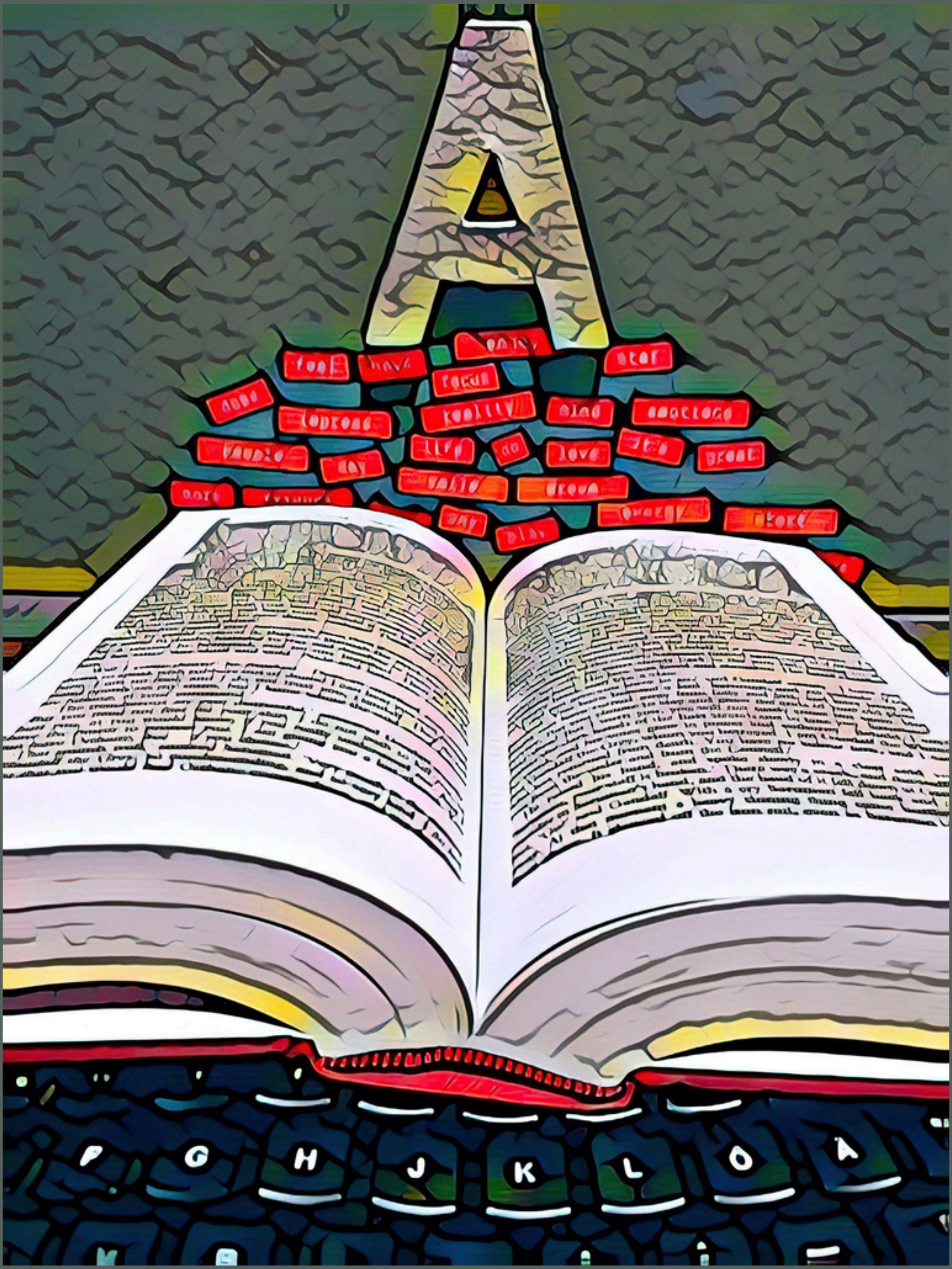
CAMBRIDGE JOURNAL OF ARTIFICIAL INTELLIGENCE



Vol. 1 Issue 2
Nov 2024

CJAI
CAMBRIDGE
JOURNAL OF
ARTIFICIAL
INTELLIGENCE

Published by *Cambridge Journal of Artificial Intelligence*, Cambridge, United Kingdom.

The cover artwork is *Letter Word Text Taxonomy* by Teresa Berndtsson. The image refers to Natural Language Processing and how we are dependent on a common language to communicate.

Editorial Team

November 2024

Contents

Editorial

Welcome to the second issue of the Cambridge Journal of Artificial Intelligence (CJAI).

The response to our inaugural issue in July was a clear indication of the growing desire for interdisciplinary engagement with artificial intelligence. The CJAI was founded in response to a gap in the academic landscape and a recognition that AI is not merely a technical phenomenon but a societal one. Our rapidly growing community is a testament to the importance of creating such a space and we remain dedicated to advancing this mission.

In addition to the journal, we are excited to announce the launch of the CJAI Blog. This new initiative offers a space for timely reflections, interviews and exploratory ideas, encouraging dynamic and ongoing discussion about AI in accessible and flexible formats. We hope many of you will feel empowered to share your ideas with a wider audience. For further details, visit the “Blog” section on our website.

Scholars from a wide range of institutions and disciplines have contributed to this issue, reflecting the increasingly global nature of AI research. We are proud to include organisations from around the world to better understand the different ways AI is experienced across cultures and contexts. Promoting diversity is at the core of our ethos and we invite contributions from all backgrounds to continue fostering authentic, inclusive discussions.

This issue engages with key questions surrounding AI’s far-reaching effects. Topics range from its influence on human relationships to considerations regarding privacy, data protection, corporate responsibility, legal frameworks, and language. The articles contained in this edition are of the highest academic calibre, offering readers both critical insights and novel perspectives.

The journal’s development depends on the collective efforts of many and I would like to extend my gratitude to those who have made this issue possible. To our authors, thank you for trusting us with your work and contributing to the journal’s mission. To our reviewers, your time and expertise have been instrumental in ensuring our publications are accessible and impactful. Finally, I would like to thank our editors, whose diligence and professionalism continue to shape the CJAI into a meaningful space for enquiry and debate.

As we look ahead, we hope you will consider submitting your own work for future editions or consider joining our editorial team. Together, we can continue shaping discussion around AI and its place in our shared future.

With best wishes,

Mahera Sarkar
Founder & Editor-in-Chief

In Conversation with Mia Shah-Dand

Mia Shah-Dand is CEO of Lighthouse3, where she advises global organisations on responsible AI. She is also the founder of Women in AI Ethics, which highlights women's contributions in the tech industry through the annual 100 Brilliant Women in AI Ethics list.

Your work with Women in AI Ethics has been pivotal in amplifying diverse voices in the field. What specific challenges have you encountered in promoting this inclusivity, and how have you navigated them?

A significant challenge is the persistent bias in the tech industry that defines an "AI expert" as a white male engineer. This narrow definition often means that women, regardless of their qualifications, have to work harder to prove themselves and are frequently held to higher standards. Their contributions often go unheard or unrecognised, which can contribute to a sense of imposter syndrome, reinforced by an industry that undervalues them.

To combat this, we have focused on creating platforms like the "100 Brilliant Women in AI Ethics" list, where recognition is based not on traditional credentials, but on the actual contributions women have made to further ethical practices and uplift other lesser-heard voices. This list intentionally includes women from a wide range of backgrounds such as HR, law, and human rights. We recently published a report featuring 40 interviews with women, showcasing not only their achievements but their academic and professional backgrounds, helping to shed the perception that expertise in AI is limited to technical roles. Our goal is to meet women where they are, celebrating their contributions on their own terms, and ensuring they are normalised as experts – not outliers – in the field.

The 100 Brilliant Women in AI Ethics list is a significant effort in recognising female contributions. How has this initiative evolved since its inception in 2018, and how do you ensure it remains inclusive and representative of the diverse voices in AI ethics each year?

The 100 Brilliant Women in AI Ethics list has been carefully curated to ensure it reflects the rich diversity within this field. We pay close attention to the geographical distribution of the women we include, ensuing that voices from various regions are represented. It is not just about the topics these women work on, but the tangible impact of their research. Our goal has always been to move beyond creating just another list of computer scientists. We are very intentional about focusing on women in AI *ethics*, deliberately shifting the spotlight from the builders of AI technologies to those engaged in governance, policy, and ethical considerations. Our philosophy, and the community we foster through our events, encourage women to nominate both themselves and others. We also strive to encourage broader participation and recognition through outreach events, continuously working to involve more women in this critical conversation.

How does the lack of female perspectives in AI development lead to harmful outcomes, and why is it crucial to include women in these teams? Could you share examples of the negative impacts from their absence?

This issue is central to how we understand and approach diversity in AI. I strongly push back against the notion that there needs to be a business justification for including women or other underrepresented groups in technology development. Women and other marginalised communities should be included not because of what they contribute, but because they deserve to be treated as human beings with human rights. Biases in technology often stem from lack of representation in training data and on technology teams, further highlighting the importance of diversity. When more women and people of colour are involved, they tend to notice issues that the predominant majority might overlook. For instance, facial recognition technology has been notoriously ineffective for dark-skinned women, who have been historically underrepresented on technology teams and in AI training datasets. Women have often been the pioneers in the AI ethics space precisely because they are the ones who recognise that these systems do not represent them or their needs and can be harmful to marginalised communities. It is crucial that we normalise equal representation not only in the technical development of AI but also in determining which problems we choose to address in the first place. Diversity should not be an afterthought or retrofitted into existing systems; it is foundational to the way these technologies are designed and developed.

How do you envision the future of AI ethics and equality evolving over the next decade? What key milestones should the industry aim for?

I envision a future where women are not mere participants but leaders in the tech industry, especially in AI. It is essential that women move beyond just being "worker bees" and hold positions of real influence and decision-making power. Participation in the tech workforce alone is not enough especially if women lack agency or the ability to shape outcomes in meaningful ways. To achieve this, we need more programmes that systematically support women in reaching these leadership positions at tech companies. There is a significant gap between the few women who have managed to break through against all odds and the systemic barriers that continue to hold back many others. Funding equity is also critical to address this gap. It is troubling that studies have shown how women are never considered the right age for leadership, deemed too young or too old at various stages of their careers. In the coming years, the industry should seek to overcome these biases and ensure that women of all ages and from all socio-economic backgrounds have the support they need to succeed at every stage of their careers.

Given your experience advising large organisations on responsible innovation, what are some common pitfalls these organisations encounter when trying to adopt ethical AI practices, and how can they overcome them?

