): 131–46. https://doi.org/10.1093/jicj/mqab015.

Ressa, Maria. "We're All Being Manipulated the Same Way." *The Atlantic*, April 6, 2022. https://www.theatlantic.com/ideas/archive/2022/04/maria-ressa-disinformation-manipulation/629483/.

Roberts, Dorothy, and nia t evans. "The 'Benevolent Terror' of the Child Welfare System." *Boston Review*, March 31, 2022. https://bostonreview.net/articles/the-benevolent-terror-of-the-child-welfare-system/.

Russell, Stuart J., and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Third edition. Boston, Indianapolis: Pearson, 2016.

Sandvik, Kristin Bergtora, Katja Lindskov Jacobsen, and Sean Martin McDonald. "Do No Harm: A Taxonomy of the Challenges of Humanitarian Experimentation." *International Review of the Red Cross* 99, no. 904 (April 2017): 319–44. https://doi.org/10.1017/S181638311700042X.

SDG Knowledge Hub. "Summit of the Future." Accessed October 29, 2024. https://sdg.iisd.org/events/summit-of-the-future/.

Shepherd, Marshall. "Repeating Misinformation Doesn't Make It True, But Does Make It More Likely to Be Believed." *Forbes*, August 17, 2020. https://www.forbes.com/sites/marshallshepherd/2020/08/17/why-repeating-false-science-information-doesnt-make-it-true/.

Stanford University DigiChina project. "Full Translation: China's 'New Generation Artificial Intelligence Development Plan' (2017)." Accessed November 13, 2024. https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/.

Teitel, Ruti G. "Transitional Justice Genealogy (Symposium: Human Rights in Transition)." *Harvard Human Rights Journal* 16 (2003).

Transitional Justice Evaluation Tools. "Homepage: Transitional Justice Evaluation Tools." Accessed August 26, 2024. https://transitionaljusticedata.org/en/.

United Nations Department of Political and Peacebuilding Affairs (UNDPPA) Politically Speaking website. "Virtual Reality Bites: Using Technology to Bring Post-Conflict Situations to Life." August 11, 2022. https://dppa.medium.com/virtual-reality-bites-using-technology-to-bring-post-conflict-situations-to-life-bd5cb98ce3f6.

UNDPPA. "Futuring Peace: Exploring the Power of Generative AI." Accessed August 26, 2024. https://www.futuringpeace.org.

Vasdani, Tara. "Robot Justice: China's Use of Internet Courts." LexisNexis Canada. Accessed September 6, 2024. https://www.lexisnexis.ca/en-ca/ihc/2020-02/robot-justice-chinas-use-of-internet-courts.page.

Vinuesa, Ricardo, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. "The Role of Artificial Intelligence in Achieving the Sustainable Development Goals." *Nature Communications* 11, no. 1 (January 13, 2020): 1–10. https://doi.org/10.1038/s41467-019-14108-y.

Annex 1: Mapping the use of AI and data-driven technologies in the humanitarian and human rights space

No	Organisation	Focus Area / Work	Organisation Type
1	ACLED https://acleddata.com/	Data collection on violent conflict and protest	Not for profit
2	AI forensics https://aiforensics.org/	Algorithmic investigations and audits	Not for profit
3	Airwars https://airwars.org/	Conflict monitoring tool to assess civilian harm	Not for Profit
4	Bellingcat https://www.bellingcat.com/	Independent open-source investigators	Charitable org.
5	Center for Human Rights Science https://www.cmu.edu/chrs/ (within Carnegie Mellon University)	Scientific research and technical assistance for human rights documentation	Academic
6	Culture Pulse https://www.culturepulse.ai/	AI powered insights for better decision-making (e.g. insights on conflict)	For profit
7	Digital Forensic Research Lab (Atlantic Council) https://www.atlanticcouncil.org/programs/digital-forensic-research-lab/	Tracking disinformation, documenting human rights abuses, and building digital resilience	Think tank
8	Digital Sherlocks (from the Atlantic Council's DFRL) https://dfrlab.org/digital-sherlocks/	Training program on Disinformation 101 to Geolocation to Social media platform analysis and monitoring	Think tank
9	Engine Room https://www.theengineroom.org/	Research and technical support to human rights organisations (intermediary)	Research / tech (not for profit)
10	Forensic Architecture https://forensic-architecture.org/	Spatial analysis and digital modelling in investigating human rights violations	Research agency
11	HRDAG https://hrdag.org/	Human rights data analysis	Not for profit

No	Organisation	Focus Area / Work	Organisation Type
12	Human Rights Investigations Lab, University of Berkeley https://humanrights.berkeley.edu/programs/investigations-program/human-rights-investigations-lab/	Research and design of open-source tools for human rights monitoring	Academic + practice
13	KoboToolBox https://www.kobotoolbox.org/	Data collection tool for research and social good	Not for profit
14	Machine learning for Peace, University of Pennsylvania https://web.sas.upenn.edu/mlp-devlab/	ML tools for data tracking and forecasting major political events	Academia + practice
15	Mnemonic https://mnemonic.org/	Digital documentation of human rights violations	Not for profit
16	OSR4Rights (at University of Swansea) https://osr4rights.org/	Research and design of open-source tools for human rights monitoring	Academic + practice
17	Tella app https://tella-app.org/	Encrypted documentation app	Not for profit
18	TJET https://transitionaljusticedata.org/en/	Transitional justice-focused database	Academic
19	Ushahidi https://www.ushahidi.com/	Using citizen-generated data to develop solutions that strengthen their communities	Not for profit
20	Witness https://www.witness.org/	Video as a medium for human rights work	Not for Profit

Note: Generic, publicly available AI-powered services such as Google Translate, ChatGPT or organisational ones such as Microsoft Copilot are not included in this table. Country-specific archiving and documentation efforts e.g. the Azadi Archives and the Syrian Archives are similarly not included. The table is non-exhaustive.

Annex 2: Guiding questions

1. Which normative framework or institutional framework should we use to frame perspectives on the use and governance of AI for transitional justice (i.e. human rights; humanitarian law; ethics)? Is a clearer separation of these conceptual distinctions necessary or will silos not help?

2. What does it mean to take a victim-centred approach in the design of AI applications related to transitional justice?

3. How can we have inclusive and meaningful victim participation?

4. What pre-requisites (e.g. knowledge) should be in place when engaging with victims/ victims' groups on AI and transitional justice?

5. How do we make sure that victims are not overly burdened when engaging with such issues?

6. How should we measure and conduct the risk-benefit analysis of AI?

7. How can we ensure that AI applications developed for transitional justice adhere to the do no harm principle and avoid revictimizing victims?

8. How do we ensure accountability for victims and affected persons, in case things do go wrong?

9. How should we think through the potentially longer-term effects of AI, including at the societal level?

RAOUL
WALLENBERG
INSTITUTE

Co-funded by
the European Union

Global Initiative
Against Impunity
for International Crimes
and Serious Human Rights Violations

Ministry of Foreign Affairs of the
Netherlands