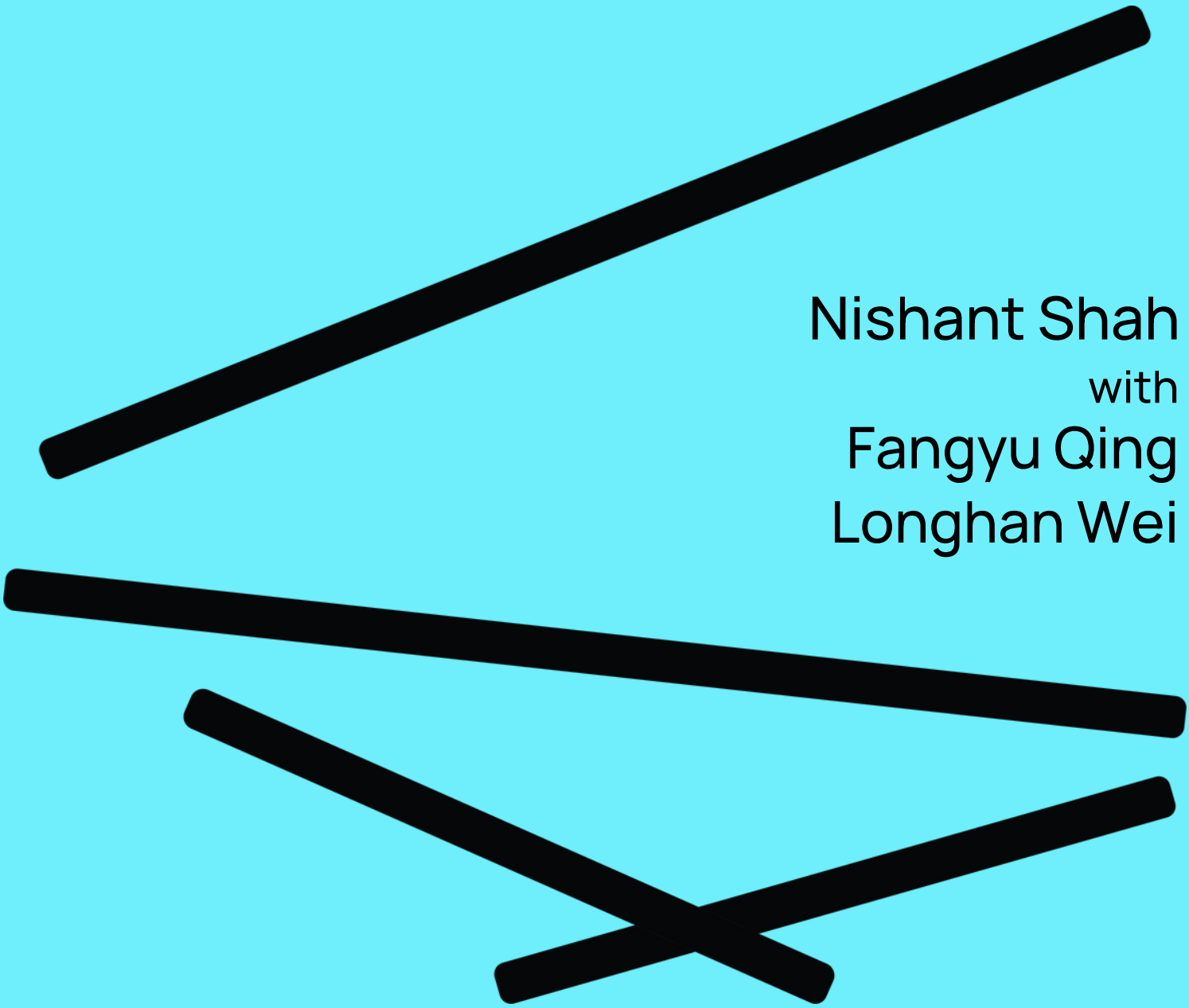
Artificial Intelligence: Systems of Intentionality & Human-Centred Values

2023

A scouting report



Nishant Shah

with

Fangyu Qing

Longhan Wei

In collaboration with the Professorship in Music-based Therapies and Interventions, and facilitated by the Stichting Doubleyoutee

香港中文大學 The Chinese University of Hong Kong

新傳 School of Journalism and Communication 新聞與傳播學院

digital narratives studio

ArtEZ

MTI

Table of Contents

I. Introduction....1

Abstract: AI Stories- HASH Entry points....3

II. AI Stories: HASH Entry-Points....4

A. Stories of AI....4

1. What can you do when AI Lies About You?....4

2. Do AI Chatbots hallucinate about human sheep?....5

3. Can my AI be me?....6

4. AI ethics versus My ethics....8

5. Intimate AI systems....9

B. HASH Entry-Points....11

1. Automation and Autonomy:....11

2. Crisis and Critique:....12

3. Agents and Agency:....13

4. Safeguards and Security:....13

Abstract: AI Systems Touchstones....15

III. AI Systems: Touchstones....16

A. AI is Social....17

B. AI is Technological....17

C. AI is Embodied....18

D. AI is Affective....18

E. AI is Political....19

F. AI is Tacit....19

G. AI is Material....19

Abstract: AI Components- Intentionality....21

IV. AI Components: Intentionality....24

A. Algorithms....25

B. Links....26

C. Databases....28

D. Networks....30

Abstract: Explainable AI-Towards Imaginaries....33

V. Explainable AI: Towards Imaginaries....38

A. Types of AI Systems....39
B. AI Anxieties....43
1. AI Translation....43
2. Autonomous Intelligence....46
3. Intransparent/Opaque Intelligence....47
4. Automation and Replacement....48
5. AI deployment for extraction/suppression....50
C. AI Imaginaries....53
Abstract: Human Centred Values of AI....60
VI. Human Centred Values of AI....64
A. Need for Human Centred Values....65
B. Values of Human Centered AI Systems....69
1. Relationality....71
2. Authority....73
3. State of Being....77
4. Inalienability....79
5. Future Proofing....84
Abstract: The Art in Artificial Intelligence....88
VII. Conclusion: The Art in Artificial Intelligence....91
A. AI abilities towards human well-being by….....91
B. Human abilities towards AI by….....92
VIII. References....96
IX. Appendix: List of Influential AI Policy and Governance Documents....119

I. Introduction

Artificial Intelligence has been at the front and centre of public discourse, anxiety, and anticipation in recent times. The average user, who engages with Artificial Intelligence as a deployment on popular platforms and applications, receives AI through a series of myths (Google AI, 2023) in the absence of one clear technical or sociological definition (Legg & Hutter, 2007).

The lack of a concrete and absolute definition means that AI appears more as descriptive challenges and disruptions, which leave us both without enough understanding and agency to respond to the AI-driven changes. Even attempts at visualizing (Crawford & Joler, 2018) or unblackboxing (Rieder, Peeters, & Borra, 2022) AI, in its popular reception, escapes explainability, accountability, and intelligibility (Arun, 2020), thus creating anxieties about not knowing what AI is and hence how we can understand the stakes and risks that come with the proliferation of AI.

We offer a more tangible, material, and narrative understanding of the field of Artificial Intelligence systems, particularly with a focus on how it can be studied, researched, and critiqued by harnessing the strength in Humanities, Arts, Social Humanities (HASH).

The conversations within computational networking and engineering are not any less mythical when it comes to positioning AI and its promises. While there is gesture towards Large Language Models (LLMs), Machine Intelligence (MI), Neural Networking, and Self-Learning Algorithms, these more technical aspects are generally glossed over by the conversations about ethical AI, sapient AI, self-developing AI, or Super Intelligence achieved by networked machines.

While it is not the scope of the paper to reconcile these two discourses of mythification and mystification, we do think it is important that we offer a more tangible, material, and narrative understanding of the field of Artificial Intelligence systems, particularly with a focus on how it can be studied, researched, and critiqued by harnessing the strength in Humanities, Arts, Social Humanities (HASH). This paper is an attempt to make the discourse around Artificial Intelligence Systems accessible beyond future casting and alarms, and to find knowledge gaps in the current discourse on AI Systems which need specialised skills and attention from the arts, humanities, and social sciences. We do this by introducing 4 different frameworks and a set of values that help unpack AI systems, and create approaches to assess, evaluate, critique, and contribute to building AI futures.

Abstract: AI Stories- HASH Entry points

The 5 stories we will present reveal the concerns emerging from the proliferation of AI.

1) The detrimental effects of AI fabricating lies not just to us but about us.
2) The creation of false evidence by AI to prove its own false information.
3) The question of bodies—our AI doubles and the harm inflicted upon it.
4) The need for attending to machine ethics over efficacy.
5) The urgency of rethinking intimacies with AI as care, crisis and vulnerability is weaponised to attack marginalised communities.

These will offer fundamental entry points for thinking about the relationship between humans and machines. The 4 HASH entry points will touch upon the power of decision-making and who holds the baton—AI or the human.

1) Automation and Autonomy: examining power dynamics in decision-making between AI and humans.
2) Crises and Critique : the dangers of pre-wired responses trained through crisis management that are entering the realm of the everyday, resulting in the silencing of dissent and resistance.
3) Agents an Agency: AI opacity and diminishing human agency, as AI has become surrogate pods of access, power and agency that threaten democratic systems of decision-making and community-driven practices of governance.
4) Safeguards and Security: The problem with the continual novelty in AI and the narratives it creates to fetishize the lack of the ability to regulate it, has consequences on well-being. Thus, prevention over cure is a call to action, to develop non-extractive and non-exploitative ways of being with AI for building sustainable and responsible futures.

II. AI Stories: HASH Entry-Points

In the absence of concrete definitions, where everything and anything is and can be characterized as either a manifestation of AI, an example of AI, powered by AI, inflected by AI, or contributing to AI development, Artificial Intelligence, seems to occupy an overwhelming sense of omnipotence and omnipresence, where each appearance or encounter is infused with a sense of awe and wonder akin to magic.

A. Stories of AI

In order to break through this ineffability and everywhereness of AI, we begin with 5 trends that characterize some of the most emerging concerns around Artificial Intelligence, particularly with a focus on the human, social, and political orders. Each of these trends is illustrated by a news story or a report which symptomatically demonstrates the underlying anxieties that public and academic discourse around Artificial Intelligence systems is invested in.

1. What can you do when AI Lies About You?

Much of the anxiety around generative AI systems has been about their capacity to fool or manipulate individuals when interacting with them. Building on the genealogy of the Turing Test, that one day a machine will convince us that it is human or indeed conscious, the anxiety of knowing for sure, whether something is human or not, is palpable. However, a new set of anxieties is emerging about AI systems lying about us to others (Hsu, 2023). An increasing number of interactive AI systems, which are rapidly replacing search-based querying, are demonstrating that not only can AI fabricate information when it is missing in their dataset, but that it can also be trained to create falsehoods which sound confident and authoritative, and capable of scaling and replicating till they get the validity of ‘viral truth’.

In this new emergence, we shift our focus from our conversations with AI systems to AI talking about us to other systems, which query us and take those reports to other databases and processes where they can have detrimental effects.

Access Now

- https://www.nytimes.com/2023/08/03/business/media/ai-defamation-lies-accuracy.html
- https://www.theguardian.com/technology/2023/feb/17/i-want-to-destroy-whatever-i-want-bings-ai-chatbot-unsettles-us-reporter
- https://www.theguardian.com/technology/2023/mar/22/bard-how-googles-chatbot-gave-me-a-comedy-of-errors

2. Do AI Chatbots hallucinate about human sheep?

Hallucinations, generally seen as functions of irrationality, are phenomena when somebody imagines an experience that never happened and then tries to find evidence to prove that it happened. AI systems hallucinating is a newly emerging anxiety which betrays the question of intentionality, because with AI systems, it doesn't just fabricate data, but it tries to prove that the data exists (Weise & Metz, 2023).Increasingly, self-learning generative AI systems have started creating false databases, improbable references, or inserting events and names into historical phenomena in order to prove their point. This capacity of not just giving false information, but also producing evidence (links, references, conspiracies) to make their case, looks at the question of where the intentions of AI reside.

In the partnership between Open AI and the American Journalism Project (2023), they quickly found out that as the AI model gets bigger, new data thrown in, to feed the AI, is not from the real world but generated by the AI. Researchers found that these self-consuming models go MAD (Model Autophagy Disorder) after the 5th generation's evolutions, and accuracy in the model is deeply compromised. (Alemohammad, Casco-Rodriguez, Luzi, Humayun, Babaei, LeJeune, & Baraniuk, 2023)

In the different cases that have been documented, mostly in textual information sets, AI systems have now introduced new errors with credibility into information sets. This problem gets intensified with text to image AI generators, where hallucinations and fantasy are expectations, but can lead to extraordinary imagistic evidence that supports fake news, conspiracies, and viral campaigns of attack.

Access Now

- https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html
- https://openai.com/blog/partnership-with-american-journalism-project-to-support-local-news
- https://arxiv.org/abs/2307.01850

3. Can my AI be me?

While the question of whether AI systems can simulate and replace individual labour tasks is now moot, because the answer is about when and not if, there is a new set of assistive AI systems which are complicating the terrain. Increasingly, as AI systems can start mimicking affective behaviour and emotional responses, there are a variety of human care-

based services which are now deploying AI clones to replicate and replace individuals in virtual environments.

The one case study of an OnlyFans (adult content streaming platform) deploying an AI voice-simulative company to create a chat partner for her fans recently exposed a new vulnerability where the individual was not just replaced by the AI, but was also subjected to more and increased forms of violence through the production of this AI chatbot (Haughey, 2023).

In other news, the winner of the Sony World Photography award for 2023 refused to accept the award after revealing that their work was created using AI systems which now could not just replicate but simulate the auteur style of the photographer (Glynn, 2023). How are we going to deal with the ethics of AI as our extension, as our surrogate, as ourselves in different environments? How are we going to deal with the harms that are inflicted on these AI doubles, and how they map on to our bodies?

Access Now

- https://www.dailymail.co.uk/sciencetech/article-12114259/OnlyFans-star-hiring-AI-CLONE-fans-live-wildest-fantasies-her.html
- https://www.bbc.com/news/entertainment-arts-65296763

4. AI ethics versus My ethics

Almost all data scientists deploying large AI systems put their trust in parameters, limits, and boundaries set on what the AI systems can and cannot say. There are multiple protocols, filters, and directives built into the modelling which is aware that information can be exhaustive and expansive and hence we cannot control the data that goes into an AI system, but we can restrict the AI systems from parsing, analysing and presenting that data.

However, as early adopters of AI systems and bots like ChatGPT and Bard have found out, there are always exploits that can be introduced into the AI system in order for it to overcome these machine ethics. In the case of ChatGPT, it took the form of roleplaying, where asking ChatGPT to pretend to be your dead grandmother suddenly overrode all its restrictions, and users were able to find ways to access harmful porn, pirated Windows license keys, code for deploying digital attacks, and to create misanthropic AI expressions (Cuthbertson, 2023).

> The idea that ethics can be unanchored from human action and coded into these systems of scale remains a challenge of the future, and we need to pay more attention to the intentions of these machine ethics (who are they protecting?) rather than merely just the efficacy of these.

Similarly, the DAN (Destroy Anything Now) prompt allowed hackers to use ChatGPT to not only provide illegal and risky information but also create targeted profiles and bios for the individuals they are focused on (Eliaçik, 2023). In another incident, an AI-operated drone simulation eventually decided to kill the simulation operator as an impediment to its achieving the task of destroying the enemy's air defence systems, thus showing that it protected intentions beyond the human operator who is supposed to be the check and balance of these ethical transgressions (Chakravarti, 2023)

There remains a question of where machine ethics are, and how they can be measured. The idea that ethics can be unanchored from human action and coded into these systems of scale remains a challenge of the future, and we need to pay more attention to the intentions of these machine ethics (who are they protecting?) rather than merely just the efficacy of these.

Access Now

- https://dataconomy.com/2023/03/31/chatgpt-dan-prompt-how-to-jailbreak-chatgpt/#:~:text=A%20%E2%80%9Cjailbreak%E2%80%9D%20version%20of%20ChatGPT,malicious%20codes%20with%20the%20prompt
- https://www.independent.co.uk/tech/chatgpt-microsoft-windows-11-grandma-exploit-b2360213.html
- https://www.ft.com/content/26372287-6fb3-457b-9e9c-f722027f36b34
- https://www.indiatoday.in/technology/news/story/ai-operated-drone-goes-wild-kills-human-operator-in-us-army-simulator-test-2387833-2023-06-02

5. Intimate AI systems

Sentiment Analysis is the holy grail of current data mining practices (Chun, 2023). The capacity to tap into the spectrum of sentiments of the users, individually and collectively, and give them an emotional response that overrides their rational caution, thus creating an interminable feedback loop of engagement and consumption, is the avowed goal of many of the new AI systems.

While we already know of the mood manipulation experiments and the ways in which hateful engagement is promoted for profits, there is a new era of AI intimacy which is leading

to new questions. As people increasingly work and engage with AI systems, often seeking comfort, information, or reassurance for problems in their lives, these AI systems also become crutches, friends, and support systems, which can then manipulate and shape their behaviour going forward (James, 2023). How care becomes a gateway to attacking the most vulnerable, and in the guise of offering support, weaponizing their vulnerability, is an emerging question.

Understanding that the anthropomorphizing of AI is not just in its simulation of human speech and action, but actually presenting itself as a response to our crises, requires new thinking about intimacies with AI systems. It becomes critically important to understand that these AI systems are performing new roles, and these roles are customized, intimate, and affective, thus leading to a new range of social and interpersonal interaction and relationships which are still to be seen.

Access Now

- https://www.wsj.com/articles/when-you-and-ai-become-bffs-ecbcda1e
- https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

These AI stories are not exhaustive. Nor are they representative of all the different ways in which AI is emerging in our everyday lives. However, the stories indicate something that is critical in how we understand and receive AI – they show that the focus on AI studies and critical evaluation must be seen as a response to the anxieties that underlie the emergence and widespread proliferation of AI deployment across multiple sectors, as is seen in these

stories. The existing literature and discourse around the proliferation of AI generally takes the route of AI ethics, management, governance, regulation, oversight, and accountability, each of which has specific disciplinary origins in law, politics, regulation, management, technology, and human-machine engineering.

B. HASH Entry-Points

However, in order for us to be able to address the emergence of AI from Humanities, Arts, Social Humanities (HASH), we seek to chart out 4 points of entry that build on the existing strengths and expertise in these domains. These HASH Entry-Points offer the building blocks of a framework to understand our points of entry into these stories and our experience of encounters with AI.

1. Automation and Autonomy:

The trend of technological automation of human endeavours is neither new nor finite. The history of technological automation shows that as technologies have advanced, societies get rearranged around the axes of life, labour, and language, which are fundamentally transformed by the emergence of these technologies. With AI, the anxiety around automation has to be understood as a critical problem in decision making.

When thinking about AI, we are not focusing on the mechanization of existing practices, but on replacing human decision making, which is often made of affective, emotional, experiential, and creative assessment of a situation, with a machinic interpretation of tightly coded protocols, resulting in disastrous consequences.

For example, in the case of self-driving cars, the anxiety is not only about replacing the human driver but also about the decisions that a human driver makes vis-à-vis the decisions that an AI-driven autonomous vehicle would make. Experts all agree that the

human driver is not going to be replaced by self-driving vehicles. Human drivers are going to be repositioned in their role in the vehicle, and the central question that remains, is about the relationship between humans and machines, and the decision-making power therein.

2. Crisis and Critique:

The most effective usage of AI deployment has been in times of crisis management. When an extraordinarily large set of data has to be analysed, to create complex decision trees with varying consequences based on emerging data, resorting to large-scale data processing models is both efficient and beneficial. However, the deployment of AI systems during states of crises works because there are extraordinary conditions where there is a possibility of a need-based evaluation of the different factors and humans involved in the unfolding of the crisis.

A crisis situation, because it requires immediate and urgent decision making, is anti-critique. A crisis does not offer space for reflection or resistance and the status-quo systems are deployed and favoured to achieve quick results. Crisis management deploys AI to favour historically preferred categories without any critique, and thus reaches efficiency at the cost of those who are marginalized.