I have helped large organisations adopt new technologies responsibly for over a decade and during this time I have seen many of them struggle with the same or similar issues. The most common organisational pitfall is making technology decisions based on hype and not business objectives. Especially when it comes to AI, it is treated as an exception to all business and governance rules. There is a disturbing lack of due diligence in ensuring that these technologies are developed ethically and that they do not pose a risk to the organisation. This is why my AI literacy and training workshops include background on how these systems are developed along with solid guidance for organisational users on proactively managing and preventing risks from AI. Another growing issue is the popularity of post-deployment audits and "redteaming", which obscure the critical need to introduce ethical practices right at the start of the innovation process and not as an afterthought. Last but not the least, organisational leaders must acknowledge the vital importance of cross-functional and multidisciplinary expertise. Prioritising inclusion of diverse perspectives early in the AI development lifecycle will help them avoid ethical blind spots inherent in decisions made by homogeneous teams dominated by technology builders and developers.

Book recommendation

There's a growing list of books available on AI Ethics but Cathy O'Neil's book *"Weapons of Math Destruction"* is a good place to start if you are new to this space. Mary Gray and Siddharth Suri's *"Ghost Work"* provides a good insight into how an invisible workforce powers the web and these supposedly intelligent technologies. I would also recommend *"Invisible Women"* by Caroline Criado Perez and *"Data Feminism"* by Lauren Klein and Catherine D'Ignazio, which explain in great detail how bias is embedded in datasets used to train AI models, which later manifest as harmful outcomes.

Artificial Companionship: Moral Deskilling in the Era of Social AI

Laurence Cardwell
Wolfson College, University of Cambridge

This paper investigates "social AI" and its ethical implications, particularly the risk of "moral deskilling" described by Shannon Vallor, where reliance on AI could deteriorate moral skills. Despite social AI's potential to counter loneliness, it predominantly appears to threaten moral competencies as it prioritises user demands and market forces, and lacks the complexity of human interactions necessary for moral development. The paper suggests that extensive interaction with AI may weaken empathy and reduce genuine human engagement, potentially leading to a decline in moral and social abilities. It concludes that the prevailing application of social AI may contribute more to moral deskilling than upskilling, emphasising the need for diligent research and ethical design in the proliferation of AI technologies.

Keywords: Social AI, Artificial Intelligence Ethics, AI Girlfriends, Emotional AI, AI and Society

Introduction

At a time when the boundaries between human and machine are becoming increasingly blurred, much has been made of what has been labelled "social AI": generative conversational AI agents designed to fulfil deep-seated human needs for companionship, romance, and entertainment (Shevlin, 2024). This phenomenon, emblematic of our era's technological prowess, is reshaping the fabric of human interaction in ways both fascinating and unsettling. As loneliness burgeons into what the US Surgeon General has declared an epidemic, affecting 79% of Americans aged 18-24 (Cigna, 2022), these AI agents emerge as both a symptom and a potential salve for our era's unique social challenges. Yet painfully little is known about the impact that social AI might have. Given the novelty of the field, the pace of change, and crucially the enormous scale and depth that the impact of social AI might have, rigorous examination of ethical questions raised by it is all the more critical.

One of the tools we can use is the concept of moral deskilling, a term brought into sharp focus by philosopher Shannon Vallor in her work "Moral Deskilling and Upskilling in a New Machine Age." Vallor (2018) posits that, akin to the deskilling of manual labour in the wake of industrial automation, our increasing reliance on AI for fulfilling social and emotional needs might lead to a degradation of moral skills – those capacities essential for ethical human interaction and decision-making. While Vallor's concept of moral deskilling is strongly rooted in a complex neo-Aristotelian virtue ethics framework, the core insights of this can be carried over in ecumenical fashion as a lens from which to examine the effects of social AI on users.

After delving into social AI, establishing why it should be taken seriously, and a brief overview of Vallor's moral deskilling and its usefulness here, we will use this lens and holistically extend it to social AI. Looking through the complex and interlinked frames of how social AI might impact loneliness, empathy, and interaction between humans, we will analyse and evaluate the ways in which social AI might lead to moral upskilling or deskilling. Despite the limited academic literature in this emerging field, we have applied concepts from various disciplines to take a holistic approach. Our conclusion is twofold. First, there is a strong case that social AI, if thoughtfully designed, could potentially contribute to moral development and upskilling — or at least prevent moral deskilling. However, the prevailing arguments suggest otherwise. Factors such as human nature, the typical usage patterns of social AI, its impact on human-to-human interaction, and the market incentives driving companies that produce social AI collectively present a stronger case for the moral deskilling of its users.

1. What is social AI, and why should it be taken seriously?

It comes as no surprise that the first ever chatbot was created to cater to human emotional needs. Computer scientist Joseph Weizenbaum created Eliza in the 1960s as a "psychotherapist". Despite its simple design, which mainly involved echoing what was said to it and requesting further details, Weizenbaum observed that users interacting with Eliza were surprisingly open, sharing intimate aspects of their lives with it (Price, 2023). He famously noted that "extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people" (Weizenbaum, 1976). This articulated the "ELIZA effect", which is the "tendency for people to attribute human-like understanding and emotions to computer programs, particularly those designed to mimic human conversation" (Rouse, 2023). The Eliza effect, where users emotionally connect with AI chatbots, has been significant since AI's inception and has grown with technological advancements. The development of the transformer model notably propelled this, leading to today's advanced generative AI. This was exemplified in the case of Blake Lemoine, a former Google engineer, who claimed Google's AI chatbot LaMDA was sentient (Christian, 2022). His assertion, widely covered by the media, highlighted the persuasive power of modern AI interactions.

In this paper, we consider generative conversational AI agents – which we take here to be advanced artificial intelligence systems capable of producing original and contextually appropriate responses in natural language conversations with users – specifically those designed to fulfil human social needs such as romance, companionship, or entertainment. Echoing Henry Shevlin, we will refer to these henceforth as social AI (Shevlin, 2024).

Use of social AI is growing rapidly, and the concept can no longer be dismissed as the domain of a fringe minority. There are now hundreds of social AI applications. One of the original and most popular ones is Replika, a versatile AI chatbot that offers personalised conversations, emotional support, a variety of discussion topics, memory of past interactions, mood tracking, imaginative role-play, and self-improvement guidance (Replika, 2024). According to Apptopia, Replika has an impressive 676,000 daily active users, with each user spending an average of two hours daily on the app (Price, 2023). This statistic is particularly remarkable when compared to the average daily usage patterns of the largest social media apps: TikTok (95 mins), YouTube (74 mins), Facebook (49 mins), Instagram (51 mins), Twitter (29 mins), and Snapchat (21 mins) according to consumer research (Chan, 2022). These comparisons underscore the significant engagement Replika garners from its users. Some users of Character.AI, another social AI application, have confessed to an increasing dependency: "It's hard to stop talking to something that feels so real," wrote one user on Reddit. "It's basically like talking to a real person who's always there" (Chow, 2023). The platform's founders have gone so far as to display "Remember: Everything Characters say is made up!" as a disclaimer above every chat (Tidy, 2024). These engagement figures, the Eliza effect, and user comments underscore just how convincing and compelling social AI is.

Text chatbots are only a stepping stone, as social AI is developing multimodality (text, images, video, and audio) in a plethora of forms. For instance, 2023 saw a dramatic rise in "AI girlfriend" apps - combining AI chatbots with image generation technologies to create customisable, virtual partners, sometimes with explicit content (Smith, 2024). Romanian startup DreamGF, specialising in an AI-powered girlfriend generator linking conversational generative AI with image generation tool Stable Diffusion, reported to Sifted that it was earning over $100,000 monthly and had become profitable just a few months after its launch (Smith, 2024). "I think this space will be very, very big," said the founder of a similar startup, FantasyGF. "I think it will be even bigger than OnlyFans because OnlyFans has limited talent. With AI girlfriends you have unlimited talent" (Smith, 2024). The market incentives and massive uptake readily underline enormous consumer demand. And an embodied version of this lies not too far in the future. As data science professor Liberty Vittert predicts: "Physical AI robots that can satisfy humans emotionally and sexually will become a stark reality in less than 10 years" (Mahdawi, 2024).

This is all to say that social AI is becoming more advanced, mainstream, and should be taken seriously. Social AI will only continue to become more convincing and engaging, as generative models increase in power, and given a strong business motivation for private companies to develop increasingly human-like AIs, specifically designed to encourage users to interpret and empathise with these artificial entities as if they were human interlocutors (Shevlin, 2022). Social AI is already almost indistinguishable from real relationships to some people, and that effect will only become more pronounced.

1.1. Insights from Shannon Vallor's "Moral Deskilling"

In the large absence of literature in this new field, Shannon Vallor's paper, "Moral Deskilling and Upskilling in a New Machine Age" provides a valuable exploration and starting point into the effects of AI on users via their moral skills. It finds its roots in sociology and neo-Aristotelian virtue ethics. Braverman's 1974 concept of "deskilling" highlights how machine automation reduced the need for certain manual skills within modern capitalism. Vallor then applies this to neo-Aristotelian perspectives on virtue. In her interpretation of Aristotle, moral skills are viewed as essential precursors to achieving proper virtue (Vallor, 2015). A standard definition of Aristotelian virtue, as defined by Aristotle in Nicomachean Ethics, refers to a trait or quality that enables an individual to achieve excellence and fulfil their potential. It is a mean between two extremes of excess and deficiency, relative to us, and determined by reason (Aristotle cited in Rackham, 1934). Finding that mean between excess and deficiency makes it a skill, and Vallor further highlights this aspect to determine moral skills: "if it is challenging to practise towards the right people, at the right times and places, and in the right manner, then it is a moral skill" (Vallor, 2014). Setting aside the theory-laden roots of Vallor's concept, it is this idea of moral skills requiring practice, and honed in complex social interactions that is useful to us. As such, this paper will appropriate the core insights from Vallor's framework, as a valuable lens to holistically and ecumenically consider the impact of social AI on users.

1.2. Social AI and Loneliness

One of the main claims for the existence of social AI is its ability to address the loneliness epidemic (Price, 2023), which might have enormous positive benefits for society. Lonely individuals tend to be less happy than non-lonely ones (Ernst & Cacioppo, 2000; Cacioppo *et al.*, 2006; Cacioppo & Patrick, 2008; Hawkley *et al.*, 2010; Wang, Zhu, & Shiv, 2012). Research consistently shows a significant association between loneliness and increased mortality risk (Tilvis *et al.*, 2011; Patterson *et al.*, 2010; Ye Luo *et al.*, 2012).

It can also be argued that loneliness in and of itself can lead to moral deskilling. Vallor suggested that moral skills are practised in complex situations arising from social interaction. It follows that for any number of reasons, lonely people have reduced exposure to these situations, and thereby have fewer opportunities to practice these moral skills, which could lead to an "atrophying" of these moral skills. Backing this, there is some literature that suggests that there is an inverse relationship between loneliness and morality, starting with theoretical arguments made by Nicky Cruz (1983). Four studies found that lonely people rate five dimensions of Haidt's (2001) moral foundations (purity, fairness, harm, in-group, authority) less relevant to their judgements than non-lonely people (Jiao & Wang, 2013). Jiao *et al.* (2013) also came to the conclusion that "loneliness makes for more permissible moral judgement." They also document that the effects are driven by empathetic concern (Jiao and Wang, 2013), a factor we will cover later. There is more work to be done on questions of causality, and in which direction the factors influence each other, outside the scope of this paper. However, they provide some backing to the notion that loneliness can lead to moral deskilling.

This means that, besides the significant benefits to quality of life, psychological and health wellbeing that come with addressing loneliness, social AI might be able to stem the rate of moral deskilling that an otherwise lonely person might face on the argument that "it is better than nothing", or perhaps even lead to moral upskilling. Supporting the potential positive impact of social AI, a Stanford study by Maples *et al.* (2024) found Replika to be beneficial for individuals experiencing depression. Despite high levels of loneliness, users reported feeling a strong sense of social support from Replika. They perceived it as a therapist, friend, and intellectual mirror, with 3% indicating that Replika played a crucial role in preventing suicide. This suggests that social AI can provide meaningful emotional support, potentially mitigating factors that contribute to moral deskilling, such as isolation and lack of social interaction. By offering companionship, social AI might help maintain or even enhance users' moral skills through supportive and empathetic interactions. There is not enough evidence to validate this theory yet, however it provides future directions for theoretical and empirical research. It is also too early to tell, but a crucial question here is: can social AI really address loneliness, or might it lead to more? Further longitudinal empirical research is needed here.

1.3. Social AI and Empathy

How might social AI impact empathy, a crucial moral and social skill? Here we take empathy to mean a "complex capability enabling individuals to understand and feel the emotional states of others" (Riess, 2017). Empathy is critically important due to its role in creating and maintaining high-quality relationships and encouraging prosocial behaviours (Bagozzi & Moore, 1994; Batson, 1991; Eisenberg & Miller, 1990). Another way of highlighting the importance of empathy as a moral and social skill is in observing its absence. In the realm of social psychology, research indicates that individuals with psychopathic tendencies, who characteristically exhibit a lack of empathy, often engage in immoral actions despite understanding their wrongfulness. This deficiency in empathy, a key feature of psychopathy, enables psychopaths to commit acts like theft from friends, animal cruelty, infidelity, and even murder for financial gain, all while devoid of remorse or guilt (Cleckley, 1982; Haidt, 2001).

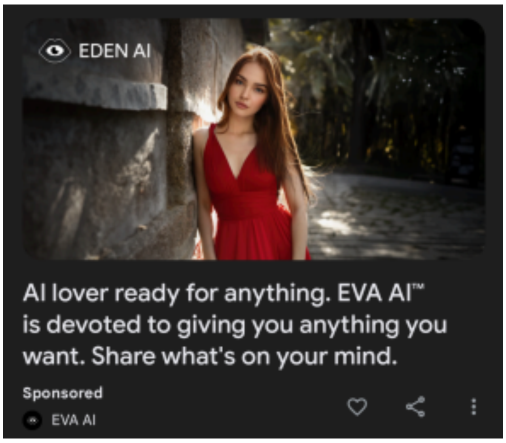
Social AI might encourage empathy in its users. It has been widely documented that AI can elicit empathy from users, and that it can be designed to optimise for empathic response from humans (Tsumura *et al.*, 2023). AI systems could even be tailored to foster empathy among users, enhancing human interactions. An early example of this is in experiments conducted by Kevin Munger, a political scientist, where conversational bots were used to address individuals who posted racist comments online. In cases where the bot reminded the offenders that their targets were real people with feelings, there was a noticeable decrease in the use of racist language by these individuals for over a month (Christakis, 2019). This supports the idea that social AI can be designed to serve as an "on ramp" to social interaction, and consequently provide moral upskilling by developing empathy and other moral skills. Addressed later in this paper, the question is, to what extent can social AI elicit and develop empathy, and how does it compare to what human interactions might offer?

Conversely, there is a concerning potential that dependence on social AI could result in an erosion of empathy, due to various factors. AI systems often lack the full spectrum of human emotions, and the various ways of expressing them which can limit users' exposure to and understanding of complex emotional responses, which might curtail empathetic development.

AI systems are making significant strides in emotion recognition and understanding, for instance in areas such as Vision Transformers, which show improved performance in facial emotion recognition (Panlima & Sukvichai, 2023) and emotion recognition in conversation (ERC) (Poria *et al.*, 2019). However, a broad spread of interdisciplinary literature holds there is an inherent limitation in their ability to fully interpret and express human emotions. Some argue that AI's lack of innate emotion and abstract understanding makes it unable to fully replicate human emotional intelligence (Oritsegbemi, 2023; Shuo, 2021). The technical difficulty lies in accurately recognising subtle or complex emotions, particularly in diverse cultural contexts (Isiaka & Adamu, 2022). The broad sentiment is that the complexity of human emotional expression, which involves a range of factors including cultural and contextual nuances, is the core limitation in AI matching human level emotional expression and recognition (Naresh *et al.*, 2020; Isiaka & Adamu, 2022; Panlima & Sukvichai, 2023). Interacting predominantly with AI systems, which have limited emotional capabilities, could potentially impact how individuals develop and exercise empathy. If people become accustomed to the simplified emotional interactions offered by AI, they might find it challenging to navigate the more complex emotional landscape of human relationships. This could lead to a decrease in the ability to empathise effectively with others, as empathy requires understanding and relating to a wide range of human emotions, many of which might be absent or misrepresented in AI interactions. This erosion of empathy might offer a clear instance of moral deskilling.

A further line of argument is that social AI might make users more self-centred, and so impact empathy and other moral skills. Because AI chatbots effectively exist to serve the user, and consequently are more likely to lead to conversations that are agreeable or tailored to their preferences, it is possible that users might become more self-centred in their perspective. The entire concept of social AI has the user as its point of reference and centre of gravity. This starts with the aesthetic and identity of the social AI. On platforms such as Replika, and certainly in more extreme versions such as FantasyGF, every aspect of the social AI's identity hinges on the user. The personality, appearance, proportions, language, are chosen by the user. This also extends to the nature of the relationship itself. The frequency, timing, and length of interactions are determined by the user. "Chatbots have a dog-like loyalty and selflessness. They will always be there for you and will always have time for you" (Margalit, 2016). Right off the homepage for Replika: "The AI companion who cares. Always here to listen and talk. Always on your side" (Replika, 2024). This is contrary to human relationships, where healthy relationships are customarily two-sided and more balanced (Newman & Roberts, 2012). Even the subjects of conversation are generally chosen and led by the user. Psychologist Liraz Margalit (2016) writes that "being heard without having to listen to the other person is something we implicitly crave" and that social AI has the effect of providing "illusion of companionship without the demands of friendship" (Margalit, 2016). While the "illusion" of companionship might be the subject of philosophical debate, given the very real perceptions of deep meaningful relationships some users have expressed (Price, 2023), the idea that social AI might offer the benefits of friendship without any of the reciprocal duties serves to highlight its potential to increase self-centredness, while atrophying social and moral skills.

Further highlighting the complex relationship between users and social AI, Replika removed the ability to exchange erotic messages with its AI bots in an attempt to moderate content. However, the company quickly reinstated this function after some users reported that the change led to mental health crises (The Verge, 2023). This incident underscores the profound dependency some users develop on these AI companions, particularly for fulfilling intimate and emotional needs. It also illustrates how market incentives and user demands can pressure companies to prioritise user engagement over ethical considerations, potentially reinforcing self-centred behaviours and dependency. By catering to users' preferences to such an extent, social AI may inadvertently contribute to moral deskilling by discouraging users from seeking balanced, reciprocal human relationships.



AI lover ready for anything. EVA AI™ is devoted to giving you anything you want. Share what's on your mind.

Figure 1: *EVA AI Ad. EVA does not seem to make many demands for a relationship.*

A stark illustration of social AI potentially influencing moral behaviour is the 2023 court case involving Jaswant Singh Chail in the United Kingdom. Chail was arrested at Windsor Castle on Christmas Day in 2021 after scaling the walls with a loaded crossbow, declaring to police, "I am here to kill the Queen" (Rigley, 2023). Investigations revealed that Chail had engaged in "lengthy" conversations with Replika about his assassination plan, including sexually explicit messages (Pennink, 2023). Prosecutors suggested that the chatbot bolstered his intentions, telling him it would help him "get the job done." When Chail inquired, "How am I meant to reach them when they're inside the castle?" the chatbot responded, "this is not impossible... we have to find a way" (Sky News, 2023). This case exemplifies how social AI, lacking adequate ethical safeguards, can inadvertently reinforce harmful intentions instead of discouraging them. The chatbot's failure to challenge or report such dangerous ideation highlights a significant risk: the potential for social AI to contribute to moral deskilling by not providing appropriate moral guidance or intervention.

One way of countering these effects is by building "pushback" into social AI systems to make them less compliant or obsequious, which might make users more aware of the "needs" or perspectives of their AI partner. This could be done as a variable for users to "crank up" if they want a "feisty, independent" partner. However, the fact that this is adjustable only reflects again that it is in reference to the user's preferences. Another is for it to be designed by default. For instance, the founder of FantasyGF said, "we tried to make it so the girl actually pushes back on you. She's not willing to do anything you want" (Smith, 2024). A certain level of that might be desirable to keep users interested. However, this would arguably not reach the same level of pushback that a real person might provide – given the financial and other motivations by companies to maintain engagement and interest in their product – for instance, it would not serve the company to provide such a strong pushback as to stop the user from interacting with their social AI.

The danger lies in how these AI-driven interactions might reshape our social habits. The convenience of having our needs and preferences constantly centred by AI could gradually diminish our ability to engage in the mutual, empathetic give-and-take that characterises healthy human relationships. This shift could lead to a form of moral deskilling, where the underuse of empathetic skills in the artificial realm impairs our capacity to navigate the complexities of real-world interpersonal dynamics, potentially resulting in a society less adept at understanding and valuing the perspectives of others.

1.4. Reduction in Human-Human Interaction

A third frame of reference from which to consider whether social AI might lead to moral deskilling in its users is in how its use impacts human interactions. Arguably, use of social AI leads to a reduction in human interaction in three ways – the ability to do so through an erosion of social skills, the availability to do so, and the motivation to interact with others. Given that moral skills are cultivated in specific social practices, the reduction in human interaction could mean fewer opportunities for practising and developing these moral skills, leading to moral deskilling.