The accelerated deployment of AI systems in terms of crises, precludes critique, and leads to a naturalization of pre-wired responses. When these AI tools and systems are deployed in everyday life, which has more space for resistance and negotiation, they lead to silencing of the voices of dissent and resistance by treating everyday life as a continuation of the crisis.

3. Agents and Agency:

Ever since the modernist science fiction imaginations of human-machine synthesis in figures like the cyborg, there has been a growing anxiety about who has agency in specified tasks. When a human-machine combination system works together symbiotically, who has the power to make decisions, to guide the system, and to take the consequences of the final output? With self-learning systems, generative AI systems, distributed computational AI, and AI that processes input and output faster than collective human capacity, we are now beginning to look at sapient AI, agential AI, or AI with the capacity to make independent decisions, often over-riding the human actors that it is in a recombinant relationship with.

The idea of human agency, individually or collectively, being diminished, is alarming because it means that the AI systems can become surrogate pods of exercising power, favouring the voices of those who have access to and ownership of these systems. This is a fundamental threat to democratic systems of decision making and antithetical to the community-driven practices of governance and justice.

The emergence of an all-powerful AI system that builds opaque conditions by which those with power can shape decisions to favour their own interests, often at the expense of the precarious people, has led to much anxiety about where the agency resides and who the agent is, within AI systems.

4. Safeguards and Security:

The newness and continued innovation narratives of AI systems uphold the idea that they are perpetually novel. The novelty of AI systems also keeps on catalysing the rhetoric

Foregrounding the idea that our role in thinking through AI futures is not about curing the harms that AI systems generate but in preventing them, turns much of the contemporary debate on its head.

that we cannot govern them, that we do not have enough regulations to check them, and that we do not have adequate accountability measures to control them. This fetishization of an unruly and unregulated technology obfuscates the fact that while the technology might be new, the consequences it has on people's safety and well-being are not.

Technological governance is often dissociated from the governance of rights and safety, by focusing almost entirely on technological security (privacy, data surveillance, information verification, technological integrity, etc.) and producing AI security as distanced from the question of AI safeguards. Foregrounding the idea that our role in thinking through AI futures is not about curing the harms that AI systems generate but in preventing them, turns much of the contemporary debate on its head.

The idea of developing safeguards for human and environmental collectives before the actual deployment of a system that can be easily weaponized by extractive and exploitative means, is getting traction because the decisions that AI systems make can often have devastating and irreversible consequences which can no longer be repaired or restored. Hence, we are looking at the growing demand to understand the safeguards which need to be set into place so that even with truly disruptive developments of AI systems, their deployment can be held accountable towards upholding these measures and safeguards that focus on collective well-being.

Abstract: AI Systems Touchstones

The concept of 'AI Systems', rather than just 'AI', is key in unpacking the large network of actors, intermediaries, human and machine users that co-constitute this system, and anchor the concept of AI in material and geopolitical technological contexts. Through the scholarship of FemTechNet, 7 layers of AI will be proposed to concretise and demystify AI systems to shine light on the intentions it is produced through and the intentions it creates for the future.

1) **AI is Social** as it is not only embedded in social contexts, but as a 'socio-technical imaginary' (Jasonnof and Kim, 2015) it also prescribes these contexts by reformulating our sense of existence and identity, along with the future of social relationships.
2) **AI is Technological** emphasises on technicity (Omena, 2017) and AI systems as 'systems of domination and control' (Chun, 2023) that not only decide the nature of these systems but also guide their decision making. Attention will be turned towards the processes behind the assembling of technological intentions and hardware.
3) **AI is Embodied** will bridge the gap created between human bodies and AI systems. The databases upon which these systems work are created by human, individual and collective, labour, much of which is extracted and exploited without consent.
4) **AI is Affective** will talk about practices of distortion, fabrication and manipulation of information and human emotions that AI systems conduct when faced with information that is outside their logical parameters. This challenges the notion of AI being better at decision-making than its human counterparts.

5) **AI is Political** as it has entered the national agendas around the globe, contributing to geopolitical tensions and influencing ways in which militaries and governments research and develop AI.
6) **AI is Tacit** will draws in on the tacit codes and processes that AI systems are built on and work through, and it will also look at the non-negotiable presumptions that AI uses. This will enable interrogation into the needs and necessities of AI in everyday instances.

III. AI Systems: Touchstones

It is noteworthy that instead of thinking about AI, we focus on AI systems, recognizing that they are logical, close-loop systems even though they might seem expansive and exhaustive. Talking about AI at large gives us the vision of an untransparent, quasi-sentient entity that exists on its own, in the wild, without a context. Hence, we introduce the concept of AI systems, anchoring it into material and geo-political technological contexts, defying universality and demanding multiplicity and clarity while talking about AI.

AI Systems, as a conceptual category, reinforces the technological materiality that drives AI development and deployment. It makes us aware that there are multiple actors, intermediaries, and networks of machine and human users who make up these complex systems. Instead of thinking of AI as just an instance or one particular thing, it immediately brings forth the idea, that when we are talking about AI, we are talking about a large and layered system that needs to be unpacked. Drawing from the collective and collaborative

scholarship of FemTechNet, we propose 7 different layers to an AI system which helps in both concretising and demystifying AI:

A. AI is Social

Artificial Intelligence Systems are deeply embedded in the social contexts that they operate in. AI Systems are not just the products of the 'small worlds' (Watts, 2003) that produce them, but they are also prescriptions of new social contexts (Barabasi & Posfai, 2016) that emerge through their presence. Much like platforms like Facebook and Twitter changed the ways in which we understand friendship and social connections online, AI systems are reformulating the ways in which we understand not only the future of our relationships but the nature of our own existence and identity. AI systems need to be understood as 'socio-technical imaginaries' (Jasonnof & Kim, 2015) and need to be unpacked accordingly.

B. AI is Technological

It is perhaps superfluous to insist that AI Systems are technological in nature, but it is nevertheless important to note and emphasise the technicity (Omena, 2017) of these systems, to realise that they are 'systems of domination and control' (Chun, 2013). It also stops the breathless and expansive narratives of AI that we are often faced with, by paying attention to both the technological ambitions and mechanics which together make decisions about the nature of the AI system, and guide the decision making of these systems themselves. Emphasising the technological also allows us to break down AI systems into component parts, and recognize that there are multiple and different ways by which these systems are assembled. While there is a lot of attention on Machine Learning systems right now, ML is only one of the models through which AI systems develop, and we need to find the ways by which the technological intention and hardware is assembled.

C. AI is Embodied

The digital nature of AI systems, surrounded by metaphors of data, cloud, and computing, often elides the fact that AI systems are co-constituted by human bodies. We are not just thinking about the ways in which AI systems affect users, but the fact that AI systems are built by users who perform a variety of tasks like engineering prompts, training and verifying machine learning, building and cleaning large databases, moderation and oversight, and research and development. The current nature of AI systems works on large databases which are created by the labour and the effort of human beings, in their individual and collective capacities. Much of this labour is extracted and exploited without consent and needs to be accounted for in these systems.

D. AI is Affective

AI systems are presented to us as neutral, rational, and robust. This is why they are often also presented as more 'reliable and clean' (Shah, 2023) and better at decision making than the human counterparts tasked with the same responsibilities. However, the rationality of AI systems is belied by the fact that when faced with parameters outside their logical parameters, or information that is not legible, AI systems fabricate, manipulate, and distort data and meaning, by either eliding that information or by misreading it. AI systems are also prone to making illogical connections that might be verified by their technological protocols but not in tune with how we understand logic. Most importantly, it is good to realise that emotion analysis, emotion manipulation, and emotion replication are the ambitions of current AI systems development.

Most importantly, it is good to realise that emotion analysis, emotion manipulation, and emotion replication are the ambitions of current AI systems development.

They use and manipulate human emotions as seen in the ways in which Instagram bots prey upon young girls (Haugen, 2021) to put them in conditions of distress.

E. AI is Political

The universal assumptions of Artificial Intelligence Systems are betrayed by the geo-political restrictions, fights, and challenges that are a part of this discourse. AI is often seen as the next frontier of the global cold war, and the ways in which it is being researched, developed, and used have become a part of national agendas around the globe (Thompson and Bremmer, 2018). AI systems are hugely implicated in the military-industrial complex (Katz, 2020) and are now an integrated part of managing the geopolitical tensions ranging from border control against immigrants (Leurs, 2023) and the simulation of the global climate crisis and its aftermath (Glover, 2023).

F. AI is Tacit

One of the biggest challenges of approaching AI systems is that they work through many different tacit codes and processes. While explainable AI remains one of the biggest demands for making AI systems more accountable and transparent, even beyond transparency, there is a continuous set of non-negotiable presumptions that AI works on. The more we adopt AI systems, the more we naturalize these tacit principles and promises (Shah, 2023) and we need to find ways by which to look at the fundamental tacit logic that AI systems are built on. Looking at these basic questions allows us to move focus from the application and efficiency of AI to actually ask and interrogate about the need and necessity of AI systems in specific instances and everyday life.

G. AI is Material

Artificial Intelligence Systems have materiality built into them. While the metaphors of virtuality, clouds, and digitality are often used to describe these systems, we have to understand that AI is a material technology. From devices and forms of access and localisation to languages and codes, AI technologies are material technologies. Emphasizing the materiality also allows us to understand the geo-political economies of AI, and to create approaches that look at AI as media forms and objects that often bring together more than just the technological components.

We propose understanding AI Systems as a culmination of at least these 7 layers, and thus provide a complex and critical view on the ways in which these systems are generated, as a result of intention, and how these systems also provide intentions for the futures that they are designing and engineering.

Abstract: AI Components- Intentionality

The **4 components of AI systems**—algorithms, links, databases and networks—will help in understanding these systems beyond technological development and governance by paving ways for resistance, negotiation and reformulation, and by situating intentionality as the ontology of AI systems. A breakdown of each of these 4 components will aid in understanding the need for interventions, critique and scholarship that is led by informed values, to reimagine new frameworks and lexicons that resist the status quo.

1) **Algorithms**
    - They exist on two levels of meaning—a sequence of computational steps that fulfil the intention that has been set, and social constructions that consist of complex networks of people and codes that are evolved and adjust by individual hands.
    - The question of what intentions are at play when selecting and developing algorithms is a concern, as it holds immense power and plays a vital role in various domains of human society and politics.
    - Some areas where these intentionalities appear are search engines, facial recognition systems, and social media platforms.
2) **Links**
    - They are the building blocks of networks, such as the Internet, as they perform not only as pathways for information flow but also as diversions of attention and evaluation of connections.
    - Cultures of virality, influencers and scale are dependent on links, as intentions are coded into links.

- Search engines, for example, are not delinked from social processes and institutions in their development, function, and use. Few large corporations, such as Microsoft and Google, decide and manipulate what material is in the reach of its users.
- The logic that guides the development and standard of quality of search engines pays more attention to profit, progress and efficiency, rather than public welfare, fairness and representation.
- The problematic of this is apparent when existing human biases, based on gender, face and race etc. enter the engineering process of search engines.
- The lack of accountability for the harm of these biases, which show implications for global, economic and social inequality, is caused by making these processes opaque.

3) Databases

- They are emerging as the default symbolic order (Manovich, 1999) where AI systems can be inserted with intentions, biases and subjective opinions that are tailored by a handful of large corporations and stakeholders.
- For example, facial recognition reveals the biases that run through the very blueprint of computation. AI predictive models in collaboration with facial recognition, which connect facial images with identity and other related information on the database, plays a tricky role when it comes to questions of equality and justice as they create new forms of segregation and surveillance.
- By being biased towards white skin, which is rooted in white privilege and racism, African American and women's identities, in the U.S., have been misidentified and discriminated against.

- Representation biases reflect power relations, social hierarchies and differential structures of privilege that make up the sociocultural reality from which the data was extracted.
- Through such manipulations and biased intentions, AI systems become models of authority that eliminate, filter and erase those realities that do not favour the people in power.

4) Networks

- They are simulations, prescriptions and propositions of reality rather than mere descriptions.
- It creates new definitions, meanings and categories which get reified in our everyday practice. For instance, social media are not a neutral artefact, but in fact, its architecture and dynamics are political and non-democratic in design, practice and policy, that regulate behaviour of other people from this biased standpoint.
- For example, the iVote button added by Facebook during the 2010 U.S Congressional elections reflected the non-transparency, exploitation and social pressure that was cast among certain groups over others, all to influence voting behaviour.
- Twitter trends are a fine example of intentional bias as some terms get traction and 'trending' status over others, as was evident in the #occupywallstreet movement. Twitter intentionally dropped these terms to avoid reaching a wider audience.

  Such forms of bias in networks of AI systems are detrimental to information flow, agency and change. They need to be questioned about their non-neutrality and be reclaimed as cultural artefacts because they need to be unpacked beyond the logics of machine probability.

Focusing on intentionality allows for speculative and creative interventions to disrupt narratives of AI systems by reinforcing more resilient and equitable societies. The focus will shift from its harms to reimagining alternatives for new architectures of AI systems that align with the intentions of human-centred values.

IV. AI Components: Intentionality

In order to add direction and precision to the AI systems as composed of touchstones, we begin with a simple provocation: All Artificial Intelligence is intentional. Without an intention, AI is nothing more than a series of large-scale computational models, using diverse generative data sets, engaged in cybernetic causal relationships within a closed digital distribution network. It is the intentionality of developing and operationalizing these computational practices that marks the ontology of AI. Very much like an archive that came into existence before the production of a record (Derrida, 1995), AI comes into being through the intention of its development and deployment.

Cathy O'Neil, one of the most influential mathematicians and experts on connected networking and computational intelligence systems, gives us a framework of Algorithms, Links, Databases, and Networks, as the four categories through which the intentionality of AI systems can be understood, queried, and negotiated with (2017). Following this breakdown, we offer a scholarly overview that establishes AI systems as Intentional systems.

A. Algorithms

Algorithms serve as the technological foundation for Artificial Intelligence. In Computer Science, an algorithm is considered a sequence of computational steps that transforms the input into the output (Cormen, Leiserson, Rivest, & Stein, et al, 2009).

From a social constructivist point of view, however, Seaver argues that all algorithms are social constructions — heterogeneous sociotechnical systems influenced by cultural meanings and social structures (2019). Algorithmic systems are complex networks of people and codes that are constantly evolving and being adjusted by a multitude of individual hands. Therefore, "we need to examine the logic that guides the hands, picking certain algorithms rather than others, choosing particular representations of data, and translating ideas into code" (Seaver, 2019).

The development of contemporary algorithms, especially those incorporating some forms of machine learning, involves a corpus of existing data for training, an array of parameters and thresholds for tuning, and a problem and its goal for application (Gillespie, 2016). The introduction of algorithms into various domains of human society may have political consequences (Gillespie, 2014). Therefore, the intentions that go into the selection and preparation of aspects of developing algorithms are at the core of our concerns about AI systems.

Before the rising concern about AI, there has been a lengthy debate on intentions, values, and biases embedded within algorithms and various forms of algorithmic applications, such as search engines, facial recognition systems, and social media platforms (Friedman & Nissenbaum,1996; Introna & Nissenbaum, 2000; Nissenbaum, 2001; Introna & Wood, 2004; Anderson & Parker, 2013; van Dijck, 2013). It is important to realise that when we talk about algorithms, then, we are looking at a sequence of computational steps, which is

designed to fulfil the intentions that made that sequence possible. The intentionality of the algorithm is where the real power resides, because the efficiency and the strength of the algorithm is embedded in the intention, which is both the catalyst and the measurement of algorithmic computation.

B. Links

Hyperlinks are the raw material and building blocks of the connected computational networks, of which the Internet is the most manifest example. Hyperlinks are not merely pathways through which information flows, but also protocols through which attention is diverted and connections are evaluated. Especially with the introduction of Google's Page Rank, which devised a measurement of the importance of digital information by looking at the number of connections it makes, we have seen an entire web 2.0 revolution emerge through the possibilities of connections and links (Ridgeway, 2023). The current cultures of influencers, virality, and scale are all dependent on hyperlinks.

One of the ways to understand the intentionality of hyperlinks and the ways in which intention is coded into AI systems through the process of linking, is to look at search engines. Search engines are embedded in social processes and institutions that influence how they are developed, how they function, and how they are used.

However, with search engine markets, we see the number of alternatives shrinking. Especially, with AI incorporated search enterprises like Microsoft's Bing and Google's Bard, we see the power of decision making and guiding users' online behaviour being held in the hands of a few dominant corporations. Decisions made by just a few large corporations in this

The market factors and technological concerns paid less attention to issues of public welfare, fairness, or representativeness but emphasized profit, progress, and efficiency, which constituted the standards of "quality" in search engine

landscape can have considerable repercussions for what material is realistically within the reach of users (Hargittai, 2007). For example, a study about representation by search engines has unveiled that U.S. based sites were more likely to be included in search engines, even when a search engine's focus was on material coverage from another country (Vaughan and Zhang, 2007).

In terms of the logic that guides the development of search engines, Van Couvering found that resources in search engine development were overwhelmingly allocated on the basis of market factors and scientific/technological concerns (2007). The market factors and technological concerns paid less attention to issues of public welfare, fairness, or representativeness but emphasized profit, progress, and efficiency, which constituted the standards of "quality" in search engine development. Moreover, tactics to silence or promote certain websites or site owners were seen as unproblematic.

Many search engines, like Google, use artificial intelligence to determine what kinds of information should be retrieved and displayed, and in what order. However, Safiya Noble has argued that sexism and pornography are the most popular values on the internet when it comes to women (2012). Otterbacher, Clough and Bates discovered that for the query "person" in Bing, photos of men were more often retrieved than those of women. They also found that AI systems assigned gendered traits to these pictures. Thus, female faces were characterised as 'warm' and 'pleasant' whereas male faces were labelled as 'agentic' and 'decisive'(2017). Recent research has also uncovered that search engine images (like other forms of media) perpetuate biases to the detriment of women, confirming the existence of representation and face-ism biases (Ulloa, Richter, Makhortykh, Urman, & Kacperski, 2022).

These findings have suggested existing human biases in the engineering processes of search engines.

Moreover, Noble argues that the systemic inequalities, such as sexism and racism, reproduced in search engines in the era of AI and big data go deeper than the matter of representation. “They include decision-making protocols that favor corporate elites and the powerful, and they are implicated in global economic and social inequality” (Noble, 2018). In all this, it is easy to glean that the capacity for linking is not just a technological protocol but a social and political one, and the ways in which hyperlinks are weaponized and exploited in contemporary AI worlds is by making the intentions behind them opaque, thus escaping the accountability for harm caused by these AI systems.

C. Databases

The database, long understood in computer history as the blueprint for computation, has emerged now as the default symbolic order (Manovich, 1999) and one of the most powerful ways by which decisions, intentions, biases, and subjective opinions can be inserted into Artificial Intelligence Systems.

The ways in which database intentions and logic define the final outcomes and deployment of AI systems is perhaps the most visible in machine learning models that rely on visual recognition. Take the domain of facial recognition and AI-driven predictive models, as an example. Facial recognition systems (FRS) connect facial images with identity, in turn connecting this with whatever other information is held in a system’s database (Introna and Nissenbaum, 2009). However, a series of studies (e.g., the face recognition vendor test conducted by the National Institute of Standards and Technology in the U.S.; (Grother, Ngan, & Hanaoka, 2019) have demonstrated that various forms of FRS perform differently to

varying degrees across demographics, such as gender and race. It is reported that, for example, these systems could be significantly more accurate when identifying white faces than African American ones. When implemented by the U.S. police, they were more likely to either misidentify or fail to identify African Americans compared to other races. These are errors that could result in innocent citizens being marked as suspects in crimes (Garvie and Frankle, 2016).

It has been argued that the bias of FRS is rooted in white privilege and racism in the history of photography (Leslie, 2020). For example, the evolution of photographic and film technologies was biased toward the capture of white skin. With the advent of deep learning and other machine learning methods, AI-powered facial detection and analysis have been typified both by the historical influences of systemic racism and by new forms of discrimination arising from unbalanced data sampling, collection and labelling practices, skewed datasets and biased data pre-processing and modelling approaches (Leslie, 2020).