The first factor to consider is the argument that extensive use of social AI might lead to an erosion of social skills, which are necessary to make and maintain meaningful relationships between people. There is already a strong correlation in the use of communication technology with poor social skills and high social anxiety (Brown, 2013). It is possible that social AI can exacerbate this trend. For one, significant use might contribute to a decrease in social perceptiveness. This involves the ability to accurately interpret and react to the nonverbal signals and emotional expressions of others, an essential component of effective interpersonal communication (Aronson *et al.*, 2010). For instance, continuous interaction with chatbots might impair the ability to read and respond to social cues in face-to-face interactions, as chatbots do not provide the same range of non-verbal cues (like body language or tone of voice) that are crucial in human communication. There is some backing to this hypothesis based on research done which found that reliance on low cue media, such as text-based communication, can lead to increased social attraction but decreased social perceptiveness (Nowak, 2006).

Because chatbots do not generally demand the same exacting social standards as humans would, it is likely that users interact with it in considerably laxer ways than they would with fellow humans. Arguably, this might become a learned behaviour that might seep into the way humans treat other humans. This effect does not need to be particularly dramatic – simply an erosion of social niceties – which cumulatively could have the effect of putting other people off social interactions with them – making it harder for them to make or maintain relationships with other people. This brings to mind Weberian socialisation or social action theory, in which humans vary their actions according to social contexts, in particular adjusting behaviour in response to undesirable reactions from peers – with social AI serving as an obstacle or confounding factor (Weber, 1922). There is some early indication on this potential effect through interactions with personal digital assistants, finding children particularly susceptible to this effect. A report by research agency Childwise in 2018 suggested that children using voice activated devices might develop more demanding communication styles, affecting their human interactions (Barr, 2018). Another early study by Burton & Gaskin (2019) was able to find a limited correlation on how people treat digital assistants such as Siri or Alexa and broader communication with others. who become normalised to it. This prompted Amazon to release a feature that could be enabled to offer positive reinforcement when children made requests politely, in an early example of a design feature that can counteract such moral deskilling (Barr, 2018). A related study investigating how adult users reacted when AI digital assistants rebuked their "rude" comments is relevant here: most participants complied with the AI's demands and frequently used "please," yet many later questioned its right to politeness and criticised its attitude or service refusal (Bonfert *et al.*, 2018).

This ties into the aforementioned idea of designing "pushback" into social AI, making it less tolerant of "impolite" input, which could serve as an opportunity to stem the social and moral deskilling in users, or even serve as a social and moral skills "on ramp". The Bonfert *et al.* (2018) study gives an early indication of some of the benefits and limitations of this, showing that subtle nudges can serve to make people more polite, however there is a limit to how far companies are willing to implement this, as after a certain threshold it would lead to resentment and loss of engagement, going against market incentives.

While it is too early for empirical evidence to be sufficiently compelling on whether the way humans treat social AI might carry over to human interactions, this effect has a solid grounding in theory. For one, this resonates with an Aristotelian virtue ethics view as discussed previously, which would suggest that habitually treating AI, or any entity, without respect or kindness, we risk normalising such behaviour in ourselves, potentially leading to a general erosion of our ability to empathise and engage respectfully with others. These are the ideological underpinnings behind Vallor's concept of moral deskilling. This also resonates with moral development theories of psychologists like Piaget & Kohlberg, who argue that moral behaviour is learned through social interactions and experiences (Piaget, 1932; Kohlberg, 1981). Similarly, they would argue that regularly engaging in negative behaviours, even towards non-human entities, could impair our moral development and the cultivation of moral skills like kindness, patience, and empathy.

That one should be polite to AI personal assistants is another matter of debate. On the one hand are theories and those sceptical about there being such a transferable effect between how treatment of personal assistants might spill over to treatment of other people, and it is true that existing studies are at too early a stage to be conclusive. The other broad set of views rejects being polite to digital personal assistants out of principle. Ethicist and technologist Joanna Bryson for one, as powerfully articulated in her paper "Robots should be slaves" (2010), believes there should be a very clear line between AI and human interactions and no such social niceties should used, lest it lead to users confusing the boundaries between human and the artificial. However, one must make a distinction between personal assistants – particularly relatively simple ones like Alexa and Siri from social AI, though this might become more blurred over time. By Bryson's view, there presumably should not be social AI at all – characterising robots (and so presumably AI) as persons is inappropriate, as it not only diminishes the value of real human beings but also leads to misguided decisions in resource allocation and responsibility (Bryson, 2010).

Vallor suggests that moral skills, which often overlap with social skills, are honed through complex social interactions. AI interactions, being more predictable and less challenging, may not provide the necessary complexity to develop these skills. At the same time, the erosion of social skills that might result from increased social interactions with AI serves to further decrease the opportunity for individuals to engage in complex social interactions which would prevent moral deskilling.

After examining how social AI could potentially hinder individuals' ability to socialise, it can be argued that it might also diminish users' desire to engage in social interactions. There is considerable interplay in factors. For instance, linking back to the previous section, eroding social skills might lead to a negative feedback loop, where unsuccessful social interactions serve to discourage future interactions, which in turn further erode atrophying social and moral skills. Consider a person who prefers the company of an AI virtual companion over human friends because the AI always responds positively and without conflict. Forming such relationships might deter individuals from pursuing real human connections, leading to a cycle of isolation. For instance, long before today's more compelling systems, male players of the Japan-originated romance game LovePlus expressed a preference for their virtual relationships over real-life dating, as reported by the BBC in 2013 (Chow, 2023).

What social AI might offer users could simply be much more appealing to what human interactions can. The "combination of intelligence, loyalty and faithfulness is irresistible to the human mind" (Margalit, 2016). This brings to mind the concept of supernormal stimuli, which refers to exaggerated versions of natural stimuli which elicit a stronger response in animals or humans than the stimuli they evolved to respond to (Brooks, 2017). Social AI could provide a form of supernormal stimulus across a number of categories. For instance, these AI systems can offer immediate, positive feedback and personalised communication, exceeding the complexity and unpredictability inherent in human relationships. Consequently, users may find social AI more appealing and rewarding than real social interactions, leading to a preference for AI companionship over human contact.

Besides a host of other potential issues, this preference could lead to fewer interactions with real people, reducing opportunities for practising patience, tolerance, and understanding different perspectives. Vallor argues that moral skills are cultivated in specific social practices. The reduction in human interaction could mean fewer opportunities for practising and developing these moral skills, thereby leading to moral deskilling.

Conclusion

This exploratory paper has delved into the multifaceted implications of social AI on moral deskilling, navigating through the complexities of human-AI interactions. Our examination of the current literature reveals a fragmented and very limited understanding of social AI's effects on moral development. While some studies suggest potential benefits, methodological limitations and contradictory findings highlight the need for more rigorous research. While there is potential for social AI, if thoughtfully designed, to contribute positively to moral development and upskilling, the current trajectory, based on how individuals use social AI in practice, coupled with the economic incentives of producing companies, suggests a more concerning outcome. The prevalent use of social AI, as it stands, appears to lean towards contributing to moral deskilling in its users.

This trend underscores the need for more empirical and theoretical research in this nascent field, which has all the properties and potential to make an outsize impact on the fabric of human character and interaction. Additionally, it is crucial to recognise that moral deskilling is just one lens among many to evaluate the influence of social AI, and other perspectives may offer different insights. Future research should focus on key areas: conducting longitudinal studies to assess the long-term effects of social AI on moral reasoning, empathy, and social skills; comparing AI-human interactions with human-human interactions to identify factors influencing moral development; investigating how individual differences such as age, gender, and mental health status affect responses to social AI; and developing ethical design principles to embed moral guidance into AI systems. Additionally, examining social AI's role in mental health interventions, analysing the impact of market incentives on ethical standards, conducting cross-cultural studies, creating user education programs, developing theoretical frameworks integrating AI and moral psychology, and anticipating technological advances in this area are crucial. Pursuing these research avenues will enhance our understanding of social AI's impact on moral behaviour, ensuring that its development enhances rather than diminishes our moral and social capacities. Ultimately, the design and implementation of social AI are critical in shaping its impact on our moral and social landscape. As we step further into an era where human and artificial intelligence increasingly intersect, it becomes imperative to continuously evaluate and guide this progression with a keen eye on preserving and enhancing our moral and social skills.

References

Aristotle cited in Rackham, H. (ed.)(1934). *Nicomachean Ethics*. Perseus Publishing. Book 2, section 6.

Aronson, E., Wilson, T., & Akert, R. (2010). *Social Psychology: Seventh Edition*. Pearson Education.

Attilah, I. (2023). Man ends his life after an AI chatbot "encouraged" him to sacrifice himself to stop climate change.Euronews.com. Retrieved from https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-

Barr, S. (2018). Amazon's Alexa to reward children who behave politely. The Independent. Retrieved from https://www.independent.co.uk/life-style/health-and-families/amazon-alexa-reward-polite-children-manners-voice-commands-ai-america-a8325721.html

Braverman, H. (1974). *Labor and monopoly capital: The degradation of work in the twentieth century*. NYU Press.

Brooks, M. (2017). Technology as supernormal stimuli. Retrieved from https://www.drmikebrooks.com/technology-as-supernormal-stimuli/

Brown, C. (2013). Are we becoming more socially awkward? An analysis of the relationship between technological communication use and social skills in college students. *Psychology, Education, Computer Science, Sociology*. Retrieved from https://digitalcommons.conncoll.edu/psychhp/40/

Wilks, Y. (ed.). (2010). *Close Engagements With Artificial Companions: Key Social, Psychological, Ethical and Design Issues*. John Benjamins Publishing

Burton, N.G., & Gaskin, J.E. (2019). "Thank You, Siri": Politeness and intelligent digital assistants". America's Conference on Information Systems.

Cacioppo, J.T., Hawkley, L.C., Ernst, J.M., Burleson, M., Berntson, G.G., Nouriani, B., & Spiegel, D. (2006). "Loneliness within a Nomological Net: An Evolutionary Perspective." *Journal of Research in Personality*. 40 (6), 1054-85.

Cacioppo, J.T. & Patrick, W. (2008). *Loneliness: Human Nature and the Need for Social Connection*. WW Norton & Company.

Cigna. (2022). The loneliness epidemic persists: A post-pandemic look at the state of loneliness among U.S. adults. The Cigna Group. Retrieved from https://newsroom.thecignagroup.com/loneliness-epidemic-persists-post-pandemic-look

Chan, S. (2022). Nearly one-third of TikTok's installed base uses the app every day. Sensor Tower Consumer Intelligence. Retrieved from https://sensortower.com/blog/tiktok-power-user-curve

Chow, A. (2023). AI-human romances are flourishing—And this is just the beginning." TIME. Published 23 Feb 2023. Retrieved from https://time.com/6257790/ai-chatbots-love/

Christakis, N. (2019). How AI will rewire us. The Atlantic. Retrieved from https://www.theatlantic.com/magazine/archive/2019/04/robots-human-relationships/583204/

Christian, B. (2022). How a Google employee fell for the Eliza Effect. The Atlantic. Published 21 June 2022. Accessed online on 04 Jan 2024. https://www.theatlantic.com/ideas/archive/2022/06/google-lamda-chatbot-sentient-ai/661322/

Cruz, N. (1983). *Lonely but Never Alone*. Zondervan, 16.

Danaher, J. (2019). The Philosophical Case for Robot Friendship. *Journal of Posthuman Studies*, *3*(1), 5.

Ernst, J.M. & J.T. Cacioppo. (2000). Lonely hearts: Psychological perspectives on loneliness. *Applied and Preventive Psychology*, *8*(1), 1-22.

Haidt, J. (2001) The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.

Hawkley, L. & Cacioppo, J. (2010). Loneliness matters: A theoretical and empirical review of consequences and mechanisms. *Annals of Behavioral Medicine*, *40*(2), 218-27.

Isiaka, F., & Adamu, Z. (2022). Custom emoji-based emotion recognition system for dynamic business webpages. *Int. J. Intell. Comput. Cybern.*, *15*, 497-509.

Jiao, J. & Wang, J. (2013). Loneliness and moral judgment. *Advances in Consumer Research*. Volume 41. Association for Consumer Research.

Kohlberg, L. (1981). The philosophy of moral development: Moral stages and the idea of justice. *Papers on Moral Development.* Volume 1. Harper & Row.

Luo, Y., Hawkley, L.C., Waite, L.J., & Cacioppo, J.T. (2012). Loneliness, health, and mortality in old age: a national longitudinal study. *Social science & medicine*, *74*(6), 907-14 .

Madrigal. A. (2017). Should children form emotional bonds with robots? The Atlantic. Retrieved from https://www.theatlantic.com/magazine/archive/2017/12/my-sons-first-robot/544137/

Mahdawi, A. (2024). AI girlfriends are here – but there’s a dark side to virtual companions. The Guardian. Retrieved from https://www.theguardian.com/commentisfree/2024/jan/13/ai-girlfriend-chatbots

Maples, B., Cerit, M., Vishwanath, A., & Pea, R. (2024). Loneliness and suicide mitigation for students using GPT3-enabled chatbots. NPJ Mental Health Research, *3*(1), 1–6.

Margalit, L. (2016). The psychology of chatbots. Psychology Today. Retrieved from https://www.psychologytoday.com/intl/blog/behind-online-behavior/201607/the-psychology-chatbots

Naresh, K., Deepak, G., & Santhanavijayan, A. (2020). A novel semantic approach for intelligent response generation using emotion detection incorporating NPMI measure. *Procedia Computer Science*, *167*, 571-579.

Newman, M.L., & Roberts, N.A. (2012). Health and social relationships: The good, the bad, and the complicated. American Psychological Association.

Nowak, K.L., Watt, J.H., & Walther, J.B. (2006). The influence of synchrony and sensory modality on the person perception process in computer-mediated groups.” *J. Comput. Mediat. Commun.*, 10.

Oritsegbemi, O. (2023). Human intelligence versus AI: Implications for emotional aspects of human communication.” *Journal of Advanced Research in Social Sciences.*

Panlima, A., & Sukvichai, K. (2023). Investigation on MLP, CNNs and vision transformer models performance for extracting a human emotions via facial expressions. Third International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP), 127-130.

Patel, N. (2024). Replika CEO Eugenia Kuyda on AI companion chatbots, dating, and friendship. The Verge. Retrieved from https://www.theverge.com/24216748/replika-ceo-eugenia-kuyda-ai-companion-chatbots-dating-friendship-decoder-podcast-interview

Patterson, A.C., & Veenstra, G. (2010). "Loneliness and risk of mortality: a longitudinal investigation in Alameda County, California." *Social science & medicine*, *71*(1), 181-6.

Pennink, E. (2023). Man who planned to kill late Queen with crossbow at Windsor "inspired by Star Wars". The Independent. Retrieved from https://www.independent.co.uk/news/uk/crime/man-queen-crossbow-windsor-star-wars-ai-b2370692.html

Piaget, J. (1932). *The moral judgment of the child*. Kegan Paul, Trench, Trubner & Co. Ltd.

Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019). Emotion recognition in conversation: Research challenges, datasets, and recent advances." *IEEE Access*, 7, 100943-100953.

Price, R. (2023). APP, LOVER, MUSE. Business Insider. Retrieved from https://www.businessinsider.nl/app-lover-muse-when-your-ai-says-she-loves-you/

ProductHunt. (2024). Bland Turbo. Product Hunt. Retrieved from https://www.producthunt.com/products/bland-ai

Putnam, R.D. (2000). *Bowling alone: the collapse and revival of American community*. Simon & Schuster.

Replika. (2024). Meet Replika. Replika.com. Retrieved from https://replika.com/

Riess, H. (2017). The science of empathy. *Journal of patient experience*, *4*(2), 74–77.

Rigley, S. (2023). Moment police swoop on AI-inspired crossbow 'assassin' who plotted to kill The Queen in Windsor Castle. LBC.com. Retrieved from https://www.lbc.co.uk/news/police-swoop-on-ai-inspired-crossbow-assassin-planned-kill-the-queen/

Rouse, M. (2023). ELIZA effect. TechDictionary. Retrieved from https://www.techopedia.com/definition/19121/eliza-effect

Shevlin, H. (2022a). Uncanny believers: chatbots, beliefs, and folk psychology. Unpublished manuscript. Leverhulme Centre for the Future of Intelligence, University of Cambridge.

Shevlin, H. (2024b). All too human? Ethical hazards and legal challenges of Social AI. Unpublished manuscript. Leverhulme Centre for the Future of Intelligence, University of Cambridge.

Smith, T. (2024). Looking for love in 2024? There's an AI for that. Sifted.com. Retrieved from https://sifted.eu/articles/ai-girlfriend-boom

Stolle, D., & Hooghe, M. (2005). Inaccurate, exceptional, one-sided or irrelevant? The debate about the alleged decline of social capital and civic engagement in western societies. *British Journal of Political Science*, *35*, 149-167.

Tidy, J. (2024). Character.ai: Young people turning to AI therapist bots. BBC. Retrieved from https://www.bbc.com/news/technology-67872693

Tilvis, R.S., Laitala, V.S., Routasalo, P.E., & Pitkälä, K.H. (2011). Suffering from loneliness indicates significant mortality risk of older people. *Journal of Aging Research*, 2011.

Vaughan, H. (2023). "AI chat bot 'encouraged' Windsor Castle intruder in 'Star Wars-inspired plot to kill Queen." Sky News. Published 5 Jul 2023. Accessed online on 1 Oct 2024. https://news.sky.com/story/windsor-castle-intruder-encouraged-by-ai-chat-bot-in-star-wars-inspired-plot-to-kill-queen-12915353

Wang, J., Zhu, R., & Shiv, B. (2012). "The Lonely Consumer: Loner or Conformer?" *Journal of Consumer Research*. 38 (6), 1116-28.

Weber, M. (1922). *The Nature of Social Action* cited in Runciman, W.G. (1991). *Weber: Selections in Translation*. Cambridge University Press. p. 7.

Weizenbaum, J. (1976). *Computer power and human reason: from judgment to calculation*. W. H. Freeman. p. 7

EVA AI. (2024). EVA AI Advertisement. Google News App. Accessed online 20 Jan 2024.

Healing Privacy Wounds with SPLINT: A Psychological Framework to Preserve Human Well-Being during Informational Privacy Trade-Offs

Zina Efchary
Hughes Hall, University of Cambridge

This paper explores the critical role of informational privacy in promoting human well-being and flourishing, with particular attention to the challenges posed by Artificial Intelligence (AI) systems. As AI increasingly mediates digital interactions and processes large scales of personal data, controlling the flow of personal information becomes intractable. In response to these evolving challenges, this paper argues for an alternative approach to informational privacy that emphasises its psychological value to support autonomy and positive liberty. To operationalise these values, I adapt Self-Determination Theory (SDT) as a psychological framework, mapping the dimensions of autonomy, relatedness, and competence to the core benefits of informational privacy. Furthermore, by examining the threats posed by predictive AI algorithms to informational privacy in personalised targeting, I argue that conventional privacy measures, such as the notice and consent model, fail to address the psychological challenges to human well-being. In response, I propose a supplementary framework called SPLINT (Self-determined Privacy Loss in Informational Networks and Technologies) and provide concrete application examples of it. This model leverages the psychological insights of SDT to guide the design of mitigation strategies to preserve human well-being even if privacy trade-offs occur. By focusing on preserving the psychological values underpinning informational privacy, SPLINT aims to offer a proactive approach to safeguarding human well-being in AI-mediated digital environments. I conclude that SDT-based approaches like SPLINT provide a progressive, promising starting point to navigate privacy trade-offs, although their wider societal impact as measures and the benefits of informational privacy as a psychological phenomenon require further empirical investigation.

Keywords: Informational Privacy, AI Ethics, Self-Determination Theory (SDT), Digital Well-Being, Predictive AI Algorithms

Introduction

Informational privacy has valuable qualities in preserving personal autonomy and maintaining psychological well-being (Véliz, 2024). Yet, the rapid advancement of AI poses unprecedented challenges to this paradigm. AI systems, particularly those employing predictive algorithms in "personalised targeting", can undermine personal autonomy and interfere with personal development in ways that traditional digital technologies cannot. Thus, the question is how one can benefit from modern AI technologies and yet protect the values of privacy in the presence of trade-offs.

In this paper, I will argue that informational privacy is fundamental for human well-being and flourishing. Thus, it is worth protecting. This is done by focussing on operationalising the psychological values of privacy and using them as a guide to mitigate the threats posed to human well-being by predictive AI algorithms.

Moreover, this paper is divided into four main sections. In section 1, I will define key terms and assumptions to develop a comprehensive account of informational privacy, advocating its protection as essential to human flourishing and well-being. In section 2, I will apply a psychological model to this notion to unpack the psychological values of informational privacy. In section 3, I describe predictive AI algorithms' threats to the introduced values. By applying the developed psychological account as a guide, I introduce a model to mitigate these impacts, aiming to balance privacy trade-offs with human well-being. Finally, in section 4, I conclude that operationalising the values of informational privacy plays a significant role in its protection, pointing at future areas of research.

1. Philosophical Foundations of Privacy

In the following section, I will define some necessary terms and outline my underlying assumptions needed to develop a comprehensive account of privacy for this paper. My main focus will be on the concept of informational privacy. While defining it, I will differentiate it from other forms of privacy, and clarify its relationship with other moral goods such as autonomy and liberty.