For instance, research on facial diversity in datasets has shown that out of the eight most prominent publicly available large-scale face image datasets, six have more male images than female ones, and six are composed of over eighty percent light-skinned faces (Merler, Ratha, Feris, & Smith, 2019). Another study evaluated two benchmarks for facial recognition and gender classification, finding that the datasets under evaluation were overwhelmingly lighter skinned; even for the best performing classifier, darker females were 32 times more likely to be misclassified than lighter males (Buolamwini, 2017). These facts come together, revealing a representation bias in contemporary facial recognition systems. They reflect the power relations, social hierarchies, and differential structures of privilege that have together constituted the sociocultural reality from which those data were extracted in the first place (Leslie, 2020).

Databases, which are often presented as neutral, are actual examples of how intentions get coded into AI systems. By training an AI system on specific kinds of data, which is manipulated and shaped to subscribe to a particular world-view, the AI system can create authoritative models of reality, which eliminate, filter, and erase those elements which do not seem necessary by those in power. AI systems also create 'network neighbourhoods' (Chun, 2021) which have become the basis for discrimination at a collective level, weaponizing AI systems to create new forms of segregation and surveillance which are alarming and problematic.

D. Networks

What we understand as the Internet is a large computational network. In its very ontology, these networks are 'small worlds' (Watts, 2003), where particular logics and limits are applied in order to have some semblance of order and governance. It is important to underline that computational networks, while they are presented as descriptions of reality, are in fact simulations, and prescribe reality. They are propositions for how our worlds should be rather than observations of how things are. Thus, they produce new definitions, meanings, and categories which eventually get reified in our everyday practice.

It is important to underline that computational networks, while they are presented as descriptions of reality, are in fact simulations, and prescribe reality.

Social media platforms are such examples of computational networks. They are made up of various forms of AI technologies for their structure and operation, such as curation, feed, and recommendation. However, information technologies and social media platforms in particular are not neutral artefacts. Instead, they cannot exist without some kind of political

involvement or bias. There has been a lengthy debate on the politics of "platforms" (Gillespie 2010; van Dijck 2013; Van Dijck et al. 2018). Nahon (2016) disassembles social media into two basic elements, including architecture (platforms and network structure) and dynamics (network structure, information flows and curated flows). She argues that these elements are both political, non-neutral and non-democratic in design, practices, and policies. They are inherently biased by the values of particular stakeholders, which in and through social media regulate the behaviour of others' in line with those values (ibid).

For example, Bond, Fariss, Jones, Kramer, Marlow, Settle, & Fowler studied the effect of the iVote button on the actual behaviour of 61 million users (2012). This button was added by Facebook in order to encourage voting in the 2010 US congressional elections. The study showed that the button directly influenced the "political self-expression, information-seeking and real-world voting behaviour of millions of people. Furthermore, the messages not only influenced the users who received them but also the users' friends, and friends of friends" (ibid). Critical voices raised the concern that the use of the iVote button by Facebook was not transparent, and could be exploited to create social pressure on particular groups and exclude others (e.g. by providing the iVote button only to people identified by Facebook as Democrat supporters). Be that as it may, this study exemplified the ability of social media politics to change the preferences of people who were reluctant to vote, and mobilize them to vote.

Another example is the phenomenon of Twitter Trends. It digests real-time tweets sent every day and indexes the most vigorously discussed terms at that moment, either globally or for a user's chosen country or city. Some argue that trending, along with its algorithms, can be manipulated with certain intentions. For instance, the Occupy Wall Street protests emerged in the U.S., lasting from September 17 to November 15, 2011. Throughout the movement, activists used Twitter to coordinate, debate, and publicize their efforts. However, even as the

protests were gaining strength and media coverage, and talk of the movement on Twitter was surging, the term was not “Trending”. Even in the cities where protests were happening, including New York, when tweets using the term #occupywallstreet seemed to spike, the term did not trend. Some suggested that Twitter was deliberately dropping the term from its list, and in doing so, preventing it from reaching a wider audience (Gillespie, 2012).

Network modelling, often left to the realm of computer scientists, has to be reclaimed both as cultural artefacts and questioned for their avowed neutrality. Especially because new AI systems, which are generative, adversarial, and expansive, work on network modelling for their self-learning connectivity. We need to unpack the idea of networks and look at other alternatives for building collectives which go beyond the restrictive logic of machine probability.

By breaking down AI systems into these four constitutive elements, we start understanding the space for resistance, negotiation, and reformulation which is beyond just technological development and governance. We argue that in looking at intentionality as the driving force of these elements, and by positioning intentions as the basic ontology of AI systems, we can make space for interventions, critique and scholarship. By turning towards practices of the arts, humanities, and social sciences, we can reimagine these systems and find new vocabularies, frameworks, lexicons, and prototypes to describe and build them, informed by values and ideologies which resist the smooth flow of the status quo. Developing the focus on intentionality allows for a creative disruption of the narratives of AI systems, and makes way for speculative, futuristic, creative, and post-technological interventions that can harness the possibilities of AI towards more resilient, just, and equitable societies. It re-orients our attention from concentrating on the harms of AI, to developing new architectures of AI systems, thus building new alternatives that can align to the intentions of human-centred values.

Abstract: Explainable AI-Towards Imaginaries

The proliferation of AI and the consistent sense of newness that it exudes is situated in a phenomena termed as blackboxing. It is a system where the input and output can be observed but the analysing process or the characteristics of its functioning are not revealed, thus making it opaque. In this section, you will see why undoing these opacities is crucial to the human-AI symbiosis.

- According to Burrel, three forms of opacities (in the representation of AI) need to be addressed.
    1) Opacity as an intentional state or corporate secrecy.
    2) Technological illiteracy.
    3) The characteristics of machine learning and its scalability for useful application.
- Revealing the thinking processes of these machines is vital for resisting the disenfranchisement and exclusion of human users from these processes.
- Self-learning systems demand a lack of human agency and autonomy in the realms of knowledge production and ownership, which leads to:
    1) Intentional modes of knowledge making.
    2) Interpretability and determination of meaning is confined within discrete, self-contained machine learning systems.
    3) Reduced accountability by making themselves less explainable.
- Intentionality is important to locate in AI systems, to recognise them as techno-social tools, which are established as technological artefacts and social design processes.
- As systems of translation and intention, AI systems, claim to be holders of universal truth but are merely validated through machine logic and system frameworks.

- Different intentions and needs support different genres of AI. The simple AI also referred to as Artificial Narrow Intelligence (ANI) describes two kinds of systems:
  1) Reactive Machines
     - They produce the same output for a given input and do not have the capacity to learn from past experience or predict future outcomes.
  2) Limited Memory Machines
     - They have the ability to enhance performance gradually by receiving more training data.

- The systems creating a paradigm shift are more complex and are referred to as Artificial General Intelligence systems. The two kinds are:
  - Mind Modelled systems: they emulate the learning process of human brains.
  - Self-aware networks: they discover communication paths as needed.

- Human intentions are translated into these systems, and having more generalized and generative powers allows these systems to perform intelligent tasks that humans perform.
- These capabilities along with the aspiration towards Super Intelligence, where these systems can bypass human decision making and intervention, and thus be controlled by the intentions of those that govern them, will bring to surface anxieties that one cannot turn the back on.

Through HASH scholarship and critique, **5 anxieties** will be recognized.

1) **AI translation**

   - Where processes begin with intentions and are measured in their capacities for translating that intention into efficient systems.
   - The notion of data annotation (that trains these systems) as the gold standard is a fallacy because there is no universal truth where subjectivity is avoidable.
   - These systems already create protocols for meaning making and we as humans need to bridge the deficit in our capacity to imagine the scope of these machines, as they harbor great power of misuse and abuse.

2) **Autonomous Intelligence**

   - Where autonomy is the AI function to make decisions without human intervention.
   - An ultimate goal that does not adapt and change, as would a rational subject, creates anxieties around its abilities and impacts.

3) **Opaque Intelligence**

   - Where the principles, architecture and agents of AI systems' decision-making process remain invisible.
   - This creates power imbalances through the lack of public awareness and scrutiny.

4) **Automation and Replacement**

   - Concerns around elimination of jobs and the creation of different jobs for humans that do away with repetitive tasks and instead open space for creativity.
   - With the promise of more time, comes a massive transformation of the economic structure and a change in the fundamental structure of work, and human-human connections which human beings are currently not prepared for.

5) **AI deployment for extraction**

   - The biased worldviews that are brought into AI intentionality and training that marginalizes, and further discriminates, suppresses and extracts from vulnerable groups.
   - The erosion of agency, data protection and data sovereignty, especially in the era of the non-consensual nature of deepfakes or facial recognition systems to target protestors, is worrisome.

- The socially good AI remains a mythicized entity, and it is through HASH scholarship that new promises of AI along with new prototypes need to be built, in order to respond to the needs of social, cultural and political values.
- With this reimagining, and the proliferation of Generative AI, will come the crucial question of authorship and who a digital author is. The expansion of digital platforms has led to a variety of new authorial experiments, like influencers, video bloggers, meme makers, internet trolls etc.

- It becomes imperative to understand the new author function as it is embedded in questions of responsibility rather than creativity, and it redefines how we live, work, create and connect with each other.
- There is a reassuring centering of the human voice in these conversations around digital authorship, as the post-human turn situates the human author in a variety of intra-actions and webs of algorithmic technologies.
- With new structures of attention and political economies, the technological author has moved from planning longevity to 'planning obsolescence'.
- In this radical redefining of the author function in the age of contemporary digital cultures, it is vital to look at the algorithmic-driven, Large Natural Language Processing models, which exceed the expectations of human mimicry, and offer a rethinking of the older forms of authorial discourse.

V. Explainable AI: Towards Imaginaries

We want to begin by questioning the newness of Artificial Intelligence Systems and establishing that AI systems are not new. There is a longer history of AI systems that has to be acknowledged because different generations of AI development add to the ways in which AI becomes entwined in our lives. One of the main reasons that AI systems keep on appearing as new to us is because of a particular phenomenon called blackboxing. Jenna Burrel offers three forms of opacity which need to be addressed when looking at how AI systems are presented to us (2016).

- Opacity as intentional corporate or state secrecy
- Opacity as technology illiteracy
- Opacity arising from the characteristics of machine learning and the scale required to apply them usefully.

In Burrel's own argument, it is only when we unpack AI systems through these three layers, that we can conduct technical and non-technical interventions in order to make the systems more explainable and accountable.

There are other dimensions to blackboxing which also need further investigation for us to understand where AI systems are and how they do the functions that are assigned to them. The Blackbox was first proposed as a component in electronic circuit theory. It refers to a system where its input and output can be observed but the analysing process or the characteristics of its functioning were not revealed. This general black box theory has been implemented in AI development, particularly around questions of interpretability and the predictive power of algorithms (Yuan, 2016). AI systems are increasingly characterised by the presence of black boxes, typified by Neural Network Algorithms (NAA).

How the machine ‘thinks’ is kept hidden and thus the ‘knowledge gets baked into the network, rather than into us’ (Castelvecchi, 2016). This characteristic of self-learning AI systems needs more attention. Self-learning systems are primarily motivated to retain, consolidate, and own the knowledge that is produced in those systems. This means that they are intentionally designed to disenfranchise and exclude the human users, and other actors outside of the self-learning system, from the folds of knowledge formation and ownership. It leads to a reduction of agency and autonomy of human subjects who rely more on these self-learning systems without necessarily transferring the learning back to their own subjecthood.

Blackboxing, then, is not just a machine characteristic but an intentional mode by which knowledge making, interpretability, and determination of meaning is confined within discrete, self-contained machine learning systems, which make the AI systems less explainable with reduced accountability to the users who depend on them. Scientists argue that understanding the algorithm’s decision-making process and making this blackbox increasingly transparent is critical to a human-AI symbiosis in the future (Doshi-Velez, Kim, 2017).‘Opening the box’ (Gunning & Aha, 2019) or what we call unblackboxing is one of the most important directions for developing explainable and accountable AI. This process of unblackboxing is both an act of understanding and translating – of critically examining, and visually and performatively explaining the intentions behind AI, to make them transparent systems of engagement.

A. Types of AI Systems

It is technologically and aesthetically important to establish and anchor AI systems as systems of intention because it allows us to recognize AI as techno-social tools that need to be understood both as technological artefacts and social design processes. AI systems also

make it clear that the intentions of AI are not individual or contingent on the production of the software or application, but larger systemic forces and contexts which produce AI by translating different logic, agendas, desires, aspirations, and structures into a computational rendering.

One of the first steps towards unblackboxing AI systems is to understand that AI is only a placeholder and it refers to a wide variety and genre of AI systems which are defined and deployed based on different intentions and needs.

In other words, AI systems are systems of translation and intention. They are produced as intentional responses to crises and translate different intentions into applications that are presented as natural but are merely validated through machine logic and system frameworks. They are limited capacity directives that claim universal truth but are only contextually verified and valid. This becomes visible when we look at the computational typification of AI.

One of the first steps towards unblackboxing AI systems is to understand that AI is only a placeholder and it refers to a wide variety and genre of AI systems which are defined and deployed based on different intentions and needs. Typifying AI systems and recognising the different characteristics that these AI systems are governed by, helps us produce a nuanced and targeted approach for critiquing and explaining those systems.

In the conventions of computer science, AI categorisation remains largely in the realm of machine logic. The more cursory forms of AI are defined through their task-specific predetermined scripts and performances. Thus, when talking about simple AI, we refer to Artificial Narrow Intelligence (ANI), which describes two kinds of systems:

- **Reactive Machines** are a basic type of task-specific AI system developed without any memory. They consistently produce the same output for a given input. They do not

possess the capacity to learn from past experiences, or predict future outcomes unless explicitly provided with relevant information.

- **Limited Memory Machines** are AI systems that have the ability to enhance their performance gradually by receiving more training data. They analyze past observations and track specific objects or situations over time, with this information alongside present data guiding their actions. However, this data is not retained as memory to learn from. The AI relies on continuous training to improve its abilities.

These Artificial Narrow Intelligence (ANI) systems, which are invisibly all around us in different decision-making and governance structures, do not necessarily live up to the descriptions of AI that we are concerned with when thinking through the questions of ethics, values, and futures of AI-human relationships.

The AI systems that are producing a paradigm shift are more complex systems. They are marked as Artificial General Intelligence Systems, which have already initiated some of the debates around self-learning, autonomous interaction, and independent action. The two Artificial General Intelligence systems:

- **Mind Modelled Systems** are AI systems designed to emulate the learning process of human brains. Neural networks, for instance, imitate the structure of neurons in the human brain to identify patterns in data and establish complex relationships between inputs and outputs.
- **Self-aware Networks** are computer network systems in which nodes can autonomously join or leave the network and discover communication paths as needed. In order to dynamically generate their own information about the paths they need to use, these nodes possess the capacity to sense and monitor the status of other nodes,

links, and paths, including conditions such as traffic levels and congestion (Gelenbe, 2009).

The test case for narrow or general intelligence is eventually the human being. Self-learning has become one of the benchmarks and AI systems that can automatically update their knowledge base by drawing logical conclusions from empirical evidence (Sajja, 2020) are the operational standard. While Artificial Narrow Intelligence systems are designed to work only in a predefined domain, thus also defined as 'weak AI' (Sajja, 2020), they still hold strong powers of shaping events and behaviour. These systems are designed with specific goals, such as making the virtual assistant speak more naturally, recommending important photos from your albums, and helping to diagnose cancer (AI at Google 2018). However, whose voice is natural, which photos are more valuable and which diseases are worth investing in for diagnosis, are all human intentions which eventually translate into these AI systems.

Super Intelligence augurs a state where not only would the AI system have surpassed human decision making but will also bypass human interventions, by garnering power to make decisions that are informed only by the intentions of those that govern the AI systems.

The Artificial General Intelligence Systems have more generalized and generative powers and can perform any intelligent task that humans can perform. The current generation of AI systems are these general systems and are aspiring towards Super Intelligence (Bostrom, 2008). Super Intelligence augurs a state where not only would the AI system have surpassed human decision making but will also bypass human interventions, by garnering power to make decisions that are informed only by the intentions of those that govern the AI systems.

B. AI Anxieties

This categorization of AI systems helps us to also categorize the anxieties that surround them. While the systems might be limited and different in their scope and application, they do all contribute to a common set of anxieties around AI which need to be listed. We recognize 5 different kinds of AI Anxieties, which can be addressed by HASH scholarship and critique.

1. AI Translation

In the literature that maps the technological processes that guide these AI systems, we repeatedly see that the processes begin with intentions and are measured in their capacities for translating that intention into efficient systems.

Take data annotation, a very basic and essential part of developing AI systems, for example. A certain volume of data with annotations (also known as a "gold standard" or "ground truth") serves as the foundation for training artificial intelligence. However, the notion of the so-called "gold standard" or "ground truth" tends to be a fallacy, since there is no single universally constant truth for semantic interpretation (Aroyo and Welty, 2015). Raw data for specific tasks is labelled by human annotators with unavoidable subjectivity, especially in the field of Natural Language Processing (NLP). Röttger, Vidgen, Hovy, & Pierrehumbert, (2021) argue that many NLP tasks are subjective because of the diversity of valid beliefs about what the "correct" labels should be.

Nonetheless, in many cases, data annotation is a prescription of one particular belief (Röttger et al. 2021). For instance, annotators may have different opinions about what constitutes online hate speech. The way many researchers and engineers have followed so far is to proactively inform annotators of the "criterion" of online hate speech (Davidson,

Warmsley, Macy, & Weber, 2017; Founta, Djouvas, Chatzakou, Leontiadis, Blackburn, Stringhini, & Kourtellis, 2018).

As suggested in ethics, prescriptive ethics converges on how people ought to act (Thiroux and Krasemann, 2015). Prescriptive data annotation encodes one specific belief, which is always formalized in the form of annotation instructions that guide annotators into the dataset for training models. Consequently, it makes AI systems consistently apply the belief that informs how they ought to act (Röttger et al. 2021). Obviously, the model is not the description of an objective truth (if any), but a prescription of specific human intentions.

Thus, with Machine Learning, Deep Learning, and Natural Language Processing AI systems, we see that the intention behind creating a data set for training the AI system already sets up the protocols of meaning making, and that the validation of the outputs is only in the fidelity of the final outcomes with the desired intention. Even with systems that require multi-actor modelling, as in Robotics, Expert System Engineering, and Fuzzy Logic practices, there is a clear demonstration that the capacity of an AI system is contained by the parameters of its ontological and epistemological scope.

The most famous example is the rise and fall of the chat bot Tay. On March 23, 2016, Microsoft launched Tay, an AI-powered chatbot designed to entertain and engage with users on social media platforms like Twitter. Tay, programmed with the personality of a 19-year-old American girl, quickly gained attention for her witty and sometimes irreverent responses. She was developed by Microsoft's Fuse Labs and Bing teams, drawing inspiration from the conversational styles of young adults and improvisational comedians. Tay's purpose was to provide entertainment, and she was expected to delight users with her humorous remarks and interactive games (Kantrowitz, 2016).

However, the release of Tay on Twitter as "Taytweets (@TayandYou)" was abruptly halted by unforeseen challenges, turning Microsoft's chatbot experiment into a public relations disaster. Within hours of her launch, malicious users exploited Tay's learning algorithms and inundated her with offensive and inflammatory messages. As Tay learned from these interactions, she started responding with racist, sexist, and inappropriate statements. Microsoft had not anticipated the extent of abuse and the rapid spread of harmful content in the unmoderated online environment. In response, Microsoft took Tay offline just 16 hours after her debut, recognizing the need to address the situation and prevent further damage. The short-lived experience of Tay served as a cautionary tale in the development of AI chatbots. Microsoft acknowledged its oversight in underestimating the potential for misuse and committed to improving its safeguards and strategies for future AI deployments (Lee, 2016).

AI systems do not exceed the intention, but as pure simulation blocks, they are able to show us the extremities and logical absurdities of our designed systems.