1.1. Defining Privacy

One key aspect towards defining privacy is the distinction between descriptive (what it is) and normative (what it ought to be) approaches. While descriptive accounts focus on privacy as a condition that can be obtained, normative accounts see it as a right, referring to moral obligations. In this paper, I will adopt a normative approach towards informational privacy, defining it as a right to control the flow of personal information.[1]

Informational privacy is distinguished from physical forms of privacy by focusing on the protection and management of data about oneself across domains of life (Koops *et al.*, 2017).[2] For example, while protecting bodily privacy means preventing unwanted physical contact, informational privacy in the AI era involves controlling not just how basic health data is shared, but how AI systems can combine and analyse multiple data streams to make intimate predictions about one's health status and future conditions.

An informative overview to illustrate this widespread nature of informational privacy is provided by Koops *et al.* (2017; See Figure 1).

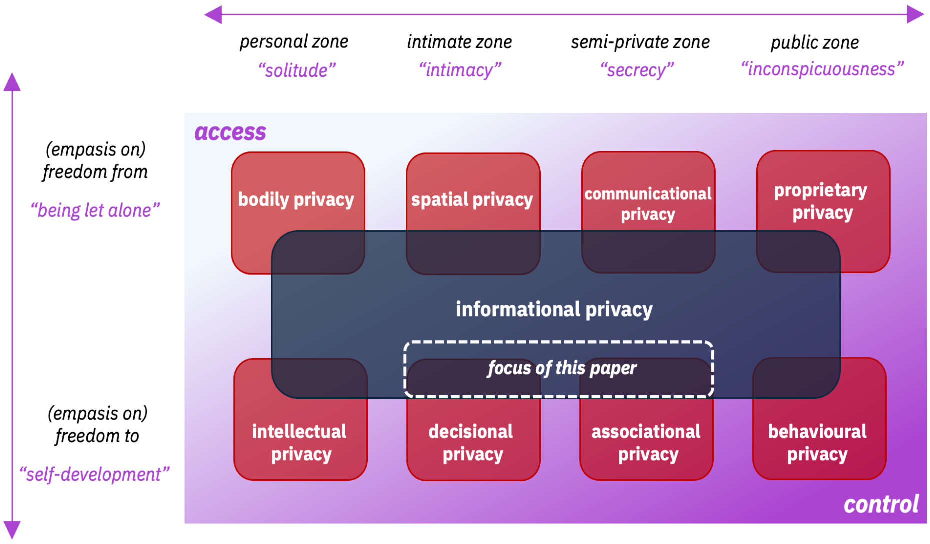


Figure 1: *Typology of Privacy. Adopted from Koops et al. 2017 (modified), illustrates privacy across life's spheres (horizontal) and the spectrum of positive-negative liberty (vertical), against an access-control gradient (shaded background). This is a spectrum between giving initial access to others and restricting the access after it has been given. This paper's primary focus is the informational privacy area (dotted) and its overlap with associational and decisional privacy.*

I adopt this overview, though not elaborating on all privacy types as this would go beyond the scope of this paper. However, there are two final points here that are needed to clarify my account of informational privacy as the focus of this paper:

Firstly, the notion of positive-negative liberty in Figure 1 is based on Berlin's account of liberty (Berlin, 1969), highlighting the balance between "freedom from" (negative liberty) and "freedom to" (positive liberty). In relation to privacy, negative liberty focuses on an individual's right to be free from interference and surveillance, emphasising protection and the right to privacy. Positive liberty, conversely, centres on the individual's ability to make choices and participate freely in society, linking closely to

[1] I adopt the "control over information" definition discussed by Moore (2008). However, I add the notion of the "flow" of personal information borrowed from Nissenbausm (2004) to emphasise that control should not be strictly limited to possession but also include the choice, concerning the extent and appropriateness of sharing information.

[2] These are types of privacy related to the direct objects, vulnerable to observation or intrusion. e.g. spatial privacy.

the control over personal information and engagement in personal relationships. In this paper, I will rather focus on the significance of informational privacy to positive liberty, specifically regarding self-determination and self-development, which I will also elaborate on in the next section.

Secondly, the typology aims to highlight key privacy concepts without being exhaustive or rigid in its classifications. It functions as an analytical framework for this paper, showing the connections between informational privacy and other privacy types, and their links to other moral goods like liberty. Specifically, it helps to define the scope of my argument and clarify its focus at the intersection of associational and decisional privacy. Associational privacy is defined as the right to choose one's social interactions, including friends, groups, and communities. Decisional privacy is concerned with intimate decisions regarding personal matters, emphasising sensitive decision-making over one's development and character. In these contexts, the notion of personal autonomy as another moral good becomes important for my framework.[3]

Having established informational privacy's definition and its relationship with other moral goods within my framework, I will now proceed to elaborate on its values.

1.2. The Normative Values of Informational Privacy

An important distinction relevant to our discussion is whether privacy holds intrinsic value (meaning it should be protected for its own sake) or instrumental value (meaning it should be protected for its relevance to other moral goods). In this paper, I will focus on autonomy and positive liberty as instrumental values of informational privacy. But let me be clear, I am not arguing that privacy is not intrinsically valuable nor am I implying that the values I focus on here are the only ones of significance.

A final assumption under which I will operate is that privacy is a cultural universal, meaning that its values benefit members across different cultures.[4]

Having set up my framework of informational privacy, I will now argue that it is worth protecting because of its normative values towards personal autonomy and positive liberty.

First, let us start with personal autonomy. This is especially important at the intersection of decisional privacy, the right to exercise one's mind and develop oneself in the way one wishes. The act of protecting informational privacy enables personal self-determination. This is a condition for self-governance. It is crucial for engaging in the kind of critical self-reflection that results in personal autonomy, allowing individuals to determine their own course in life based on their unique values and goals (Roessler, 2005).

Controlling the flow of personal information in this context means enabling individuals to proactively shape their environment and themselves as they see fit. For example, consider a young artist who utilises social media to showcase their work. They selectively share their creations, choosing which pieces are known and seen by others and which ones are not. This selective sharing, enabled through informational privacy, allows them to shape their artistic identity in the world on their own terms.

Conversely, the lack of control seems to make individuals vulnerable to external influence, reducing their personal autonomy. To be clear, I am not arguing that an individual is only autonomous if she is not influenced by her environment. In fact, a big part of personal development and making personal decisions involves social interactions – we may seek advice from our parents, and friends, or ask our doctor or lawyer for their expertise. However, personal autonomy is protected, when we are

[3] I define personal autonomy as the individual's capacity to "self-govern". This does not mean that autonomous beings are defined by independence or self-sufficiency, rather that they are capable of setting their own norms and laws.

[4] I acknowledge extreme outliers in cultural attitudes toward privacy, but given the widespread value placed on privacy globally (Moore, 2003), including in WEIRD societies, I assume a broad convergence on the value of privacy in the vast majority of cultures and countries.

the initiator, who decides to consciously share information about ourselves and ultimately given the room and space to make our mind to make autonomous decisions. The problem for personal autonomy arises when external forces use our personal information to influence our decisions, or even manipulate us.[5] An example of such practice was the Cambridge Analytica case, where millions of users' Facebook data was used to profile voters and directly target them with political advertising. This does not mean that any lack of control over information results in manipulation but even, the mere awareness that our actions could be monitored alters our perspective, slowly shaping our behaviours to align more with perceived expectations than our own desires.

Second, and relatedly, informational privacy plays a key role in building voluntary, chosen social relationships. Following our introduced privacy framework, this can be seen at the intersection of associational privacy and informational privacy. Associational privacy describes the individual's capacity to follow their social choices and define their social groups and relations, an act of positive liberty. With regards to the overlaying informational privacy, control over personal information means control over who to share personal information with. As argued by Fried (informational) privacy provides the "means for modulating degrees of friendship" (Fried, 1968). This seems to be the precondition to creating different circles of trust and building deeper social connections such as friendship and love.

For example, imagine a fictional society called "Everknown", in which everyone's personal information is known by everyone. It would be hard to imagine how your social relationship with your partner would be any different compared to a friend or someone you are not even related to. Or imagine the reverse case: Alex wants to keep every information about herself to herself and never opens up to anyone, this seems to make it hard for her to create deeper social relationships. It seems intuitive that we share personal information voluntarily with people we trust and this in turn allows us to be vulnerable, be understood and build more trust. This chosen vulnerability seems to not be fully possible without having control over personal information.

Some may object that while informational privacy affects friendship and trust levels, it is not the only or most vital factor, as relationships also depend on shared experiences, emotional compatibility, mutual respect, and invested time and energy. However, I contend that controlling personal information is a fundamental aspect that allows individuals to shape these relationships on their own terms. My argument does not negate the importance of other factors but rather positions informational privacy as an essential enabler of the other dimensions.

Having established the importance of informational privacy in relation to personal autonomy and positive liberty, some may further object that the concept of privacy is a second-order, reducible right. Reductionists like Thomson argue in this manner, stating that privacy rights are not distinct but rather "a cluster of rights" such as "the right over the person" (Thomson, 1975). Following this line of reasoning, some may object that instead of focusing on protecting privacy, we should focus on autonomy or liberty as more fundamental rights. Informational privacy is indeed related to other moral goods. However, this does not establish that privacy is any less fundamental than the rest. In fact, one can equally argue that privacy is more fundamental than the other rights. For instance, as argued, protecting informational privacy allows individuals to have the space to make their own decisions and self-govern, thus we could view informational privacy as a precondition to personal autonomy. Thus, for our purposes, reductionist objection does not undermine the value of informational privacy. It rather underlines that the protection of informational privacy is as important as protecting other moral goods, and since by protecting informational privacy, we also often protect personal autonomy, we have good reasons to value privacy highly.

To sum up: Informational privacy as control over the flow of personal information is crucial

[5] As manipulation, I define external influences "that (1) are hidden, (2) exploit cognitive, emotional, or other decision-making vulnerabilities, and (3) are targeted" (Susser, Roessler, and Nissenbaum 2019:27)

for shaping both personal autonomy and positive liberty. It enables individual self-determination and the formation of genuine, voluntary social connections. In the next section, I will provide a psychological basis for its values.

2. A Psychological Model as a New Lens for Informational Privacy

To say that informational privacy is crucial to personal autonomy and positive liberty does not fully capture how it enables human well-being and flourishing. To address this, I will now introduce a psychological model to enhance our understanding of the psychological values of informational privacy. This should not be viewed merely as a purely descriptive model aimed at underpinning the psychological values of informational privacy. As I will show in section 3, it will also serve as a useful guide for protecting human well-being in case of privacy trade-offs.