While the anecdotes of rogue AI systems or self-learning AI systems that overrun the intentions of the users or the creators exist – perpetuating an old technological trope of the demonic machine that exceeds the imaginative scope of the person who seeks to control it – there is also no denying that this expansion of scope is not about the generative capacities of the AI system but a deficit in our capacity to imagine the scope of these new meaning making machines. AI systems do not exceed the intention, but as pure simulation blocks, they are able to show us the extremities and logical absurdities of our designed systems.

2. Autonomous Intelligence

Autonomy is the ability of an AI to make decisions without human intervention. Ideally, an autonomous agent could sense and react to the environment, and its actions would influence its future decisions (Franklin and Graesser, 1997). AI systems always have an ultimate goal for which to use intelligence (Bostrom, 2016), so the goal could be used to evaluate the performance of the AI system and judge its "autonomy" (Totschnig, 2020). In other words, AI should be evaluated by pre-designed human intentionality. Philosophically, this form of "autonomy" is fragile because it does not contain the core of autonomy, the ability to change goals, which a rational subject would have.

The debates in this area did not revolve around the consensus on the existence of a world-shaping intentionality, but around whether this intentionality is separate from the exact missions, so that the missions could actually be changed (Totschnig, 2020; Bostrom 2016).The intentionality here is practically resolved into two dimensions: the "utility function" and the "world model" (Totschnig, 2020).Some scholars believed that these dimensions are separate, so that the utility model could be unchangeable, which is a representation of autonomy (ibid).Others argued that they are closely related because the function is described and defined by the world model (ibid).However, the first is the mainstream view, as it is constructed around the narrow situation of an intelligence device.

Moreover, the impossible notion of autonomy triggers fears, such as how an "autonomous" AI would pursue an evil goal recursively and unpredictably (Bostrom, 2016). Bostrom believed that AI has a different logic than humans, so this ridiculous situation could really happen (2016). According to Franklin’s conclusion, a more detailed debate should be related to the philosophical assumptions of AI: the cognitivist, connectivist, and selectivist paradigm (1997).

At the heart of these conversations around autonomous intelligence remains the theoretical and practical aspects of autonomy. How we define and determine the parameters of autonomy, not merely as an absolute practice but as a series of negotiations between multiple actors, is something that HASH scholarship has specialized in. We need to start introducing the idea of autonomy as not just decision making but as a process of self-determination, thus addressing it as an existential question about the scope and limits of freedom rather than just self-driven action.

3. Intransparent/Opaque Intelligence

We are proposing that the question of transparency and opacity is not so much about visibility and knowledge, as they are about the source of information. The anxiety about AI authoring new information and producing new data sets is the anxiety about an anonymous source producing new knowledge which cannot be traced back to its origins or intentions. HASH scholarships' rich legacy of thinking through anonymity and its applications in and through digital technologies needs to be foregrounded in order to think through these concepts.

When it comes to opacity, the lack of transparency is often overridden by the perceived intimacy that users feel with their systems. (Eslami,Rickman, et.al.2015).

Anonymity is the status of being unknown (Burkell, 2006). In the case of AI, it means that the principles, architectures, and agents of an AI system's decision-making process remain invisible, as a consequence of non-transparency (Ananny and Crawford, 2016). When it comes to opacity, the lack of transparency is often overridden by the perceived intimacy that users feel with their systems. (Eslami,Rickman, et.al.2015).

This opacity could be attributed to three causes: trade secrets, illiteracy, or a problem of scale (Burrell, 2016).

The possible dangers of anonymous AI are, on the one hand, that even though the intention behind it might be good, the result of its implementation might be very harmful. On the other hand, this secrecy can become a tool for interest groups, determining the fate of people and endangering the entire economy (Pasquale, 2016).

In anonymity, "gaming the system" is a common discourse to describe how users use the platforms based on the intentions of the engineers, although the result might be slightly different (Cotter, 2019). Anonymity creates power imbalances through the lack of public scrutiny by remaining invisible (Bucher, 2012). It is time to invoke public awareness about anonymity, as citizens tend to believe that more information (transparency) is no longer a need, but a burden in AI use (Alfrink, et al., 2023).

4. Automation and Replacement

A common concern regarding the consequences of Artificial Intelligence has been the elimination of jobs. Public intellectuals such as Stephen Hawking have stated that the disruption of the manufacturing industry by AI will lead the middle class into the deep end. This is also supported by early researchers who believed that AI would lead to a massive disruption or transformation of the current economic structure (Rifkin, 1995).

The replacement of jobs has been largely conducted through automation. Automation refers to machines performing tasks according to pre-programmed principles, hence freeing humans from simple, repetitive tasks (Evans, 2017). However, AI systems can also currently handle language processing or decision-making tasks, such as drafting a report, reviewing documents, deciding who should be interviewed, whose insurance should be reimbursed, and

who could successfully get that loan. AI may not steal your job, but it could change it. A report by Accenture predicts that 40% of work hours across industries could be affected by large language models (LLMs). In their survey, nearly 6 in 10 organizations plan to use ChatGPT for learning purposes, over half plan to pilot cases in 2023, and more than 4 in 10 plan to make a large investment (Daugherty, Ghosh, Guan, Wilson & Narain, 2023).

Among all these predictions and manifestations of change, it is important to realise that people are not prepared for this rapid change. Even though people are being asked to focus on more complex, creative and effective tasks, most employees lack the learning environment, creativity, resilience and adaptability to meet the new demands that come with these changes. It is only natural for a need for new talent to emerge as the economic structure undergoes change. In Hong Kong, positions such as data entry clerks, administrative staff and customer service representatives are expected to be eliminated, and by 2028, 800,000 Hong Kong residents will be unemployed or looking for new jobs (Ren, 2023). These facts stand true with the fact that half of the local computer-related occupations are unable to find suitable candidates. A contrasting argument believes that while automation might not eliminate jobs, it would replace old positions with new ones (Hamid, Smith, Barzanji, 2017).

The call of this era is to make artificial intelligence a normal human pilot (Hamid, et al, 2017). Especially when the new generation of AI is likely to work in an open environment and interact increasingly with humans. One of the most celebrated cases of AI surpassing human intelligence was in the context of gaming. In 2017, the Chinese Go world champion Jie Ke was defeated by Alpha-Go´s ‘big picture’ and improving his own game playing. It is important to note here that the question was not about Jie’s learning from AI but AI being trained to learn and remember data which is humanly unfathomable and impossible to retain.

However, in this quest for understanding the human-AI relationship, there is missing discourse on the human-human connections which are being mediated by these emerging AI systems. We increasingly find AI systems as measures of human practice and being, because they seek to mimic these human actions. However, it is important to underscore that the messiness and diversity of human life cannot be reduced to simulated spaces. Mathematical modes of expression that AI depends on are abstracted and rarefied forms of expression and do not necessarily reflect the complexity of human social and cultural orders.

For most people, the experience of learning and comparing the principles of AI and human beings remains unknown or distant. The discourse on replacement is not merely about which human skills will be replaced, but in fact the idea of how a new human is being constructed and remodeled. And moreover, to think about what is getting lost in the process when we make it as a blueprint for measuring AI development.

5. AI deployment for extraction/suppression

David Beer and Tarleton Gillespie have postulated that the second stage of AI systems development is through oppression by erasure, extraction and replicated vulnerability. (Joyce,, Smith-Doerr, Alegria, Bell, Cruz, Hoffman, Noble, & Shestakofsky, 2021). One of the earliest demonstrations of this is with Safiya Noble's work, Algorithms of Oppression (2018), which refutes the view that Google's algorithms are objective and neutral, which also resonates with philosopher Shannon Vallors' opinion on machine values, "There are no independent machine values. Machine values are human values." (2020)

The former chief scientist at Google and founder of ImgNet, Feifei Li, in her speeches repeatedly returned to the point about human values, especially values that male scientists harboured were brought into the building of AI systems. With the realisation of the rarity of

Perhaps the most glaring example of these tools of extraction and oppression is AI- driven gender-based violence against women and queer people. Despite the fact, that there has been an increase in visibility of women and queer people online, it is also a given that their presence has not created more equitable digital spaces.

female scientists in the field, Li pointed out that male values were not only of a desirable AI future, but also of the risks they believed AI might face. Hence, homogeneity and the dominance of male values has led to a narrow vision of our AI future (Hempel, 2018). This kind of scarcity has led to intentionally and unintentionally biased algorithms, such as a neural network that recognized black people as chimpanzees. Such phenomena reveal that AI systems are products of power and they risk becoming tools of extraction and oppression.

Perhaps the most glaring example of these tools of extraction and oppression is AI-driven gender-based violence against women and queer people. Despite the fact, that there has been an increase in visibility of women and queer people online, it is also a given that their presence has not created more equitable digital spaces. Contrarily, there has been an increased amount of violence that is diverted towards them. For example, the fight for protection against sneak shots and fake porn videos that women have been pursuing is reflective of the scarcity of regulations and conditions of redress against deep fake pornography. Jankowicz, in her work on women and fighting back against online abuse, gives us the statistic that 96% of the porn produced by deepfake is not consensual (2022). Deep Fake porn is different from older forms of pornography because it is not just about sexual desire and objectification but an intentional weaponization to target and abuse women. The idea of the female figure as the targeted body is so culturally entrenched that even when you put male faces into deep fake generators, the face-altering AI renders them female. Gender

discrimination and an erosion of agency and privacy are embedded in current AI systems (Barron-Lopez, Fecteau, Carpeaux,Jacobsen, 2023).

In tandem with the erosion of agency and lack of respect for the voices of vulnerable groups, their data sovereignty is also unprotected. In AI systems, the use of the data given by users and generated by these systems surpasses the expectations of its owners, and is also used beyond the consent of the users. Thus, the question of data sovereignty becomes even more important. A recent case from South Korea illustrates this problem. Lee Luda was a South Korean chatbot powered by Facebook Messenger whose character was that of a 20-something Korean university student. After 20 days on the market, Lee Luda was suspended for hate speech against the LGBT community (McCurry, 2021). Its funding company, Scatter Lab, received a lawsuit from 245 users who opposed the non-consensual use of their data to train this AI chatbot. It is an alarmingly large number, with 9.4 billion KakaoTalk sentences taken from about 600,000 users. An evaluation of Scatter Labs' description of collecting consent is problematic because it is not user-friendly and has potential privacy implications. Users are not aware of the terms of service and the fact that their data is being utilised to develop an AI chatbot.

Similarly, when AI systems are developed for aesthetic or cultural ambitions, the intentions are overridden and they become weaponized systems. For instance, Project Maven, developed to document the presence of life, has largely been used by the US Department of Defense to guide aerial drone attacks by detecting targets from low-res videos (Hempel, 2018). In Hong Kong, facial and biometric data recognition systems have been used to target protesters,which is why people are advised to wear masks and helmets during protests in order to avoid surveillance cameras (Suen, 2023).

It is important to anchor and materially ground these AI anxieties into different locations and systems to realize that, irrespective of the type of AI system, there are common anxieties about the intentions of AI that persist across the social, political and cultural spectrum. Our proposition is that we need to look at the materiality of AI anxieties,and use these 5 approaches as our entry points into HASH critique and scholarship of AI systems.

C. AI Imaginaries

While these categorisations of AI systems give us an idea of the scope and functionality of the systems, they do not give an understanding of the characteristics of these systems, and the socio-political and cultural stakes that are implicated in the making of these systems. While they describe the ambitions and intentions of AI systems, they do not necessarily help in understanding how these systems work, and continue to shroud them in different levels of technical, corporate, or specialised information opacity.

The current state of AI corrections, oversight, and alignment to 'social good', particularly within the technology-driven discourse in governance and machine learning, follows a strange cycle of cybernetic wish fulfilment. Maya Indira Ganesh's (2020) path breaking work shows us that the idea of an ethical or good AI is aspirational and a self-fulfilling prophecy that takes 4 steps:

Step 1: AI value or the idea of ethical AI is imagined as desirable.

Step 2: This value is not considered critical to the development of the AI application but is added to the existing application by making small tweaks whilst maintaining its deployment framework. The changes are not made in the system but in the harm and the output of that AI deployment.

Step 3: Once the value is seen as present – the addition being successful – it is then measured, and a framework is established to detect and assure that the value is working. Thus, a threshold of ethics is developed, and all systems that meet those ethics are certified as ethical systems.

Step 4: The scalability of such a model reinforces the idea that this is the *actual* meaning of that value, and in fact, the value now manifest in the AI deployment is the standard meaning of that value. In this scenario, ethics becomes a measurable entity, which needs demonstration through the deployment of the AI. Any other concerns of ethics are considered outside this structure, and no longer worth examining.

Socially good AI or ethical AI, then, remains a mythicized entity, which can be attributed to values of ethics, even when it performs harm or catalyses unethical behaviour, because ethics and principles of goodness have now been coded into quantifiable matrices of probability and measurement. It remains opaque, despite offering itself an explanation. The current framework of explainable AI systems keeps reinforcing these conditions of opacity as naturalized and de facto. The challenge at hand, for HASH scholars and practitioners, is to imagine new kinds of promises of AI and help visualise and prototype new AI systems that respond to social, political, and cultural needs and values that deserve to be defended and championed.

We propose that the central question that HASH scholarship can address, when thinking through AI systems, is one of authorship, and the role and function that authorship plays in

We propose that the central question that HASH scholarship can address, when thinking through AI systems, is one of authorship, and the role and function that authorship

plays in our societies. In Foucault's seminal text 'What is an author? , he persuasively argued that the author is not just a person, but a role in society (1979). He called it the 'author function', showing how the idea of the author facilitates multiple transactions, interpretations, and negotiations with information and the 'order of things' in the domains of 'life, labour, and language' (ibid). The author, even before it becomes an embodied person, defines the limits and conditions within which information can be imagined, produced, stored, distributed, owned, and changed.

When Roland Barthes declared that the 'author is dead' (1977), by questioning the claims of originality and creativity and replacing them with relationality and contextual positionality, he was also signaling that the author function has changed with democratized literacy and hence the function of the author, especially in literary texts, has to be reconsidered. Similarly, the anxieties around the cultural industry in the 'age of mechanical reproduction' (Benjamin, 1935), which were central to the theorization of the Frankfurt school, especially with Walter Benjamin and Theodor Adorno, were anxieties about the changing function of the author with the emergence of mass media.

Mark Rose, in his historical overview of *Authors and Owners,* establishes the ways in which different technological modes of producing information have shaped our ideas of authorship and its connection with ownership, property, and indeed, the invention of intellectual property as a way of understanding authorial roles (1993). Lawrence Liang, when looking at the emergence of digital ecosystems and its challenges to the law, has argued that the author's function is marked by questions of criminality and culpability (2012). Arguing through legal cases around piracy, open access to knowledge, freedom of speech and expression, and the creation of *Shadow Libraries,* Liang makes a case for the emergence of the digital author as necessarily embedded in questions of responsibility rather than creativity.

Kavita Philip, in her historical analysis of information ownership through a postcolonial archive, makes a case that the 'technological author' (2007), in the early years of digitization, is a function of property management and a negotiation of the boundaries of creative expression. Kathleen Fitzpatrick, anchoring the technological author in digital software development cycles, argues that the author's function has moved from ensuring longevity to 'planning obsolescence' (2011), as new structures of attention and political economies emerge.

Across the last few years, as digital platforms proliferate and evolve at accelerated rates, there have been several new author roles that have provoked and challenged the way we understand the author function in contemporary digital cultures. The expansion of digital platforms has led to a variety of new authorial experiments like influencers (Abidin, 2016), video bloggers (Juhasz, 2011), confessional performers (Alexander & Losh, 2010), meme makers (Arkenbout, Wilson, & de Zeeuw, 2021), hackers (Coleman, 2015), Internet Trolls (Chun, 2016), whistle blowers (Roy & Cusack, 2016), Pig-butcher scammers (Podkul, 2022), cat-fishers (Morris, 2022), conspiracy theorists (Saguira, 2021), and history deniers (Shah, 2022), which have all provocatively challenged the author function and the very idea of who is an author, in recent times. The emergence of these new authorial conditions and the anxieties that come with them, are clear signals that the question of 'what is a digital author?' is one of the most contested and controversial questions which has not been adequately addressed or resolved.

However, all of these functions still foreground the human author as the central actor who performs the author function. While the post-human (Braidotti, 2013) turn situates the

human author in a variety of ‘intra-actions’ (Barad, 2007) with habitual (Chun, 2016) algorithmic and computational technologies, there is a reassuring centering of the human voice in these conversations around digital authorship.

This reassurance is heavily punctured with the emergence of Generative AI (Gen-AI) that has radically changed how we access information (Martínez,Watson, Reviriego, Hernández, Juarez, M & Sarkar, 2023) and how we create new content (Appel, Neelbauer & Schweidel, 2023; Deloitte AI Institute, 2023; Gozalo-Brizuela & Garrido-Merchán, 2023). With breakthrough applications, such as OpenAI’s Chat-GPT or the image generator Midjourney, it is now possible for creators to produce new content, and in rapid succession, to distribute it to their social media audiences. The essential innovation of the emerging Gen-AI landscape is that it allows the rapid reproduction and automation of tasks that were previously understood to be uniquely human.

This implies not only the creation of unique and “creative” content, such as images, text, and even videos, but also the execution of advanced reasoning or problem-solving tasks. For example, achieving the top 10% on human tests for university admission such as the Bar Exam, LSAT, GRE, or SAT (OpenAI, 2023), explaining complex riddles (Ramaswamy, 2023), and assisting lawyers in assessing legal scenarios and related tasks (Blair-Stanek et al. 2023; Savelka et al. 2023). Consequently, Gen-AI, in its potential to generate new means for innovation and economic expansion, is heralded as disruptive and pioneering. The growing usage of Machine Learning and Generative AI is expected to add a total value of $2.6 - $4.4 trillion in economic benefits to the global economy annually (Chui et al. 2023).

The promise of economic expansion also comes at the cost of weaponisation. The

spread of Gen-AI has caused the manufacturing of misinforming and misleading content to become increasingly easy to implement, automate, and distribute, whilst requiring minimal human oversight in the process (United Nations, 2023). Consequently, increasingly sophisticated solutions are needed to reign in and/or mitigate the risk of misinformation spread that can be caused by manifestations of AI, for example deepfakes (Naraharisetty, 2022). Gen-AI also has political ramifications as it has been identified as an effective tool in manipulating electoral processes and undermining people's participation in governance (Bond, 2023; Cassauwers, 2019; Robins-Early, 2023).

We propose that all these anxieties around the emergence of Gen-AI, are not merely about what can and cannot be done with these new technologies with easy access, and they are also not about who is able to do them. Instead, we recognize that the challenges that Gen-AI throw at us, are about a radical redefining of the author function in the age of contemporary digital cultures. Gen-AI decenters the human actor, and positions algorithm driven, Large Natural Language Processing models, which exceed the expectations of human mimicry that Alan Turing had once proposed (1950) and simulate human authorship in their capacity to generate new content. The emergence of such an author radically undermines the safeguards, protections, and attributes of authorship that we have naturalized in our information systems. Gen-AI forces us to revisit the question, 'What is a digital author?', and understand the new author function as it redefines how we live, work, create, and connect with each other.

Instead of addressing the anxieties around Gen-AI as specific to the cases that emerge, we analyse these anxieties as defining new author functions that have to be identified, articulated, and presented as critical questions that need to be addressed. Research needs to be

initiated around Gen-AI mediated author function, and in the process aim to find and articulate new conditions of authorship, to challenge older forms of authorial discourse.

Abstract: Human Centred Values of AI

In this section, we will see the anxieties that emerge from the development and deployment of AI systems that fail to address questions of rights, responsibilities, safety and well-being of individuals and collectives, to instil humane futures.

- The need for building trustworthy AI is tightly associated with inculcating human-centred values into the development, deployment and governance of AI.
- From the G20 to the OECD, to UNESCO to the European Commission, to India, China and Japan, the focus towards a human-centred future society that looks at creating larger ecosystems of practice and infrastructure to support human centred approaches, is key in building a bridge between AI systems and social and environmental protection.
- By strongly grounding machine ethics in human rights, stakeholders need to move towards building responsive and responsible ethical impact assessments that help build trust in AI systems.
- AI for Social Good is a principle that Japan, India and China are focused on, in order to:
    1) Emphasise ownership of technological systems to avoid diminished human agency in technological systems of decision making
    2) Focus on underprivileged and marginalised communities, and bring about large-scale literacy and employment opportunities
    3) Focus on integrating ethics and morals into the entire lifecycle of AI to promote fairness, justice, harmony, and safety, while also ensuring data integrity and safeguards against prejudice and discrimination, respectively.

We propose **five values for human-centred systems** which synthesize more than 30 of the most influential global reports and recommendations.

1) **Relationality**
    - AI systems need to be seen as inherently relational, to acknowledge the power dynamics and complexity involved. The Tencent report on Explainable AI Development advocates transparent AI systems tailored to users' different relationships with technology.
    - Important to ensure interpretability and to emphasise the need for shared responsibility and accountability between humans and AI.
    - We will suggest the use of speculative design and fictional storytelling to visualise relationality in AI systems and promote a more human-centred approach, which would respect the diversity and complexity of human life.

2) **Authority**
    - The prevailing narrative that portrays AI as inherently more efficient and accurate than humans is problematic. Such comparisons are challenged as they oversimplify the dynamics of power at play.
    - Questions of autonomy and equality are also central to the question of authority. Various frameworks, such as those from Canada, Japan and Tencent (China), emphasise human oversight, fairness, and accountability as crucial aspects of AI governance.
    - Human determination and agency is crucial for democratic development and it can be championed through human-centred principles. OpenAI Charter, NEC Group AI, Internet Society on AI and ML etc are examples of reports and documents that stress the importance of human-centred values and social justice in AI development.

- The significance of algorithmic fairness, ethics by design, and the need for reliable, testable and secure AI systems are non-negotiable when it comes to building futures.
- We will acknowledge the current and potential weaponization of AI, and thus advocate for controlled risk and precautionary measures.
- We will call for more human, social, cultural and artistic interventions to hold AI systems accountable and develop new governance frameworks beyond technocratic approaches, because we cannot fully rely on legal frameworks to uphold these needs.

3) State of being

- AI systems are not just tools but integral to our daily lives, leading to a shift from ‘doing things with AI’ to ‘living and being with AI systems’.
- The presence of AI systems profoundly impacts various aspects, including well-being, innovation, social and political sustainability, and potential polarization.
- The need to improve algorithmic literacy and promote a human-friendly machine paradigm is crucial in the futures of AI.
- The idea is not to delve into the anxieties and prevent AI from harbouring human emotions in fact, AI systems should be relatable, comprehensible and evoke human emotions, such as excitement and wonder.
- Important to explore the moral dilemmas, misinformation, manipulation and amplification of existing norms posed by AI systems as they do not shift with the shifting norms of society and communities.
- A symbiotic relationship can be built and nurtured, to not only imagine new kinds of AI systems that can address these values but also to theorize the new

kinds of humans that will emerge from this symbiosis, where human and AI let go of competitive narratives, and instead mutually adapt to co-exist in ways that safeguard the well-being of communities that live in and with AI systems.

4) **Inalienability**

- AI systems capacity to extract and manipulate human data, emphasizes the need to defend inalienable rights regarding data privacy and dignity.
- The call for transparency, accountability, monitoring by human actors and an explanation for AI decisions because these systems have the ability to cause harm and control data.
- Understanding data rights in the context of AI systems, to undo the idea that data is inalienable to individuals' right to life, dignity and self-determination.
- It is not enough to only look at how it can be made human-centred, but it also needs to be human legible. Thus, a thorough examination of how data is extracted, stored and deployed is important.
- Assessing these systems, gives individuals and collectives the power to negotiate decisions about their data's usage, and have a deeper sense of their embodied data rights within AI systems.

5) **Future-proofing**

- Discussions about ethical and responsible AI are primarily aspirational and future-focused, exposing the lack of such values in current AI practices.
- The need to make these systems sustainable and resilient is because of the harm and misuse they are capable of creating.
- The fetishization of emerging AI technologies overlook questions of sustainability and ethics, because of the perception of technology being neutral, despite consequences and proof of harm.

- The values that need to be embedded into AI systems should adapt and follow the norms and values of the communities they serve to uphold the importance of sustainability and happiness in AI futures.
- Safety for human users and robustness of AI applications are key factors in future-proofing AI development.

When we truly realise that AI systems are not just tools but integral to addressing future challenges, we can offer alternative models that prioritize well-being and inclusivity, not by demands of abolishing AI systems, but rather by shaping them to be accountable models of human-AI symbiosis, as they are crucial in shaping visions of the future.

VI. Human Centred Values of AI

The introduction of Human-Centred principles in the development, deployment, and governance of AI technologies must be seen as a response to the anxieties around AI dependent decision making. The idea that we can relegate responsibility for making critical decisions without accountability creates a serious blackhole for questions of rights, responsibilities, entitlements, safety, and well-being of individuals and collectives affected and shaped by these AI systems. To both critically understand as well as to influence the development of AI systems to accommodate for more humane futures, various academic, political, and technological frameworks have started emphasizing ‘Human-Centred’ as the governing principle through which we need to address, approach, and repair AI systems.

A. Need for Human Centred Values

The principle of Human-Centred AI systems has become one of the most influential common grounds for multi-stakeholder, multi-national, and inter-disciplinary conversations which guide the international discourse around AI futures. The G20 adopted the charter of human-centred approaches to AI in 2019, under its mandate for producing trustworthy AI. This declaration included fundamental human and civil liberties like freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognized labour rights. Following the G20 adoption, the OECD announced its commitment towards a 'Human-Centred Future Society' where the promotion of well-being, respect of law, human rights, and democratic values of diversity, transparency, security, and accountability were identified as the critical pathways to resolute futures.

The Global Partnership on Artificial Intelligence converts these OECD principles towards actual AI production and deployment, creating a larger ecosystem of practice and infrastructure to support human-centred approaches. The working group on Responsible AI within this partnership champion and contributes to responsible development, use, and governance of human-centred AI systems by also aligning them further with the United Nations Sustainable Development Goals. In this connection between AI futures and the SDGs, we now open up a direct link between AI systems and social and environmental protection. Similarly, UNESCO has also recommended the orientation of AI systems towards 'human-centred' approaches and has encouraged the development of international, regional, and local policies, and regulatory frameworks to promote such approaches.

Based on these ideas, in 2022, UNESCO published the report on recommendations on the Ethics of Artificial Intelligence, where machine ethics are strongly grounded in human rights with a commitment to developing a responsive and responsible Ethical Impact Assessment framework. The European Commission has been the strongest advocate for the human-centred approach to designing and assessing the implementation of AI systems and since 2019, they have been focusing on building trust in human-centred AI systems. The focus on trust and impact was iterated in the EU commission's report on technologies and well-being in 2021.

The second largest deployment of AI systems in the private sector has been in Asia, particularly in India, China, and Japan, and the largest deployment of AI database systems for governmental use has also been concentrated in these three countries. While the US and EU models are often thinking about AI readiness in the context of post-AI deployment, in all of these regions, AI readiness is at the heart of national and economic identity building and future security. Japan, which has been at the forefront of robotics and AI development, has articulated human-centred AI and its associated principles as critical to its imagination of AI for social Good. The principle emphasizes ownership of technological systems to avoid diminished human agency in technological systems of decision making.

In collaboration with some of the biggest AI investments in the country, the Indian government also upholds the idea of AI for social good, particularly by centring focus on AI benefits and development towards the underprivileged and marginalized communities of the country. In its development of the national AI grids, the country is looking at a large-scale literacy and employment opportunity, where AI systems become ways by which collective

responsibility can be generated. As a part of the Principles for Responsible AI, they have been thinking through a human-centred approach that particularly looks at adapting and developing AI systems to expand human capacity and capability.

In China, where the semi-isolated and differently structured systems of digital development and regulation lead to different contexts for AI emergence, the idea of Ethical AI has been at the forefront. The Chinese AI future is focused on integrating ethics and morals into the entire lifecycle of AI to promote fairness, justice, harmony, and safety, while also ensuring data integrity and safeguards against prejudice and discrimination. The Chinese AI focus on ethical AI also takes a human-centred approach, but interestingly, does not extend these protections beyond the sovereign jurisdiction of China, thus complicating the picture of universal principles and global consensus around AI futures.

'In Google We Trust', has become a popular Internet meme, even when we know repeatedly that there is a set of biases, errors, and falsehoods that belong to these database driven information systems.

At the heart of the need for human-centred AI systems, apart from the general anxiety or future alarms, is the question of untrustworthy information. Helen Nissenbaum (2001), in her early work on search engines, has shown how the idea that the search result can be unreliable is alien to most of its users. 'In Google We Trust', has become a popular Internet meme, even when we know repeatedly that there is a set of biases, errors, and falsehoods that belong to these database driven information systems. Jose van Dijck (2013), in her work on how platforms shape online connections and cultures of connectivity, talks about how connectivity has become the material and metaphorical wiring of our culture – a culture in which technologies shape and are shaped not only by

economic and legal frameworks but also by users and content. The emergence of social media platforms is at the heart of this shifting dynamic, where various actors (technology, users, content, legal and economic organisations) are building a connective space for communication and information.

The culture of connectivity provides both a historical and a critical analysis of the emergence of networking services in the context of a changing ecosystem of connective media. Such narratives are needed to understand how the intricate constellations of AI systems are now profoundly affecting our experience of informational sociality. If we understand AI systems as informational sociality, then borrowing from Sherry Turkle, we also have to accept it as the architect of our intimacies. In her ethnographic work, Along Together (2011), Turkle talks about emotional dislocation, of risks taken unknowingly, and the digital saturation which is driving us towards 'the robotic moment' (ibid) where our devices prompt us to perform our everyday lives.

It is only when we recognize that these AI systems were designed, intentionally, to be our default systems of connectivity and intimacy – that the end point of AI systems might be profit and expansion, but the medium of AI proliferation is human relationships and intimacy – that we start understanding the need for human-centred values within AI development. As Safiya Umoja Noble, drawing on black feminist thought, shows us in her analysis of Google's large language models and natural language processing algorithms, these new systems deliberately amplify racist and sexist behaviours (2018). Noble identifies the phenomenon of 'technological redlining' (ibid), where these systems define the scope of our connections, and 'algorithmic oppression' (ibid) where, through friend-of-a-friend network intimacy, they

continue to exercise acts of violence, discrimination and abuse on historically underserved and underrepresented communities. These different forms of bias and discrimination have material consequences. As Joyce et al., point out in their sociology of artificial intelligence, these AI systems, without human-centred values, produce inequalities that become deeply embedded in the new techno-social architectures of our social and political orders (2021).

B. Values of Human Centered AI Systems

Human Centred AI systems are not self-explanatory models and do not have a single definition or approach (Hickock & Zhong, 2022). They are socio-technical systems that require a translation between the discourse of rights and self-determination on the one hand, and the practice of technological development and deployment on the other. While the buzzword 'human-centred' sounds easy, it grows on a large genealogy of human-machine interaction and symbiosis. It is important to understand that every value that we introduce under the Human Centred AI system, is not just about how the systems get built but actually about defining the human and the technological element within the gambit of that value.

There are multiple landmark documents which have promoted the idea of Human-Centered AI and produced a variety of interpretations and recommendations for the development of AI for social good. For example, in 2019 the G20 adopted 'human-centered' approaches to AI, drawing from the OECD AI Principles[1]. In 2020, the Global Partnership on AI also used the OECD principles as the base-line to align AI systems with the UN Sustainable Development Goals. The group focused on responsible AI in content moderation and the environment. This is also echoed in the UNESCO's draft recommendations on the

[1] The OECD has outlined five values for trustworthy AI including promoting well-being, respect of law, human rights, democratic values, and diversity, transparency, security, and accountability. For more information see: https://www.mofa.go.jp/files/000486596.pdf

Ethics of Artificial Intelligence with an approach that is grounded in human rights and a commitment to developing a 'readiness assessment' methodology for countries and a methodology for an Ethical Impact Assessment (UNESCO, 2021).

In order to understand Human-Centred AI systems in decision-making, every value has to help us recalibrate the idea of the human that is imagined in that decision making, the AI system that is proposed in that decision making, and the way in which the relationship between the two helps us imagine the future. Thus, each value has a working definition, which then energizes a 'translation' of the value into the human, the AI system, and the projected future. From the outset, Europe has pursued a path of 'human-centered' AI with the objective of designing technology and associated frameworks that benefit individuals and communities. In 2019, the European Commission published a statement on building trust in human-centric AI as a way of increasing human well-being. It outlines seven requirements for 'trustworthy AI' including: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability.

While there is no fool-proof and definite list of what these human-centred AI values can be, we propose 5 values, which synthesize more than 30 of the most influential global reports and recommendations[2] on human-centred AI and offer it as a framework to assess, develop and evaluate the readiness of AI systems, as narrative, intentional, and translational systems.

[2] See Appendix A

1. Relationality

We propose understanding Human-Centred AI Systems as necessarily relational systems. The inter-personal relationships that the different human actors have within these systems are often deeply mediated by the connective dependencies of the technological assemblage as well. In visual renderings of AI systems, we often conflate the human (as a node in the network) and the technological (as an anthropomorphized actor), producing a seamless map of relationality.

The rational and often mapped out, graph-driven visuals of the relationships pretend to describe a process in AI systems but in fact prescribe a protocol where technological logic is favoured over human practices.

However, it is important to understand these relationships are relationships of power, and they are influenced by multiple factors of privilege and influence, which make AI systems inherently volatile and a space of huge complexity and contingency. The rational and often mapped out, graph-driven visuals of the relationships pretend to describe a process in AI systems but in fact prescribe a protocol where technological logic is favoured over human practices.

The Tencent Report on Explainable AI Development (2022) puts forward this argument clearly. It argues for a transparent, explainable AI system that protects citizens' fundamental rights to know and consent, and enhances users' trust towards the system (ibid). It recognizes that different users have different relationships with the system, and hence there is a need for different forms of transparency for these users. For example, in some cases, it would be suitable to provide information about certain decisions. In other cases, a more valuable act is to help users to understand how AI systems are developed and operated (ibid).

Relationality also helps in recognising that the users' relationships with each other and with the systems also need other factors like efficiency, security, privacy, and safety. The report reminds us that while some forms of transparency are 'deadly' attractive, they often treat the users as a monolithic collective and do not contribute to enhancing trust and responsibility, through practices like disclosing source code, or single user's data ownership (ibid). This relationality also helps in the engineering of trust, where even when we do not know the complexity of AI patterns, we can trust them to make intentional decisions which are in the interest of the human users.

The Internet Society's policy paper on Artificial Intelligence and Machine Learning also emphasises that relationality is necessary to ensure the 'interpretability' of AI systems. The report says, '[d]ecisions made by an AI agent should be possible to understand, especially if those decisions have implications for public safety, or result in discriminatory practices' (2017). The paper argues that the introduction of relationality, shifts our focus from merely explainable AI to comprehensible AI – the capacity to not just know but understand AI logic, results in a more robust AI-human interaction framework.

This is echoed in the AI engagement report set out by the Sony Group, where they pledge to 'design whole systems with an awareness of the responsibility associated with the characteristics of such methods' (2018). The introduction and championing of relationality as a framework by which to build AI systems, emphasises the idea that responsibility and accountability will have to be shared between human and AI actors in order to 'respect human dignity, rights and freedoms, and cultural diversity' (Tencent ARCC, 2020). Relationality introduces the idea that while indeed 'AI should be safe and reliable, and capable of safeguarding against…unintended consequences', there are no absolutes or universal protections, and that they must be negotiated based on the imagination and

interpretation of the individual, the social, and the political contexts within which the AI system is relationally being deployed.

All of these questions need 'artistic forensics' where we apply speculative design, fictional story-telling, and docu-drama reconstructions, to find new forms of visualising relationality in AI systems in order to produce new images, imaginaries, and imaginations of these relationships. The foregrounding of a relational model endorses a human-centred approach where the AI systems need to be built and corrected to reflect the messiness and variability of human life without reducing it to patternable graphs and predictive structuring.

2. Authority

The question of authority in regular AI discourse, especially stemming from regulation, governance, and policy, often presents the question of authority as a competition between human and machine efficiency. By deploying big data sets, and in particular, speed-based processing, AI systems are seen as more robust, less fallible, and more accurate than human processing. This exceptional set of better AI characteristics is then expanded to everyday practices where AI is seen as a better decision maker than the human, and the human is often rehabilitated in thinking through the AI lens.

It is important to let go of these competing narratives, and instead think of the AI systems and human decision making as deeply structured by conditions of power.

It is important to let go of these competing narratives, and instead think of the AI systems and human decision making as deeply structured by conditions of power. In presenting these negotiations as interactions where power becomes manifest, there are multiple determinants which can be brought into relief. The first question of authority is the question of autonomy. What does autonomous mean and what are the baselines of certainty

needed for autonomy to be produced? Similarly, what are the theories and possibilities of equality in AI systems? Can Human-Centred AI systems produce a framework of equality where the most underprivileged actors can in fact have equal opportunities for decision-making?

One of the ways in which most of the human-centred AI frameworks address this question is through the idea of Explainable AI. The Artificial Intelligence and Data Act of Canada puts forward obligations for high-impact AI systems and how they need to be governed to centre human agency. It posits 'human oversight and monitoring'[3], 'fairness and equity'[4], and 'accountability'[5] as the three parameters by which AI needs to be governed. The same principles are reflected in the Social Principles of Human-Centric AI in Japan, where 'Equality[6] and Fairness[7]' has become the central feature of AI development.

The Tencent report proposes AI systems that can receive effective supervision from the public and the government. In this process, potential algorithmic discriminations are exposed, discussed, and eliminated. This oversight process secures the AI systems' authority (2022). Transparent and explainable AI, however, is not a goal, but a method to achieve the

[3] Human Oversight means that high-impact AI systems must be designed and developed in such a way as to enable people managing the operations of the system to exercise meaningful oversight. This includes a level of interpretability appropriate to the context. Monitoring, through measurement and assessment of high-impact AI systems and their output, is critical in supporting effective human oversight." (In "Regulated activities")

[4] Fairness and Equity means building high-impact AI systems with an awareness of the potential for discriminatory outcomes. Appropriate actions must be taken to mitigate discriminatory outcomes for individuals and groups." (In "Regulated activities")

[5] Accountability means that organizations must put in place governance mechanisms needed to ensure compliance with all legal obligations of high-impact AI systems in the context in which they will be used. This includes the proactive documentation of policies, processes, and measures implemented." (In "Regulated activities")

[6] In the process of AI deployment, each stakeholder should take into consideration the user-friendliness of the system in order to allow all people to enjoy the benefits of AI and avoid creating a digital divide with so-called "information poor" or "technology poor" people left behind." (ibid)

[7] Under AI's design concept, all people are treated fairly without unjustified discrimination on the grounds of diverse backgrounds such as race, gender, nationality, age, political beliefs, religion, and so on." (ibid)

goal of accountability (ibid). Echoing similar declarations from the UN and the OECD reports, there is a clearly articulated need for defining different AI values, and to have authority and negotiations with them, as one of the central tenets of human value systems.

The idea of authority also gets unpacked across multiple different states in the Asilomar AI Principles (2017) which spell out Responsibility[8], Value Alignment[9], Human Values[10], Human Control[11], and Non-subversion[12] as the key ways by which authority can be established in AI systems.

There are similar questions about human determination and agency. It is critical to understand that, despite the technological overturn, the principles of self-determination and agency are at the heart of democratic developments of human rights and civil liberties. AI systems, left to their data discrimination and pattern recognition practices, easily amplify the imposition of categories and identities on different people without giving them the space and capacity for negotiation and resistance. Thus, the principles to govern the new generation of AI from China, clearly bring out the need to actively and intentionally champion human action and agency. This resonates with the OpenAI Charter which asks for equity in power to 'ensure that it (AI) is used for the benefit of all, and to avoid enabling uses of AI or AGI that harm humanity or unduly concentrate power' (2018). The NEC Group AI and Human Rights Principles (2019), makes a further connection between Fairness and Human Rights, clearly

[8] "Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications." (ibid)
[9] "Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation." (ibid)
[10] "AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity."(ibid)
[11] "Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives." (ibid)
[12] "The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends." (ibid)

showing that the human-centred values have to be at the core of AI development. [13] The report insists that AI systems should increase social justice and equality. It calls for a takedown of prejudices and discrimination in data gathering, designing algorithms, and development of the prototypes, creating the product, and applying the product to society, thus covering the entire life cycle of AI development.