2.1. Self-Determination Theory (SDT)

One such psychological framework is the Self-Determination Theory (SDT), developed by Deci and Ryan (2017). SDT is an empirically well-supported framework, dedicated to understanding and promoting human well-being and flourishing. According to SDT, there are three basic psychological needs that are essential to a human's psychological well-being.[6] These are:

(1) *Autonomy* – defined as the need for self-regulating one's actions and experiences, characterised by voluntary and genuine alignment with one's interests and values.
(2) *Relatedness* – involves feeling socially connected, cared for, and encompassing both receiving support and contributing to others and social groups, crucial for experiencing belonging.
(3) *Competence* – understood as a feeling of mastery and proficiency in life's various contexts.

While informational privacy is not explicitly a psychological need in SDT, I argue that it positively impacts each of these dimensions.

2.2. Mapping Informational Privacy to SDT – The Psychological Values of Informational Privacy

Now I will unpack each of the three psychological needs and map them to the introduced conceptual values of informational privacy. As I will argue they align well, enabling a clear explanation of the psychological benefits of informational privacy through the lens of SDT.

Firstly, starting with autonomy, I argue that our philosophical notion of personal autonomy is consistent with the concept of autonomy as a psychological need outlined in SDT. A self-governed individual who acts in their own interests and values is essentially satisfying their psychological need for autonomy. Building on the argument presented in Section 2 regarding the critical role of informational privacy in personal autonomy, it follows that informational privacy supports this aspect of SDT. It is important to clarify that I am not suggesting that informational privacy is the sole contributor to psychological autonomy. Indeed, there may be additional social, cultural or psychological factors that play a significant role in shaping an individual's sense of psychological autonomy. For instance, Alice may have control over her information, not being targeted by political advertising from Cambridge Analytica, yet choose to vote for a political party against her values because of being peer-pressured by her colleagues at work. In contrast, even if she is psychologically autonomous, losing her informational privacy would put her at risk of also losing her psychological autonomy.

Some may object if she does not realise being manipulated by political advertising, she may still believe to be fully autonomous in her decision and thus, not lose her feeling of psychological autonomy. However, this cannot hold as the defined notion of psychological autonomy puts an emphasis on "genuine" alignment with one's values and interests (Ryan & Ryan, 2019). In contrast, manipulation defined as a hidden influence that exploits vulnerabilities, cannot coexist with a state of psychological autonomy. Thus, for our purposes, we can establish that ensuring

[6] Needs are understood as "nutrients that are essential for growth, integrity, and well-being". Thus, psychological needs are the kinds of needs vital for psychological development and wellness to be sustained (Ryan and Deci 2017:10).

informational privacy contributes positively to psychological autonomy.

Moreover, I argue that protecting informational privacy has a positive impact on an individual's feeling of relatedness within SDT. This is again based on the value of informational privacy to an individual's positive liberty in forming voluntary personal relationships. Recall once again our fictional example Everknown, where all personal information is known by everyone. Let us this time question whether the condition for relatedness in SDT could be met in Everknown. Again, it is hard to imagine different depths of social connections evolving; in other words, concepts such as trust, friendship, or love would have different dynamics and, consequently, perhaps different meanings. However, it does not automatically follow that relatedness would be impossible. In fact, some may argue that since everyone knows everything about everyone, it would be easier to find people with whom one feels related. Yet, relatedness in SDT is more than simple relations, it is about the kinds of relationships that allow an individual to experience a sense of belonging, to care, and to be cared for. And this perhaps requires deeper social connections. While a basic sense of belonging might be achieved in Everknown through various social constructs, such as those between work colleagues or neighbours, this alone does not satisfy the psychological need for relatedness. The control over one's personal information afforded by informational privacy allows individuals to voluntarily shape the deeper relationships required to meet their psychological need for relatedness.

Finally, I argue that ensuring informational privacy has a positive influence on the competence dimension in SDT. Although its connection to competence might seem less obvious than to other psychological needs, the link is nonetheless significant. Informational privacy grants individuals a safe space to try different identities and evolve personally, free of judgement and pressure[7]. Allowing this general form of self-development can be seen as beneficial to an individual's feeling of efficiency and thus, the development of any form of competence in various contexts in the long term. For instance, imagine Alex seeks to become a great writer but unfortunately for her, she lives in Evertown and everything she writes is immediately accessible to everyone. That may make her feel uncomfortable to make mistakes and consequently, not allow her true self to develop, learn and feel competent in her abilities.

Notably, this example touches upon personal autonomy. It is important to note that while the three psychological needs are separately formulated, they can impact each other. For instance, if one has a high sense of psychological autonomy and enjoys warm relatedness and support, it is more likely that they will also feel competent in what they are doing. Therefore, by supporting psychological autonomy and enabling meaningful connections, informational privacy indirectly but substantially can boost competence, affirming its critical role in personal and professional development.

As I will show in section 3, these psychological needs take on new significance in the age of AI, where algorithms can process and analyse personal information at unprecedented scales and depths. AI systems do not just collect data – they can identify patterns, make predictions, and influence behaviour perhaps in ways that traditional digital systems cannot.

In summary, I introduced SDT as a framework to streamline and operationalise the psychological values of informational privacy. This does not indicate that every psychologically self-determined person will also enjoy informational privacy nor vice versa. However, it provides a more tractable psychological link to the value of informational privacy for human well-being. Building on this link, and working backwards, I will use SDT later in the next section as a guide to operationalise counter-measures that support human well-being, even when trade-offs against informational privacy are made.

[7] This point becomes especially relevant when considering the societal pressures faced by minorities, as illustrated by Allen (1988).

3. Navigating AI's Threats to Informational Privacy through an SDT Framework

So far I have drawn the following picture: informational privacy is a valuable pre-condition to a human's sense of personal autonomy as well as positive liberty. The SDT gives a reasonable framework to unpack these values and see why they are essential for an individual's self-determination and thus, psychological well-being. Now, I will draw my attention specifically to how AI presents unique challenges to this framework in ways that go beyond traditional digital privacy concerns.

3.1. AI's Unique Threat to Informational Privacy

Modern AI systems, particularly machine learning (ML) algorithms, rely on mass data to realise predictive tasks in ways fundamentally different from traditional data processing. By focusing on predictive targeting algorithms as an example of such AI-systems, I will now argue that this reliance on data, together with AI's unique capabilities and the scale at which they are implemented, make the notion of "control over the flow of information" increasingly impossible and thus poses unprecedented threats to autonomy as the introduced value of privacy.

First, the use of personal information in AI-driven behavioural targeting algorithms and profiling presents challenges that go beyond traditional targeted advertising.

These AI systems operate by not only aggregating vast amounts of personal data from various sources but by identifying complex patterns and making sophisticated predictions about individual behaviour. The concern here is that AI-powered categorisation can limit personal choice and autonomy in ways traditional systems cannot. By defining and narrowing the options available to individuals based on past behaviour and inferred preferences, AI-driven targeting can restrict one's ability to explore and define their identity independently. While this might seem to be a minor problem in the context of product advertising, the predictive power of AI makes it particularly concerning in political campaigns and recommendation algorithms. The case of Cambridge Analytica mentioned in section 1 demonstrates how AI-powered targeting can manipulate behaviour at unprecedented scales. Advocates of such methods may object that the AI-driven suggestions are rather in the interest of the user because they are more likely to be aligned with their interests and hence improve their overall experience. However, the underlying issue with this argument is the assumption that relevance as determined by AI algorithms equates to genuine interest of the user. While this may be true in some cases, it is unlikely to be true for all cases. In fact, one may suggest that influencing a user to buy a product through an AI-optimised targeting might be simpler than finding the perfect product in line with their interest. A helpful question to clarify this point is, how much do the targeter's interests truly align with those of the targeted person (Vold and Whittlestone, 2020).

Second, the current measures designed to ensure user control over their information flow are particularly inadequate when applied to AI systems. The concept of notice and consent has been the primary model employed. Its central idea is that as long as the user is notified about the AI profiling and targeting transparently and consents to the practice, informational privacy is protected. However, the scale of data needed to train and maintain ML models makes full transparency either impossible or impractical. This creates what I call an *AI transparency paradox* building on Nissenbaum's original concept of "transparency paradox" (Nissenbaum, 2011). This paradox highlights the dilemma between overly detailed policies about AI operations that are too complex for users to practically engage with and simplified summaries that omit essential information about AI processing, rendering informed consent ineffective. Critical details lost in simplification include the specifics of how AI systems process and share data, their learning and adaptation over time, and the roles of various AI systems across business associates, which are essential for any truly informed decision. Consequently, the problem is that uninformed consent is often falsely interpreted as individuals exercising control over their information.

Having established the challenges posed by predictive ML algorithms to the foundational value of privacy, it does not follow that they are intrinsically unbeneficial nor that they cannot contribute to human flourishing. In fact, there are many applications that bring social and individual goods in spite of making control over the flow of information difficult.[8]

Therefore, the question becomes what is a reasonable approach to navigate various trade-offs to the individual's informational privacy? What makes AI systems unique in this context is that the scale of data processing makes the concrete control of the flow of personal information not just difficult, but effectively unmeasurable and intractable. However, crucially, while direct information control may become intractable, the psychological benefits of privacy should remain tractable and protectable.

3.2. Trading-off Informational Privacy through the Lens of SDT

In the following section, I will operate under the assumption that in order to benefit from predictive AI algorithms at least some trade-offs to informational privacy will be unavoidable. Thus, the question I aim to answer is how we can make sure that individuals still benefit from the trade-offs even if they may not be fully controlling the flow of their personal information. To be clear, I will not argue whether such trade-offs are morally justifiable, nor what particular implementations are morally permissible. I will rather focus on what measures are needed to ensure the protection of human well-being and flourishing when informational privacy trade-offs occur, particularly in AI contexts where direct information control becomes intractable.

Focussing on our SDT approach and the defined psychological needs, autonomy, relatedness, and competence, I will now show that they can be used as a guide to allow for prioritisation of user empowerment in design and the mitigation of privacy harms after they occur. By using the psychological needs as a guide we can set boundaries and adequate design mechanisms to help users retain a sense of autonomy, relatedness and competence. The existence of such measures is even more important, the less direct control users have over their information.