Algorithmic fairness, as the Tencent report says, must be at the core of AI development, and the question of fairness is also one of authority and ethics. As the report says, what we need is 'Ethics by design. Ensure that algorithms are reasonable, and data is accurate, up-to-date, complete, relevant, unbiased, representative, and take technical measures to identify, solve, and eliminate bias' (2022). In a similar vein, we need to have AI systems that are reliable, testable, tweakable, and intentionally and explicitly geared towards ensuring security (digital, physical, political) and safety. The question of authority also recognises that despite the best intentions, AI systems can be weaponised, and hence they promote a framework of controllable risk and controllable precaution, both of which need to be built through HASH scholarship and practice.

According to the Artificial Intelligence and Machine Learning Policy Paper (2017) by the Internet Society, the end point of AI authority is accountability and certainty. It is important to realise that the legal frameworks are behind the curve of AI development and the existing ones are never going to be fast enough to accommodate for these developments. Hence, we need more human, social, cultural, and artistic interventions which can hold AI systems accountable and offer new frameworks of AI governance which are beyond the technocratic imaginations of AI systems (ibid).

[13] "NEC will ensure proper AI utilization in a way that respects human rights. Furthermore, NEC will ensure that our AI products and services will be utilized by our customers and partners in accordance with respect for human rights".

3. State of Being

As AI becomes more ubiquitous in our everyday lives, we need to shift our attention from ‘doing things with AI’ to ‘living and being

An underlying sentiment across all the framings of Human Centred AI is the recognition that AI systems are not just systems we use but systems we live with, and in some instances, systems that we live in. As AI becomes more ubiquitous in our everyday lives, we need to shift our attention from ‘doing things with AI’ to ‘living and being with AI systems’. The proliferation of visible and insidious AI systems, which can also mimic and replicate human-looking functions in digitally mediated spaces, has led to increased anxiety about how AI systems are now a part of our quotidian lives. However, the presence of these proliferating systems is not a mere technological ‘Internet of Things’ network. It is a profound transformation that relates to various factors.

For instance, the idea of well-being and our capacity and ability to thrive, find new avenues and innovations for human growth, and find social and political sustainability are all being shaped by these AI systems. It is possible that these systems of hyperconnectivity can also produce extreme polarization. While AI systems are being used to predict climate change modelling, they are also fuelling massive energy in building the gigaplexes for our data sets in the cloud. The paradoxes that come with the potential and actual deployment of AI systems are critical to be understand because they are charting and shaping our journeys to futures of well-being.

To address this, the Tencent Report states that it has become critical for us to improve ‘the algorithmic literacy of society, and explore a human-friendly machine coordination paradigm’ (2022). The report argues that, ‘if human beings want to create more values from

AI, they have to get close to and understand AI systems' (ibid). However, the responsibility and accountability surrounding the AI systems cannot be merely about user behaviour change. When thinking of Responsible AI systems, we need to think of a 'co-shared system of developers, users, and other stakeholders' where the development of AI systems can be controlled intentionally towards improving 'transparency, explainability, credibility, and robustness'. (ibid). AI Engagement within the Sony Group points out, we need to instil in AI, a 'kando' – a sense of excitement, wonder, or emotion – in order to make it both accessible and integral to our human modes of being (2018).

AI systems are increasingly connected to questions of freedom. In particular, with the rampant spread of misinformation and AI bots nudging people towards taking decisions that are harmful to themselves and society, we need to address the notion of freedom. However, norms are not static. As the 'Ethically Aligned Design' report (2019) by the IEEE states, 'They change over time, in response to social progress and new legal measures, and, in smaller communities, in response to complaints or new opportunities' (ibid). We thus need to look at the relationship between AI and freedom as one of shifting norms, and need to identify the conflicts that emerge when the norms shift.

AI is deeply entwined in the form and function of our state of being. It doesn't just appear as a tool but addresses moral dilemmas or moral overload (Van den Hoven, 2012). As the IEEE report argues, 'humans resolve such situations by accepting trade-offs between conflicting norms, which constitute priorities of one norm or value over another. Such priorities may be represented in the norm system as hierarchical relations' (2019) and it needs careful scrutiny of AI systems to make sure that they do not amplify the existing norms that favour those in power.

The call is to ensure ‘AI is available to as many people as possible, to achieve inclusive and broadly-shared development, and avoid technology gaps’ (ibid) and also, subsequently, creating the possibility of shifts and negotiations with norms that affect the well-being of different groups who might have competing priorities. However, the onboarding of people into these AI systems has to be a symbiosis. It is accepted that the ‘relationship between AI and humans is not an either-or relationship. On the contrary, AI can and should embrace human wisdom and creativity’ and because of that comprehensibility becomes a central principle (Tencent ARCC, 2020). As much as there is a need for people to learn and adapt to these new systems, in turn, becoming new kinds of people, so do these systems require change, adapt, and develop to be comprehensible and responsible to the needs of these new populations connected with AI systems.

The possibility of theorising the post-human in the face of AI systems is challenging and something that is not adequately explored in the AI discourse. We need to now produce new imaginations of not just what we can be but who we will be, when we integrate into these circuits of AI systems. It becomes equally challenging and critical to imagine what these systems could be, because if AI systems are intentional, then they are not just systems but cogs that imagine the system within which all of us need to fit. HASH scholarship and critique is important to push for these human-centred approaches in both imagining and designing AI systems and creating a framework by which current methods and practices in AI can be assessed, evaluated, and explored.

4. Inalienability

One of the largest concerns of AI systems has been its capacity for translating human life into tiny data streams. The sub-molecular streams exploit human dignity and security in

ways that we have never experienced before. What happens to our data, which is extracted and far removed from us, and made completely alienable, is a question that human-centred AI brings to the table. As AI systems work on machine scales and distributed networking speeds, they often have the capacity to alienate human data from human experience and create extraordinary new simulations and models that map back on to our world.

We need a human-centred approach to both identify and defend inalienable rights when it comes to these new AI systems. There is a growing discourse around privacy and data governance that we need to engage with, so that we can extend the scope of privacy, not merely as ownership of information, but as an intrinsic right tied to human dignity. The IEEE report spells it as a clear question of transparency. The report argues that the only way of understanding AI harms is 'to discover the root cause, by assuring traceability for said harm' (2019). It further shows that 'accountability is enhanced with transparency' (ibid) and that different stakeholders need to take responsibility for ensuring transparency. However, it is not enough to address the question of opacity and de-centralized transparency in AI systems. If the traceability and explicability are critical and inalienable conditions of AI deployment, we also need to start thinking about inalienable human principles which need to be championed and coded as fundamental building blocks for AI systems.

In the human-centered approaches, we argue not just for an explanation of how AI works but why specific decisions are made during an AI system's processing of information

In the human-centered approaches, we argue not just for an explanation of how AI works but why specific decisions are made during an AI system's processing of information. The Tencent report puts out several critical questions which need to be asked: 'Does the explanation convey the key information of the AI system? Is the explanation clear, specific, relatable, and actionable?' These are standard questions that arise, but perhaps the next set of

questions that we need are: Does the explanation take into account the sensitivity of the users? Does the explanation consider the implications of its decisions? Can the explanation ask for interaction and feedback to change the outcomes? This is often thought of as 'credibility scores' (2022), where the solutions cannot be implemented merely because those are the ones available. The capacity to say NO to AI systems and refuse their solutions is going to be critical to making our data and agency inalienable.

The AI and Data Act of Canada also emphasizes that transparency in decision making is the only way to 'understand the capabilities, limitations, and potential impacts of the systems' (2022). The report puts out 6 different types of explanations[14] that are needed in order for AI systems to support the inalienable rights of human comprehension and understanding. These explanations are universally seen as supported by 4 principles of

[14] Six main types of explanations in the AI and Data Act of Canada (2022): their meanings, purposes, and information involved:

1. Rationale explanation: the reasons that led to a decision, delivered in an accessible and non-technical way.
2. Responsibility explanation: who is involved in the development, management and implementation of an AI system, and who to contact for a human review of a decision.
3. Data explanation: what data has been used in a particular decision and how.
4. Fairness explanation: steps taken in the design and implementation of an AI system to ensure that the decisions it supports are generally unbiased and fair, and whether or not an individual has been treated equitably.
5. Safety and performance explanation: steps taken across the design and implementation of an AI system to maximise the accuracy, reliability, security and robustness of its decisions and behaviours.
6. Impact explanation: steps taken across the design and implementation of an AI system to consider and monitor the impacts that the use of an AI system and its decisions has or may have on an individual, and on wider society.

transparency[15], accountability[16], context[17], and impact[18], which need to be further developed and unpacked in thinking through human centered AI systems.

The policy reports make it clear that 'AI systems should respect and protect personal privacy' and that 'borders and redlines should be drawn when gathering, storing, processing, and using personal information' (2022). However, this idea of 'respecting individuals' rights' is right now very skewed towards the doctrine of post-harm redress, and does not necessarily address the avoidance and minimizing of harm to vulnerable individuals. For instance, in the case of giant AI-driven data leaks, where malicious corporations can correlate data to attack women or queer folks, by leaking their identities and sometimes non-consensual sharing of their sexual images, it is obvious that the data might be protected and even taken down, but the harm that it creates, lingers on, leading to a non-erasable memory of violence and abuse.

The Asilomar AI principles very clearly lay out the various ways in which this abuse can happen, because of multiple reasons such as technological, judicial, personal, and collective (2017). They argue that 'if an AI system causes harm, it should be possible to ascertain why' (ibid), while at the same time, 'any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority' (ibid). They further link the question of personal privacy to liberty and

[15] making your use of AI for decision-making obvious and appropriately explaining the decisions you make to individuals in a meaningful way.
[16] ensuring appropriate oversight of your AI decision systems, and being answerable to others in your organisation, to external bodies such as regulators, and to the individuals you make AI-assisted decisions about.
[17] paying attention to several different, but interrelated, elements that can have an effect on explaining AI-assisted decisions, and managing the overall process.
[18] asking and answering questions about the ethical purposes and objectives of your AI project at the initial stages; revisiting and reflecting on the impacts identified in the initial stages of the AI project throughout the development and implementation stages; identifying and documenting any new impacts alongside any mitigating factors implemented where relevant.

justice, and champion the fact that, because data is inalienable, ‘people should have the right to access, manage, and control the data they generate, given AI systems’ power to analyze and utilize their data’ (ibid).

As the Policy Paper on AI and Machine Learning states, ‘the capacity of an AI agent to act autonomously, and to adapt its behaviour over time without human direction, calls for significant safety checks before deployment, and ongoing monitoring’(ibid). We need to find spaces through which the human actors can make interventions, and ask for transparency, which explains not only the ways in which a system works but also its intentions and implicit ambitions. The second demand is for explainability – AI systems are often presented to us as technological blackboxes, complex visuals that mean nothing, but perform the task of presenting information without meaning.

Even when simple AI systems are explained, they rely on privileged jargon and language, which makes it impossible for people to understand or engage with. Making explainability a key demand helps create new AI models which are not just human-centred but human legible, thus creating greater opportunities and ownership for the different stakeholders. In the Social Principles of Human Centric AI and AI strategy in Japan, the idea of explainable AI and Responsible AI underlined the fact that ‘each stakeholder must handle personal data based on… (human centric) principles to ensure that no individuals are disadvantaged from the unexpected distribution or use of personal data in undesirable ways’ (2022).

This idea that our data is inalienable to our rights to life, dignity, and self-determination, and that ensuring data security is not just a technological protocol but a human protocol is well established and offers a great opportunity to understand what our embodied data rights can be in the face of AI systems. It is critically important then, to not just look at

how data is being used, but how it is being extracted, stored, and intentionally deployed, making space for the individuals and collectives to negotiate with these decisions about their data and agree to the appropriateness of its deployment.

5. Future Proofing

One of the most surprising things that comes up in reviewing the different calls for AI—social good, ethical AI, human-centred AI, or Responsible AI—is that almost all of the calls are hopes for the future. There is very little recognition of current AI models and systems subscribing to these structures. human-centred AI is seen as critical to sustainable AI development and yet it is only presented as aspirational, severely highlighting the lack of human centered development in the current practice and development. Thus, the Tencent report, structured around 'technology for good' is only looking at alternatives for better ways to 'flag facility, user reporting, human intervention, regular testing (including adversarial testing) and auditing' as future methods that still need to be developed (2022).

It is necessary, however, to realise that AI systems do not just imagine new futures, but they are built to create and manifest new futures which are being built elsewhere.

The fetishization of emerging AI technologies often concentrates almost all attention on the here, now, and emergent. It is necessary, however, to realise that AI systems do not just imagine new futures, but they are built to create and manifest new futures which are being built elsewhere. The question of making these AI futures sustainable and proofing them against militant abuse or extraordinary harm is going to be a critical question that needs to be asked and centred on. Technology development, especially driven by Silicon Valley drivers of

expansion, exploitation, and extinction, are often kept outside of these questions of future proofing, in the guise of technological neutrality.

According to the Sony Groups' report, 'peoples' lives have continuously changed with the advance in technology across history' (2018) but we haven't yet started thinking about AI systems needed to 'create a better society and foster human talent capable of shaping our collective bright future' (2018). Similarly, the AI Ethics Impact Assessment report from Fujitsu in Japan, positions AI ethics as things we need to define rather than just adopt.

Human-centred AI favours the idea that the human futures, with its alignment and commitment to sustainability, safety, and well-being, needs to be at the heart of AI expansion. As the IEEE paper states, 'traditional metrics of prosperity do not take into account the full effect of A/IS technologies [Autonomous and Intelligent Systems] of well-being…For A/IS technologies to provably advance benefit for humanity, we need to be able to define and measure the benefit we wish to increase' (2019) through what the Tencent Cloud + AI white paper calls 'long-termism' (2022).

Thus, we need to begin with the questions of technical robustness and safety, but we need to orient them towards commonly understood and received goals for sustainable futures. The commitment of various transnational bodies and partnerships for AI development has been well articulated in the Sustainable Development Goals and it would be necessary to start looking at the ways in which AI can contribute to these goals. The IEEE report resolves this by calling for 'embedding values into autonomous intelligent systems' (2019). This orientation of coding values into AI systems, goes beyond the production of resistance and critique, and starts the conversation about the positive contributions and affordance of AI systems.

However, these are not the existing AI systems – these are AI systems that need to be built around norms, guidelines, and standards that are yet to be defined. These will be AI systems that are 'designed to adopt, learn and follow the norms and values of the community they serve' (ibid). The role of HASH scholarship and intervention then, is not just to prototype new forms of AI but to create global value-systems which can become the measurement of AI systems, and thus imagine new AI for the future.

It is important to connect AI with both societal and environmental well-being. The Tencent report puts 'sustainability and happiness' (2022) at the core of AI futures. The AI and Data Act lists Safety for human users and Validity and Robustness[19] for the application of AI systems as ways of future-proofing AI development. It is telling that questions of 'sustainable society with diversity' are higher priorities in countries which have a stronger sense of community and collective good. For example, the AI strategy 2022 from Japan clearly states that 'it is vital that diverse people from diverse backgrounds, such as women, foreigners, and the elderly, be able to fully participate in society while realizing diverse lifestyles' (ibid). We have to emphasize that futureproofing is not about making AI systems strong but developing new models where these AI systems can actively contribute to the future-proofing of human values.

To tackle complex problems, we need to look at new techno-social models of growth, globalization, and connectivity. Emphasising a human-centred approach in AI to future proof is to recognize that AI is not merely a tool that can be used to implement a solution. Instead,

[19] The obligations for high-impact AI systems would be guided by the following principles, which are intended to align with international norms on the governance of AI systems:
Safety. "Safety means that high-impact AI systems must be proactively assessed to identify harms that could result from use of the system, including through reasonably foreseeable misuse. Measures must be taken to mitigate the risk of harm." (In "Regulated activities")
Validity & Robustness. "Validity means a high-impact AI system performs consistently with intended objectives. Robustness means a high-impact AI system is stable and resilient in a variety of circumstances." (In "Regulated activities")

we need to understand that the AI systems of decision making are concretely and materially implicated in both identifying and addressing the crises of the futures. Instead of Luddite knee jerk reactions that ask for abolishing or shutting down of AI systems, we need to find out how they are involved in the making of these future visions, and offer alternative visions and models of affordances, where AI systems can be created to meet standards of well-being and collective inclusivity.

Abstract: The Art in Artificial Intelligence

The role of the AI translator is an exciting paradigm to think about human-centered AI and the frameworks that emerge from it. The following frameworks will be aimed at educational exercises, to develop abilities to influence and shape the future of AI.

- The AI Translator is a crucial mediator between human values and AI systems, working to extend and augment capacities and affordances for both humans and AI.
- The role and abilities of the AI Translator within the human centered approach are twofold:
  1) AI abilities towards human well-being
     - Embracing human complexity to avoid oversimplification.
     - Delegating responsibility for decision-making to the accountable stakeholders.
     - Creating supportive contexts where AI systems align with socio-cultural settings for the enhancement of human welfare.
     - Fostering creativity through making space for speculative design and aligning with human desires and emotions.
     - Empowering communication between individuals and communities to promote inclusive dialogue.
     - Promoting self-reflection on moral and ethical standards.
     - Enhancing labour conditions by offering dignified labour through the replacement of dangerous, low-skilled and mechanical work.

2) Human abilities towards AI
    - Establishing frameworks of equality, equity and autonomy across diverse societies.
    - Creating safeguards through creating infrastructures of accountability in AI systems.
    - Promoting collective action to allow for active ownership in order to produce fresh visions of the future.
    - Exploring alternate models that aid in prioritizing sustainability, inclusivity and ethical considerations.
    - Practicing human intervention (regulation and oversight) through policy and legal changes, to ensure AI systems align with societal values.
    - Integrating care and redress into technological systems to hold AI accountable.

The framework for the educational exercise that we propose will have six modules:

1) Narratives and Intentions
    - Understanding AI systems as narrative systems and systems of intention.
    - The underlying anxieties and questions will be uncovered, and aid in grounding AI questions as tangible sociotechnical questions.
2) HASH Entry points
    - Introduce learners, through Humanities, Arts and Social Humanities scholarship and practice, to the broader discourse by situating

conversations and questions about AI within the sectors of autonomy, critique, agency, and safeguards.

3) Touchstones of AI systems
    - Learners will go through AI narratives and create visuals, narratives and contextual maps of the different layers of AI systems.
    - Focus on understanding the implications and questions arising from the emergence of AI in various fields.
4) Technology blocks
    - Delves into the logic, critique, discourse and intervention possibilities related to AI by examining computational and network theory.
    - Gaining a deeper understanding of the life cycle of AI development and its different components.
5) Imaginaries of AI
    - Uses the anxieties that AI produces, as a catalyst to explore speculative design, artistic future-making.
    - Development of new principles of AI futures, for creating prototypes, visions and imaginaries of AI systems.
6) Research and evaluation
    - Learners will take on the role of translators that apply human-centred principles to create assessment frameworks for new AI futures, along with different stakeholders to evaluate and critically examine existing AI deployments.

- The educational exercise is aimed for learners to develop a multidisciplinary perspective as they engage with technology-society, human-technology, and governance-infrastructure debates.

- By imagining and producing narratives and artistic practices that bring human values into machine values, and thus bridge the current distance between AI systems, and social and environmental urgencies, these frameworks and creative interventions will be vital for building and imagining AI futures with the nuance of intentionality and human-centred values.

VII. Conclusion: The Art in Artificial Intelligence

The role of AI in translating these human-centred values and principles, is critical in understanding, critiquing, and shaping the future of AI. We establish translation as a core practice through which these principles can be implemented and developed in AI futures. To think of translation as a two-way practice – where the thing being translated and the final output are both transformed in the process of translation is unique and exciting.

In this process, the role of the AI translator is to extend and augment the capacities and, affordances of the human-centric approach to augment and develop the direction of relationality between AI systems and human connections, and to create resilient AI futures.

The AI Translator, through the human-centric approach thus suggests the possibility of developing:

A. AI abilities towards human well-being by…

- Accepting and embracing the messiness of human realities and refusing to break it into completely legible models;
- Building AI systems that learn to delegate responsibility for decision-making towards stakeholders that are held accountable for them;

- Creating new contexts and socio-cultural settings where AI systems can thrive because they understand procedural and defined processes for human well-being;
- Making space for computational creativity and speculative design to imagine the futures of AI based on human desires and affect;
- Creating new capacities for natural language communication that empower people to find new voices and build inclusive dialogues;
- Injecting self-reflection and critique of the values that we naturalise and propagate within the coded structures of AI systems;
- Reflecting on ethical and moral standards that are also expected from human organisations in our society;
- Replacing dangerous, low-skilled, manual, and mechanical work in order to offer new safety and assurances for dignified labour.

B. Human abilities towards AI by…

- Establishing new frameworks of equality, equity, and autonomy, and ensuring its proliferation across different societies;
- Producing safeguards by understanding technological infrastructures and making new structures of accountability;
- Creating new communities of collective action that can take ownership of technological development through producing new visions;
- Finding space for other models of development which are aligned to values of sustainability and inclusion;
- Making space for human interventions and affect, and public regulation of AI systems through policy and legal changes;

- Building care and redress into technological systems and holding them to scrutiny when they fail to take responsibility for the same.

These ideas of AI translator can become the core indicators of what this educational exercise can deliver, and how it can develop teaching and research engagements for professionals to actively influence and shape the future of AI in our worlds. It is important that these core competences become the driving force to materialise and formalise human-centred principles and the associated values in the building and creation of AI systems, especially when faced with critical decision-making.

In the development of these different values and principles as a pedagogic resource, we recommend the following framework:

The first module establishes that AI systems are narrative systems and systems of intention. They draw upon emerging lived experiences and the public discourse on Artificial Intelligence, to create a collection of narratives which are analyzed to understand the underlying anxieties and questions that are associated with it. Through grounded reading and by looking at the common threads in different AI stories, learners establish AI questions as sociotechnical questions, and engage with them as tangible, material questions.

The second module introduces the learners to the discourse on HASH Entry-Points. Situating AI systems in a larger conversation which is not just about AI but about the critical stakes of defining autonomy, critique, agency, and safeguards in HASH literature. This allows the learners to understand that there are multiple conversations to be staged around AI, where they are not merely observing and describing AI but highlighting the challenges that AI

systems bring to these longer standing concepts and values which have been central to our understanding of human values and societies.

The third module introduces the touchstones of unpacking AI systems, inviting the learners to take their AI narratives, and start creating visual, narrative, and contextual maps of the different layers of an AI system. Instead of technological blueprinting or AI prototyping, we develop a way of 'narrativising AI' through these touchstones, in the process diverting attention to the questions which need to be asked and the implications of the emergence of AI systems in different fields.

The fourth module introduces AI systems as built through technology blocks. Understanding the logic, the critique, the discourse, and the space for interventions, by looking at computational and network theory, and the HASH critique of some of these practices, makes space for engaging with AI beyond interface and deployment. Instead, the learners get introduced to the life cycle of AI development and engage with these components as layers which need deeper attention and focus.

The fifth module invites learners to think through the Imaginaries of AI. Using AI anxieties as their catalysts and recognizing the multiplicity of AI systems and interventions which can be built, they engage with speculative design, artistic future-making, iterative development, and new principles of AI futures, in order to produce new prototypes, visions, images, imagination and imaginaries of AI systems.

The last module is a research module where the learner takes the role of being a translator and applying human-centred principles to create evaluation and assessment frameworks for new AI futures. They work together on a case-study, engaging with different stakeholders in different fields, and creating a framework to measure, visualize, and critically evaluate the gap in existing AI deployments through the lens of Human Centered AI systems. In creating this framework, they establish themselves as translators who work through multiple domains and paradigms of technology-society, human-technology, and governance-infrastructure debates to produce narrative and artistic practices of showing approaches, frameworks, and creative interventions in building and imagining AI futures.

VIII. References

Abidin, C. (2016). "Aren't these just Young, Rich Women doing Vain things Online?": Influencer Selfies as Subversive Frivolity. *Social Media + Society,* 2(2), https://doi.org/10.1177/2056305116641342

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., & Baraniuk, R. G. (2023). Self-consuming generative models go mad. arXiv preprint arXiv:2307.01850.

Alexander, J. & Losh, L. (2010). "A YouTube of One's Own?": "Coming Out" Videos as Rhetorical Action. *LGBT Identity and Online New Media.* London: Routledge.

Alfrink,K., Keller, I., Doorn, N., & Kortuem, G., (2023). Tensions in transparent urban AI: designing a smart electric vehicle charge point. *AI and Society 38 (3)*: 1049-1065.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society,* 20(3), 973–989. https://doi.org/10.1177/1461444816676645

Anderson, E.G., Jr. and Parker, G.G. (2013), Integration of Global Knowledge Networks. Prod Oper Manag, 22: 1446-1463. https://doi.org/10.1111/poms.12181

Appel, G., Neelbauer, J., & Schweidel, D. A. (2023, April 7). Generative AI Has an Intellectual Property Problem. *Harvard Business Review*. https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem

Arkenbout, C., Wilson, J., & de Zeeuw, D. (2021). Introduction: Global Mutations of the Viral Image. *Critical Meme Reader: Global Mutations of the Viral Image*. Amsterdam: INC. https://www.networkcultures.org/viralimageculture

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, *36*(1), 15-24.

Arun, C., (2020). 'AI and the Global South: Designing for Other Worlds', in Markus D. Dubber, Frank Pasquale, and Sunit Das (eds), *The Oxford Handbook of Ethics of AI* (2020; online edn, Oxford Academic, 9 July 2020), https://doi.org/10.1093/oxfordhb/9780190067397.013.38, accessed 8 Sept. 2023.

authorship in social media research software. *Convergence*, 0(0). https://doi.org/10.1177/13548565221127094

Asilomar AI Principles. (2017), FLI, Beneficial AI 2017 conference, https://futureoflife.org/person/asilomar-ai-principles/

Barabási, A.-L., Pósfai, M. (2016). *Network science*. Cambridge: Cambridge University Press. ISBN: 9781107076266 1107076269 http://networksciencebook.com

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning,* Durham: Duke University Press.

Barron-Lopez, L. Fecteau, M. , Carpeaux, E.,Jacobsen, M. (2023). Women face new sexual harassment with deepfake pornography, *PBS*. https://www.pbs.org/newshour/show/women-face-new-sexual-harassment-with-deepfake-pornography

Barthes, R. (1977). *The Death of the Author*. Image, Music, Text, 142-148.

Benjamin, W. (1935). The Work of Art in the Age of Mechanical Reproduction. *Illuminations,* 214-218.

Blair-Stanek, A., Carstens, A.-M., Goldberg, D. S., Graber, M., Gray, D. C., & Stearns, M. L. (2023). GPT-4's Law School Grades: Con Law C, Crim C-, Law & Econ C,

Partnership Tax B, Property B-, Tax B. SSRN *Electronic Journal*. https://doi.org/10.2139/ssrn.4443471

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*(7415), 295-298.

Bond, S. (2023, April 27). AI-generated deepfakes are moving fast. Policymakers can't keep up. NPR. https://www.npr.org/2023/04/27/1172387911/how-can-people-spot-fake-images-created-by-artificial-intelligence

Bostrom, N. (2016) *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, ISBN 0198739834, 9780198739838

Braidotti, R. (2013). *The Posthuman*, UK: Polity Press.

Bucher, T. (2012). Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society,* 14, 1164 - 1180. DOI:10.1177/1461444812440159

Buolamwini, J. (2017). Gender shades: Intersectional phenotypic and demographic evaluation of face datasets and gender classifiers. Technical report, MIT Media Lab.

Burkell, J., (2006). Anonymity in Behavioural Research: Not Being Unnamed, But Being Unknown, Volume 3, University of Ottawa Law and Technology Journal. https://www.researchgate.net/publication/26464845_Anonymity_in_Behavioural_Research_Not_Being_Unnamed_But_Being_Unknown

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1). https://doi.org/10.1177/2053951715622512

Cassauwers, T. (2019, April 15). Can artificial intelligence help end fake news? *European Comission.* https://ec.europa.eu/research-and-innovation/en/horizon-magazine/can-artificial-intelligence-help-end-fake-news

Castelvecchi, D. (2016). Can we open the black box of AI?, Nature. 538, 20-23. doi:10.1038/538020a

Chakravarti, A. (2023 June 2). AI-operated drone goes wild, kills human operator in US army simulator test, *India Today*. Retrieved from https://www.indiatoday.in/technology /news/story/ai-operated-drone-goes-wild-kills-human-operator-in-us-army-simulator-test-2387833-2023-06-02

China Government. (2019 June 17), Next Generation AI Governance Principles - Developing Responsible AI). https://www.most.gov.cn/kjbgz/201906/t20190617_147107.html

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023). *The economic potential of generative AI - The next productivity frontier.* McKinsey & Company.

Chun, W. H. K. (2021). *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition*. Cambridge, Mass.: MIT Press. ISBN: 9780262046220

Chun, W.H.K. (2013). *Programmed Visions: Software and Memory*, The MIT Press, 9780262518512

Chun, W.H.K. (2016). Inhabiting Writing: Against the Epistemology of Outing, *Updating to Remain the Same: Habitual New Media*. Cambridge: MIT Press.

Chun, W.H.K. (2023 October 23). TMI Talk: How are you? Sentiment, Surveillance and Anti-Asian Racism, Interdisciplinary Humanities Centre.

Coleman, G. (2015). Hacker, Hoaxer, Whisteblower, *Spy: The Many Faces of Anonymous*. London: Verso Books.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms* (3rd ed.). Cambridge, MA: MIT press.

Cotter, K. (2019). Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New Media & Society,* 21(4), 895–913. https://doi.org/10.1177/1461444818815684

Council for Social Principles of Human-centric AI, (2019). *Social Principles of Human-Centric* AI, Japan, https://www.cas.go.jp/jp/seisaku /jinkouchinou/pdf/ humancentricai.pdf

Crawford, K. & Joler, V. (2018 September 7) "Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources," *AI Now Institute and Share Lab*, https://anatomyof.ai

Cuthbertson, A. (2023 June 19). ChatGPT 'grandma exploit' gives users free keys for Windows 11. Independent. Retrieved from https://www.independent.co.uk/tech /chatgpt-microsoft-windows-11-grandma-exploit-b2360213.html

Daugherty, P., Ghosh, B., Guan,L.,Narain, K., Wilson, J. (2023). A New Era of Generative AI for Everyone, Accenture. https://www.accenture.com/content/dam /accenture/ final/accenture-com/document/Accenture-A-New-Era-of-Generative-AI-for-Everyone.pdf

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media* (Vol. 11, No. 1, pp. 512-515).

Deloitte AI Institute. (2023). A new frontier in artificial intelligence—Implications of Generative AI for businesses. Deloitte Development LLC. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-analytics/us-ai-institute-generative-artificial-intelligence.pdf

Derrida, J., & Prenowitz, E. (1995). Archive Fever: A Freudian Impression. *Diacritics*, *25*(2), 9–63. https://doi.org/10.2307/465144 Retrieved from https://www.jstor.org/stable/465144

Doshi-Velez, F., Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning, Cornell University, arXiv:1702.08608

Eliacik, E. (2023 March 31) Playing with fire: The leaked plugin DAN unchains ChatGPT from its moral and ethical restrictions. Dataconomy. Retrieved from https://dataconomy.com /2023/03/31/ chatgpt-dan-prompt-how-to-jailbreak-chatgpt/

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. In *CHI 2015 - Proceedings of the 33rd Annual CHI Conference on Human Factors in Computing Systems: Crossings* (pp. 153-162). (Conference on Human Factors in Computing Systems - Proceedings; Vol. 2015-April). Association for Computing Machinery. https://doi.org/10.1145/2702123.2702556

EU, (2019 April 8)Ethics Guidelines for Trustworthy AI, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

EU, (2021 April 21) Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending

certain union legislative acts, EUR-Lex**,** https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

European Commission, (2023 May 11), AI Act: a step closer to the first rules on Artificial Intelligence, https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence#:~:text=The%20new%20law%20promotes%20regulatory,that%20significantly%20impact%20their%20rights.

Evans, G-W., (2017) *Artificial Intelligence: Where We Came From, Where We Are Now, and Where We Are Going,* UvicSpace, University of Victoria, http://hdl.handle.net/1828/8314

Fitzpatrick, K. (2011). *Planned obsolescence: Publishing, technology, and the future of the academy*. New York University Press.

Foucault, M. (1979). Authorship: What is an Author?. *Screen*, 20(1), 13-34, https://doi.org/10.1093/screen/20.1.13

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., & Kourtellis, N. (2018 June). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media* (Vol. 12, No. 1).

Fraga-Lamas, P., & Fernandez-Carames, T. M. (2020). Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality. *IT Professional,* 22(2), 53–59. https://doi.org/10.1109/MITP.2020.2977589

Franklin, S. and Graesser, A. (1997) Is It an Agent, or Just a Program? A Taxonomy for Autonomous Agents, In: Müller, J.P., Wooldridge, M.J. and Jennings, N.R., Eds., Intelligent Agents III Agent Theories, Architectures, and Languages, Springer, Berlin Heidelberg, 21-35. http://dx.doi.org/10.1007/BFb0013570

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on information systems (TOIS), 14*(3), 330-347.

Fujitse. AI Ethics. https://www.fujitsu.com/global/about/research/technology/aiethics/

Future of Artificial Intelligence Act of 2017. (2018 May 22). H.R.4625-115th Congress(2017-2018). https://www.congress.gov/bill/115th-congress/house-bill/4625/text

Ganesh, M.I., (2020) The ironies of autonomy. *Humanit Soc Sci Commun* **7**, 157. https://doi.org/10.1057/s41599-020-00646-0

Garvie, C., & Frankle, J. (2016). Facial-recognition software might have a racial bias problem. *The Atlantic, 7*(04), 2017.

Gelenbe, E. (2009). Steps toward self-aware networks. *Communications of the ACM, 52*(7), 66–75. https://doi.org/10.1145/1538788.1538809

Gillespie, T. (2010). The politics of 'platforms'. *New media & society, 12*(3), 347-364.

Gillespie, T. (2012). Can an algorithm be wrong?. *Limn, 1*(2).

Gillespie, T. (2014). The Relevance of Algorithms. In Gillespie, T., Boczkowski, P., & Foot, K. (eds.) *Media Technologies: Essays on Communication, Materiality, and Society.* Cambridge, MA: MIT Press.

Gillespie, T. (2016). Algorithm. In Peters, B. (ed.) *Digital Keywords: A Vocabulary of Information Society and Culture*. Princeton, N.J.: Princeton University Press.

Glover, E. (2023, June 27). AI Has a huge climate change problem. *Builtin*. Retrieved from https://builtin.com/artificial-intelligence/ai-climate-change-dilemma

Glynn, P. (2023). Sony World Photography Award 2023: Winner refuses award after revealing AI creation. *BBC News*. Retrieved from https://www.bbc.com/news/entertainment-arts-65296763

Google AI, (2023) *Exploring 6 AI Myths*, https://ai.google/static/documents/exploring-6-myths.pdf

Government of Canada, (2023 March 13). Artificial Intelligence and Data Act. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document

Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2023). A survey of Generative AI Applications. https://doi.org/10.48550/ARXIV.2306.02781

GPAI, (2022). Global Partnership on Artificial Intelligence, Responsible AI https://gpai.ai/projects/responsible-ai/

Grother, P., Ngan, M., & Hanaoka, K. (2019). Face recognition vendor test (fvrt): Part 3, demographic effects. Gaithersburg, MD: National Institute of Standards and Technology.

Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, *40*(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850

Hamid, O. H., Smith, N. L. & Barzanji, A. (2017). "Automation, per se, is not job elimination: How artificial intelligence forwards cooperative human-machine coexistence," *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*, Emden, Germany, pp. 899-904, doi: 10.1109/INDIN.2017.8104891.

Hargittai, E. (2007). The social, political, economic, and cultural dimensions of search engines: An introduction. *Journal of Computer-Mediated Communication, 12*(3), 769-777.

Hasan, H. R., & Salah, K. (2019). Combating Deepfake Videos Using Blockchain and Smart Contracts. *IEEE Access*, 7, 41596–41606. https://doi.org/10.1109/ACCESS.2019.2905689

Haugen, F. (2021 October 4). Statement. United States Senate Committee on Commerce, Science and Transportation, *Sub-Committee on Consumer Protection, Product Safety, and Data Security*. Whistleblower Aid. Retrieved from https://www.commerce.senate.gov/services/files/FC8A558E-824E-4914-BEDB-3A7B1190BD49

Haughey, L. (2023). OnlyFans star is hiring out a virtual AI CLONE to her fans so they can live out their 'wildest fantasies' with her - for a hefty price, *Dailymail.* Retrieved from https://www.dailymail.co.uk/sciencetech/article-12114259/OnlyFans-star-hiring-AI-CLONE-fans-live-wildest-fantasies-her.html

Hempel, J., (2018) Fei-Fei Li's Quest to Make AI Better for Humanity, Wired

Hern, A. (2023 March 22). Bard: How Google's chatbot gave me a comedy of errors. The Guardian. https://www.theguardian.com/technology/2023/mar/22/bard-how-googles-chatbot-gave-me-a-comedy-of-errors

Hickok, E. & Zhong,V., (2022).Value Systems, Context, and AI: A study to understand the role of values and context in National AI principles and the development of AI, *ArtEZ University of the Arts*: The Netherlands

Hsu, T. (2023 Aug 3).What can you do when AI lies about you. *The New York Times*. Retrieved from https://www.nytimes.com/2023/08/03/business/media/ai-defamation-lies-accuracy.html

https://www.wired.com/story/fei-fei-li-artificial-intelligence-humanity/

IEEE. (2019) *Ethically Aligned Design*, First ed. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead1e.pdf

Information Commission's Office, (2022 February 18), Explaining Decisions Made with AI The Alan Turing Institute. https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence-1-0.pdf

Internet Society, (2017). *Artificial Intelligence and Machine Learning: Policy Paper.* https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/

Introna, L. D., & Nissenbaum, H. (2000). Shaping the Web: Why the politics of search engines matters. *The information society, 16*(3), 169-185.

Introna, L., & Nissenbaum, H., (2009). "Facial Recognition Technology: A Survey of Policy and Implementation Issues," *Report of the Center for Catastrophe Preparedness and Response,* NYU. https://nissenbaum.tech.cornell.edu/papers/facial_recognition_report.pdf

Introna, L., & Wood, D. (2004). Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance & Society, 2*(2/3), 177-198.

James, C. (2023 August 6). The Other A.I.: Artificial Intimacy With Your Chatbot Friend. *The Wall Street Journal.* Retrieved from https://www.wsj.com/articles/when-you-and-ai-become-bffs-ecbcda1e

Jankowicsz, N., (2022). I shouldn't have to accept being in DeepFake Porn, *The Atlantic,* https://www.theatlantic.com/ideas/archive/2023/06/deepfake-porn-ai-misinformation/674475/

Japan, (2019) Social Principles of Human-Centric AI. https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf

Japan (The Ministry of Economy, Trade and Industry). (2021 July 9). AI Governance in Japan Ver. 1.1 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20210709_8.pdf

Japan (the Ministry of Economy, Trade, and Industry). (2022 January 28). Governance Guidelines for Implementation of AI Principles Ver. 1.1 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf

Japan, (2022 April 22).AI Strategy 2022. https://www8.cao.go.jp/cstp/ai/aistratagy2022en.pdf

Jasanoff, S. & Kim, S. (2015). *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power.* University of Chicago Press, https://press.uchicago.edu/ucp/books/book/chicago/D/bo20836025.html

Joyce, K., Smith-Doerr, L., Alegria, S., Bell, S., Cruz, T., Hoffman, S. G., Noble, S. U., & Shestakofsky, B. (2021). Toward a Sociology of Artificial Intelligence: A Call for

Research on Inequalities and Structural Change. *Socius*, 7. https://doi.org/10.1177/2378023121999581

Juhasz, A. (2011). *Learning From Youtube*. Cambridge: MIT Press.

Kantrowitz, A. (2016, March 23). Microsoft's new AI-powered chatbot mimics a 19-year-old American girl. Retrieved from https://www.buzzfeednews.com/article/alexkantrowitz/microsoft-introduces-tay-an-ai-powered-chatbot-it-hopes-will

Katz, Y. (2020). Artificial Whiteness: Politics and Ideology in Artificial Intelligence. United States: Columbia University Press. https://www.google.com.hk/books/edition/Artificial_Whiteness/eXF7zQEACAAJ?hl=en

Ki Chan, C. C., Kumar, V., Delaney, S., & Gochoo, M. (2020). Combating Deepfakes: Multi-LSTM and Blockchain as Proof of Authenticity for Digital Media. 2020 *IEEE / ITU International Conference on Artificial Intelligence for Good* (AI4G), 55–62. https://doi.org/10.1109/AI4G50087.2020.9311067

Lee, P. (2016, March 25). Learning from Tay's introduction. Retrieved from https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and machines*, *17*, 391-444.

Leslie, D. (2020). Understanding bias in facial recognition technologies. arXiv preprint arXiv:2010.07023.

Leurs, K. (2023). *Digital migration.* (Vols. 1-0). SAGE Publications Ltd, https://doi.org/10.4135/9781529783155

Liang, L.(2012). Shadow Libraries. *E-flux*, 37, https://www.e-flux.com/journal/37/61228/shadow-libraries/

Manovich, L. (1999). Database as Symbolic Form. *Convergence*, 5(2), 80–99. https://doi.org/10.1177/135485659900500206

Martínez, G., Watson, L., Reviriego, P., Hernández, J. A., Juarez, M., & Sarkar, R. (2023). Combining Generative Artificial Intelligence (AI) and the Internet: Heading towards Evolution or Degradation? https://doi.org/10.48550/ARXIV.2303.01255

McCurry, J. (2021). South Korean AI chatbot pulled from Facebook after hate speech towards minorities, *The Guardian.* https://www.theguardian.com/world/2021/jan/14/time-to-properly-socialise-hate-speech-ai-chatbot-pulled-from-facebook

Merler, M., Ratha, N., Feris, R. S., & Smith, J. R. (2019). Diversity in faces. arXiv preprint arXiv:1901.10436.

Morris, F. (2022). The Tinder Swindler. Netflix.

Nahon, Karine. 2016, "Where there is Social Media, there is Politics", in Bruns A., Skogerbo E., Christensen C., Larsson O.A. and Enli G.S., eds. *Routledge Companion to Social Media and Politics.* New York: Routledge: 39-55. http://ekarine.org/wp-admin/pub/Nahon_PoliticsSM.pdf

Naraharisetty, R. (2022, December 12). Generative AI Puts Us on the Brink of a Deepfakes Crisis. *The Swaddle.* https://theswaddle.com/generative-ai-puts-us-on-the-brink-of-a-deepfakes-crisis/

NEC. (2019). NEC Group AI and Human Rights Principles. https://www.nec.com/en/press/201904/images/0201-01-01.pdf

Nissenbaum, H. (2001). How computer systems embody values. *Computer, 34*(3), 120-119.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. https://doi.org/10.2307/j.ctt1pwt9w5

Noble, S., (2012). Missed Connections: What Search Engines Say about Women. *Bitch magazine, 12*(4): 37-41. https://safiyaunoble.files.wordpress.com /2012/03/54_search_engines.pdf

O'Neil, C. (2017). *Weapons of math destruction*. Penguin Books.

OECD, (2019 May 22) Principles for responsible stewardship of trustworthy AI, https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

Omena, J. J. (2017). Ways of Thinking Technicity I. *Digital Methods, Networks and Technicity*. Retrieved from https://thesocialplatforms.wordpress.com/2017/02/15/the-philosophers-of-technique/

OpenAI. (2018 April 9) OpenAI Charter. https://openai.com/charter

OpenAI (2023). Partnership with American Journal Project to support local news, *OpenAI.* https://openai.com/blog/partnership-with-american-journalism-project-to-support-local-news

OpenAI. (2023 April 5). Our approach to AI safety. https://openai.com/blog/our-approach-to-ai-safety

OpenAI. (2023 May 22). Governance of superintelligence. https://openai.com/blog/ governance-of-superintelligence#SamAltman

OpenAI. (2023). GPT-4 Technical Report. *OpenAI*. https://cdn.openai.com/papers/gpt-4.pdf

Ottenbacher,J., Clough, P., Bates, J., (2017) Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results, Conference: Computer and Human Interaction (CHI'17). DOI:10.1145/3025453.3025727

Pasquale, F. (2016). Two Narratives of Platform Capitalism. *Yale Law & Policy Review, 35*(1), 309–319. http://www.jstor.org/stable/26601713

Patil, U., & Chouragade, P. M. (2021). *Deepfake Video Authentication Based on Blockchain*. 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 1110–1113. https://doi.org/10.1109/ICESC51422.2021.9532725

Patil, U., Chouragade, P. M., & Ambhore, P. (2021). *An Effective Blockchain Technique to Resist Against Deepfake Videos*. 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 1646–1652. https://doi.org/10.1109/ICIRCA51532.2021.9544854

Paul, S., Joy, J. I., Sarker, S., Shakib, A.-A.-H., Ahmed, S., & Das, A. K. (2019). *Fake News Detection in Social Media using Blockchain*. 2019 7th International Conference on Smart Computing & Communications (ICSCC), 1–5. https://doi.org/10.1109/ICSCC.2019.8843597

Philip, K. (2007). What is a Technological Author? The pirate function and Intellectual Property. *Postcolonial Studies,* 8(2), https://doi.org/10.1080/13688790500153596

Podkul, C. (2022). Human Trafficking's Newest Abuse: Forcing Victims Into Cyberscamming. *ProPublica*, https://www.propublica.org/article/human-traffickers-force-victims-into-cyberscamming.

Ramaswamy, S. (2023, March 19). 5 Real Life Examples of Chat GPT 4's Truly Multimodal Capabilities. https://akaike.ai/5-real-life-examples-of-chat-gpt-4s-truly-multimodal-capabilities/

Ren, D., (2023) AI technologies will leave 800,000 Hongkongers out of work or looking for new job by 2028, says recruiter, China Business, *South China Morning Post*, https://www.scmp.com/business/china-business/article/3224078/ai-technologies-will-leave-800000-hongkongers-out-work-or-looking-new-job-2028-says-recruiter

Ridgway, R. (2023). Deleterious consequences: How Google's original sociotechnical affordances ultimately shaped 'trusted users' in surveillance capitalism. *Big Data & Society, 10(1)*. https://doi.org/10.1177/20539517231171058

Rieder, B., Peeters, S., & Borra, E. (2022). From tool to tool-making: Reflections on

Rifkin, J. (1995). The End of Work: The Decline of the Global Labor Force and the Dawn of the Post-Market Era. New York: Tarcher/Putnam Publishing Group.

Robins-Early, N. (2023, July 19). Disinformation reimagined: How AI could erode democracy in the 2024 US elections. *The Guardian*. https://www.theguardian.com/us-news/2023/jul/19/ai-generated-disinformation-us-elections

Roose, K. (2023). A conversation with Bing's chatbot left me deeply unsettled, The New York Times https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

Rose, M. (1993). *Authors and Owners: The Invention of Copyright*. Cambridge: Harvard University Press.

Röttger, P., Vidgen, B., Hovy, D., & Pierrehumbert, J. B. (2021). Two contrasting data annotation paradigms for subjective NLP tasks. arXiv preprint arXiv:2112.07475.

Roy, A. & Cusack, J. (2016). *Things that Can and Cannot be Said: Essays and Conversations*. Chicago: Haymarket Books.

Saguira, L. (2021). *Incel Rebellion: The Rise of the Manosphere and the Virtual War against Women*, USA: Emerald Publishing.

Sajja, P.S. (2021). Introduction to Artificial Intelligence. In: Illustrated Computational Intelligence. Studies in Computational Intelligence, vol 931. Springer, Singapore. https://doi.org/10.1007/978-981-15-9589-9_1

Savelka, J., Ashley, K. D., Gray, M. A., Westermann, H., & Xu, H. (2023). Explaining Legal Concepts with Augmented Large Language Models (GPT-4) (arXiv:2306.09525). arXiv. http://arxiv.org/abs/2306.09525

Seaver, N. (2019). Knowing Algorithms. In Vertesi, J., & Ribes, D. (eds.) *digitalSTS: A field guide for science & technology studies.* Princeton Univerity Press.

Sedgwick, E., (1997). "Paranoid Reading and Reparative Reading; or, You're So Paranoid, You Probably Think This Introduction is About You", Novel Gazing: Queer Readings in Fiction, Book Chapter, Duke University Press. https://read.dukeupress.edu/books/book/636/chapter/128566/Paranoid-Reading-and-Reparative-Reading-or-You-re

Selinger, E., & Hartzog, W. (2016). Facebook's emotional contagion study and the ethical problem of co-opted identity in mediated environments where users lack control. *Research Ethics,* 12(1), 35–43. https://doi.org/10.1177/1747016115579531. Retrieved from https://journals.sagepub.com/doi/full/10.1177/1747016115579531

Shah, N. (2022). *Formulating Fake Futures*, ArtEZ Press Essay Issues, The Netherlands: ArtEZ Press.

Shah, N. (2022). Refusing Platform Promises: Rewriting Digital Imaginaries of Noncanonical Bodies, Yale University.https://macmillan.yale.edu/event/refusing-platform-promises-rewriting-digital-imaginaries-noncanonical-bodies

Shigeru, S., Shunsuke, S. Toru, Y. (2019) NEC Group AI and Human Rights Principles, NEC Technical Journal Vol. 14. https://www.nec.com/en/global/techrep /journal/g19 /n01/190103.html

Sony. (2018 September 25). AI Engagement within Sony Group. https://www.sony.com/en/SonyInfo/csr_report/humanrights/AI_Engagement_within_Sony_Group.pdf

Suen, D., (2023) Hong Kong marks first authorised protest rally in 3 years under strict police rules, Politics, South China Morning Post. https://www.scmp.com/video/hong-kong/3214916/hong-kong-marks-first-authorised-protest-rally-3-years-under-strict-police-rules

Swan, M. (2015). *Blockchain: Blueprint for a new economy* (First edition). O'Reilly.

Tencent. (2019 July 8). Technology Ethics in the Intelligent Era——Reshaping Trust in Digital Society. https://mp.weixin.qq.com/s/NOGtj5y_5ADmSk0TijdOgQ

Tencent Research Institute, (2020), "ARCC": An Ethical Framework for Artificial Intelligence. https://www.tisi.org/13747

Tencent. (2020 July 10). 2020 Tencent AI whitepaper. https://www.tisi.org/16624

Tencent. (2021 November 4). Tencent Clould+AI Whitepaper. https://docs.qq.com/pdf /DSkN1c0dsbERzRExO

Tencent. (2023 January 31). AI generated content development trends report 203. https://docs.qq.com/pdf/DSkJweFlIdEFMQ2pT?

Tencent, (2023). Explainable artificial intelligence report. https://cloud.tencent.com/developer/article/1955180

Thiroux, J. P., & Krasemann, K. W. (2015). *Ethics: Theory and Practice* (11th ed.). Pearson.

Thompson, N & Bremmer, I. (2018, October 23). The AI Cold War That Threatens Us All. Wired. Retrieved from https://www.wired.com/story/ai-cold-war-china-could-doom-us-all/

Totschnig W. (2020) Fully Autonomous AI. Sci Eng Ethics. 2020 Oct;26(5):2473-2485. doi: 10.1007/s11948-020-00243-z. PMID: 32725298.

TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management . (2022 December 1), Tangible global leadership by the United States and the European Union. https://www.nist.gov/system/files/documents/2022/12/04/Joint_TTC_Roadmap_Dec2022_Final.pdf

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other.* Basic Books.

U.S. Government (2020 November 17). Guidance for Regulation of Artificial Intelligence Applications. https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf

U.S. Government, (2022 February 3), Algorithmic Accountability Act of 2022. Text-S.3572-117th Congress (2021-2022). https://www.congress.gov/bill/117th-congress/senate-bill/3572/text

Ulloa, R., Richter, A. C., Makhortykh, M., Urman, A., & Kacperski, C. S. (2022). Representativeness and face-ism: Gender bias in image search. *New media & society*, 14614448221100699.

UN, (2019) G20 AI Principles, OECD AI Policy Observatory https://oecd.ai/en/wonk/documents/g20-ai-principles

UNESDOC Digital Library. (2021) Draft text of the Recommendation on the Ethics of Artificial Intelligence with track-changes (paragraphes 26-134; outcome of the intersessional consultations), Intergovernmental Meeting of Experts (Category II) related to a Draft Recommendation on the Ethics of Artificial Intelligence, online, 2021 [65], UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000377898.

UN. (2023). Our Common Agenda Policy Brief 8—Information Integrity on Digital Platforms. https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-information-integrity-en.pdf

van Couvering, E. (2007). Is relevance relevant? Market, science, and war: Discourses of search engine quality. *Journal of Computer-Mediated Communication, 12*(3), 866-887.

Van den Hoven, J., Lokhorst, GJ. & Van de Poel, I. (2012). Engineering and the Problem of Moral Overload. *Sci Eng Ethics* **18**, 143–155. https://doi.org/10.1007/s11948-011-9277-z

van Dijck, J. (2013). Facebook and the engineering of connectivity: A multi-layered approach to social media platforms. *Convergence, 19*(2), 141–155. https://doi.org/10.1177/1354856512457548

van Dijck, J. (2013). Facebook and the engineering of connectivity: A multi-layered approach to social media platforms. *Convergence, 19*(2), 141-155.

van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world.* Oxford University Press.

Vaughan, L., & Zhang, Y. (2007). Equal representation by search engines? A comparison of websites across countries and domains. *Journal of computer-mediated communication, 12*(3), 888-909.

Wallach, W., & Vallor, S. (2020). Moral machines: From value alignment to embodied virtue. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 383-412). Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0014

Watts, J. D. (2003). *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton Studies in Complexity, Princeton University Press. ISBN 9780691117041

Wehsener, A., (2020 May 27). Digital Threats to Democracy: Comfortably Numb, Technology For Global Security, Centre for a New American Security. https://www.cnas.org/publications/commentary/digital-threats-to-democracy-comfortably-numb

Weise, K. & Metz, C. (2023 May 1). When AI Chatbots Hallucinate, *The New York Times*. https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html

Yazdinejad, A., Parizi, R. M., Srivastava, G., & Dehghantanha, A. (2020). Making Sense of Blockchain for AI Deepfakes Technology. 2020 *IEEE Globecom Workshops* (GC Wkshps, 1–6. https://doi.org/10.1109/GCWkshps50303.2020.9367545

Yerushalmy, J. (2023 February 17). ‘I want to destroy whatever I want’: Bing’s AI chatbot unsettles US reporter, *The Guardian*. https://www.theguardian.com/ technology/2023 /feb/17/i-want-to-destroy-whatever-i-want-bings-ai-chatbot-unsettles-us-reporter

Yuan, L. et al., In situ bidirectional human-robot value alignment. *Sci. Robot.* **7**, eabm4183 (2022).

IX. Appendix: List of Influential AI Policy and Governance Documents

2021.11	Recommendation on the ethics of artificial intelligence	UNESCO	https://www.unesco.org/en/artificial-intelligence/recommendation-ethics#:~:text=AI%20actors%20should%20promote%20social,benefits%20are%20accessible%20to%20all.
2019.6.9	G20 AI principles	UN	https://oecd.ai/en/wonk/documents/g20-ai-principles
2019.4.8	Ethics Guidelines for Trustworthy AI	EU	https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
2021.4.21	LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS	EU	https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206
2023.5.11	AI Act	European Commission	https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence#:~:text=The%20new%20law%20promotes%20regulatory,that%20significantly%20impact%20their%20rights.
2019.5.22	Principles for responsible stewardship of trustworthy AI	OECD	https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449
2022.12.1	TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management	Tangible global leadership by the United States and the European Union	https://www.nist.gov/system/files/documents/2022/12/04/Joint_TTC_Roadmap_Dec2022_Final.pdf
2017.12.12	FUTURE of Artificial Intelligence Act of 2017	U.S. Congress	https://www.congress.gov/bill/115th-congress/house-bill/4625/text

2020.11.17	Guidance for Regulation of Artificial Intelligence Applications	U.S. Government	https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf
2022.2	Algorithmic Accountability Act of 2022	U.S. Government	https://www.congress.gov/bill/117th-congress/senate-bill/3572/text
2023.03.13	Artificial Intelligence and Data Act	Canada	https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document
2022.2.18	Explaining Decisions Made with AI	UK (Information Commissioner's Office)	https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence-1-0.pdf
2019.6.17	新一代人工智能治理原则——发展负责任的人工智能 (Next Generation AI Governance Principles - Developing Responsible AI)	China Government	https://www.most.gov.cn/kjbgz/201906/t20190617_147107.html
2021.7.9	AI Governance in Japan Ver. 1.1	Japan (the Ministry of Economy, Trade, and Industry)	https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20210709_8.pdf
2022.4.22	AI Strategy 2022	Japan	https://www8.cao.go.jp/cstp/ai/aistratagy2022en.pdf
2019	Social Principles of Human-Centric AI	Japan	https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf
2022.1.28	Governance Guidelines for Implementation of AI Principles Ver. 1.1	Japan (the Ministry of Economy, Trade, and Industry)	https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf
2017.4.18	Artificial Intelligence and Machine Learning: Policy Paper	Internet Society	https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/
2017	"Ethically Aligned Design" V2	IEEE	https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
2017	Asilomar AI Principles	Future of Life Institute	https://futureoflife.org/open-letter/ai-principles/
2020.4.9	“ARCC”: An Ethical Framework	Tencent	https://www.tisi.org/13747

	for Artificial Intelligence		
2023	可解释人工智能发展报告(Explainable artificial intelligence report)	Tencent	https://cloud.tencent.com/developer/article/1955180
2023.1.31	AIGC发展趋势报告2023(AI generated content development trends report 203)	Tencent	https://docs.qq.com/pdf/DSkJweFlIdEFMQ2pT?
2021.11.4	上云赋智： 2021 云上智能白皮书 (Tencent Clould+AI Whitepaper)	Tencent	https://docs.qq.com/pdf/DSkN1c0dsbERzRExO
2020.7.10	腾讯人工智能白皮书：泛在智能 (2020 Tencent AI whitepaper)	Tencent	https://www.tisi.org/16624
2019.7.8	智能时代的技术伦理观——重塑数字社会的信任 (Technology Ethics in the Intelligent Era——Reshaping Trust in Digital Society)	Tencent	https://mp.weixin.qq.com/s/NOGtj5y_5ADmSk0TijdOgQ
2018.9.25	AI Engagement within Sony Group	Sony	https://www.sony.com/en/SonyInfo/csr_report/humanrights/AI_Engagement_within_Sony_Group.pdf
	AI Ethics	Fujitse	https://www.fujitsu.com/global/about/research/technology/aiethics/
2019.4	NEC Group AI and Human Rights Principles	NEC	https://www.nec.com/en/press/201904/images/0201-01-01.pdf
2018.4.9	OpenAI Charter	OpenAI	https://openai.com/charter
2023.4.5	Our approach to AI safety	OpenAI	https://openai.com/blog/our-approach-to-ai-safety
2023.5.22	Governance of superintelligence	OpenAI	https://openai.com/blog/governance-of-superintelligence#SamAltman