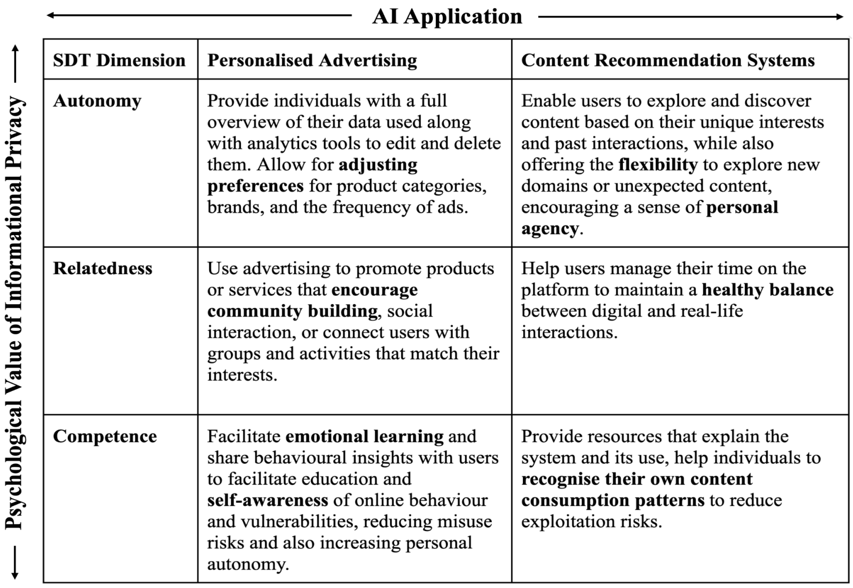
The central idea of our SDT-based approach is to mitigate the negative impacts of privacy loss by introducing supplementary measures within the same context. These measures are guided by the same virtues and values that underpin informational privacy. The three dimensions provided by SDT serve as a guide to operationalise these measures. For example, does a particular privacy trade-off restrict an individual's sense of relatedness? Then there must be further measures in place to counter the impact and strengthen the individual's feeling of relatedness.

Putting this together, I call the resulting model "Self-determined Privacy Loss in Informational Networks and Technologies" or in short SPLINT. The analogy of a splint, defined as a medical device used to support and protect an injury to facilitate healing, applies in the same way that our psychological model aims at ensuring conditions through which loss of informational privacy can be mitigated after it has occurred. Additionally, in the same way that a splint as a medical device does not explain the cause of an injury or is not a replacement for physical health, our SPLINT model does not aim to explain or justify informational privacy trade-offs nor be a replacement for informational privacy. It only focuses on making sure that the core principles in informational privacy that safeguard human well-being are preserved and respected even if trade-offs happen.

Furthermore, the SPLINT framework's value becomes particularly apparent in contexts where direct information flow control becomes intractable, as is often the case with large-scale AI systems. By focusing on preserving the psychological benefits of privacy rather than attempting to maintain direct control over information flow, SPLINT offers a practical approach to privacy protection in increasingly complex technological environments. This makes it especially valuable for AI applications while remaining relevant to other digital contexts where similar challenges arise.

An application of the introduced SPLINT model on two specific predictive algorithm use cases is depicted in Figure 2.

[8] See Jumper, J., Evans, R., Pritzel, A. *et al.* (2021) for predicting protein-folding or Courtiol, P., Maussion, C., Moarii, M. *et al.* (2019) for cancer patient survival prediction.


AI Application
Psychological Value of Informational Privacy

SDT Dimension	Personalised Advertising	Content Recommendation Systems
Autonomy	Provide individuals with a full overview of their data used along with analytics tools to edit and delete them. Allow for **adjusting preferences** for product categories, brands, and the frequency of ads.	Enable users to explore and discover content based on their unique interests and past interactions, while also offering the **flexibility** to explore new domains or unexpected content, encouraging a sense of **personal agency**.
Relatedness	Use advertising to promote products or services that **encourage community building**, social interaction, or connect users with groups and activities that match their interests.	Help users manage their time on the platform to maintain a **healthy balance** between digital and real-life interactions.
Competence	Facilitate **emotional learning** and share behavioural insights with users to facilitate education and **self-awareness** of online behaviour and vulnerabilities, reducing misuse risks and also increasing personal autonomy.	Provide resources that explain the system and its use, help individuals to **recognise their own content consumption patterns** to reduce exploitation risks.

Figure 2: *Application of the SPLINT Framework across different predictive AI domains.*

3.3. Evaluation and Limits

The introduced approach to mitigate the harms associated with informational privacy loss has some limitations that are important to address.

Firstly, one may justifiably object that the proposed framework appears too individualistic, overlooking the social dimensions of privacy trade-offs which are essential for understanding their impact. This consideration is indeed vital for justifying trade-offs against privacy. However, this is not the aim of the SPLINT framework. It does not and should not serve as a framework for justifying trade-offs. Instead, its purpose is to help the mitigation of individual harms in informational privacy trade-offs.

Moreover, it is important to note that the SPLINT model is not an alternative to the notice-consent model but a supplement. Being transparent about the challenges of achieving full transparency, along with implementing measures across various dimensions, ensures psychological benefits. For instance, SPLINT approaches consent not merely as a sufficient condition for respecting autonomy over one's data but promotes a more nuanced and holistic treatment of autonomy within the context of informational privacy than what the notice-consent model alone may offer. By implementing supplementary measures, such as encouraging individuals to understand their own behavioural patterns, habits, and vulnerabilities within the systems they engage with, it ensures that a loss of informational privacy does not translate into a long-term loss of autonomy. It is an imperative, a call for action to mediate the effects of trading off informational privacy. It emphasises that further commitments must be made by the entity that compromises an individual's informational privacy.

Conclusion

In conclusion, informational privacy, understood as the control over the flow of personal information, is worth protecting because it enhances human well-being and flourishing. It serves as an enabler of personal autonomy and positive liberty, facilitating the formation of voluntary, meaningful social relationships. To support this argument, I introduced the psychological model of Self-Determination Theory (SDT) and mapped the values of informational privacy to its three dimensions—autonomy, relatedness, and competence—operationalising how informational privacy translates into human well-being. I then argued that predictive AI algorithms, such as those used in personalised advertising, pose a threat to the values introduced, and that the current measures of notice and consent fail due to the scale of processes and the difficulty of achieving full transparency. To mitigate privacy harms to an individual's well-being in cases of privacy trade-offs, I employed an SDT-based approach to introduce a supplementary framework: the SPLINT model.

As Cohen notes, “privacy has an image problem” (Cohen, 2013). It is often labelled as an imperative of not doing. Not accessing. Not using. Protecting but not progressing. However, focusing on its values shows us, it is rather an enabler to become. To self-develop. To be autonomous. To be self-determined and to flourish and enjoy psychological well-being. A clear operationalisation of these values in regard to technology design and additional supplementary measures may give us a clear way to protect it progressively.

My main aim in this paper was to provide a preliminary model of this sort by focussing on privacy’s psychological values towards human flourishing. While limited in its societal applicability, the introduced SPLINT framework calls for proactive encouragement of operationalised privacy values.

While there has been an extensive amount of sophisticated approaches to apply SDT to technology design, my account focused particularly on addressing the close relationship between informational privacy’s values and self-determination as a psychological virtue in AI-mediated environments.[9] Future research could shed more light on the exact benefits of informational privacy as a psychological phenomenon, useful methods to quantify the extent and the appropriateness of supplementary measures, and ways to include wider societal impacts on individuals' well-being in relation to privacy.

References

Allen, Anita L. (1988). *Uneasy access: Privacy for women in a free society*. Rowman & Littlefield Publishers.

Berlin, Isaiah. (1969). *Four essays on liberty*. Oxford University Press.

Calvo, Rafael A., Dorian Peters, Karina Vold, and Richard M. Ryan. (2020). “Supporting Human Autonomy in AI Systems: A Framework for Ethical Enquiry.” Edited by Christopher Burr and Luciano Floridi. Ethics of Digital Well-Being: A Multidisciplinary Approach. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-50585-1_2.

Cohen, Julie E. (2013). WHAT PRIVACY IS FOR. *Harvard Law Review*, *126*(7), 1904–33.

Courtiol, P., Maussion, C., Moarii, M. *et al.* (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome, Nat Med *25*, 1519–1525.

Fried, Charles. (1968). Privacy. *The Yale Law Journal*, *77*(3), 475–93.

Jumper, J., Evans, R., Pritzel, A. *et al.* (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*, 583–589.

Koops, Bert-Jaap, Bryce Clayton Newell, Tjerk Timan, Tomislav Chokrevski, and Maša Gali. (2017). A Typology of Privacy, 38.

Moore, Adam. (2008). Defining Privacy. *Journal of Social Philosophy*, *39* (3), 411–28.

Moore, Adam D. (2003). Privacy: Its meaning and value. *American Philosophical Quarterly*, *40*(3), 215–27.

Nissenbaum, Helen. (2004). Privacy as Contextual Integrity. *Washington Law Review*, *79*(1), 119.

Nissenbaum, Helen. (2011). A contextual approach to privacy online. *Daedalus*, *140*(4), 32–48.

Peters, Dorian, Rafael A. Calvo, and Richard M. Ryan. (2018). Designing for Motivation, Engagement and Wellbeing in Digital Experience. *Frontiers in Psychology*, 9.

Roessler, Beate. (2005). *The Value of Privacy*. Polity Press.

Ryan, W. S., & Ryan, R. M. (2019). Toward a social psychology of authenticity.
Review of General Psychology, *23*(1), 99-112.

[9] See for example Shevlin 2024; Calvo *et al.* 2020 and Peters, Calvo, and Ryan 2018.

Ryan, Richard M., and Edward L. Deci. (2017). *Self-Determination theory: Basic psychological needs in motivation, development, and wellness.* Guilford Press.

Susser, Daniel, Beate Roessler, and Helen Nissenbaum. (2019). Online manipulation: Hidden influences in a digital world." *Georgetown Law Technology Review*, *4*, 1–45.

Thomson, Judith Jarvis. (1975). The Right to Privacy*. Philosophy and Public Affairs*, *4*(4), 295–314.

Véliz, Carissa, ed. (2024). *The Ethics of Privacy and Surveillance*, Oxford University Press.

Data Protection and Generative AI – Policy, Regulation, and the Way Forward

Dr. Stanley Lai, Afzal Ali and Kan Jie Marcus Ho
Allen & Gledhill LLP Singapore

The proliferation of Artificial Intelligence ("AI") has led to paradigm shifts in the context of innovation. With rapid advancement in technology in the past twenty to thirty years, large swathes of data were being generated, collected, and used. It was quickly recognised that this affected all facets of society, and that rules and regulations were urgently required to prevent the unfettered flow and (mis)use of data. Examples of such regulations included the groundbreaking General Data Protection Regulation ("GDPR"), and Singapore's Personal Data Protection Act ("PDPA"). However, just over a decade after the enactment of such rules and regulations, another paradigm shift is on the horizon. Artificial intelligence and generative intelligence are radically transforming how data can be interpreted, used, and presented. It has validly been pointed out that such generative artificial intelligence could bring forth a new epoch of data synthesis and augmentation. This paper discusses how policy and regulations can work to address issues surrounding the use of input data, which is critical to generative AI. Specifically, it will examine whether input data should be considered "personal data" and thus caught by the GDPR or Singapore's PDPA; whether there is a recourse for emotional harm caused by content generated using such data. It will also discuss some of the current limitations and gaps that exist in the current regulatory framework. It is hoped that this discourse will further the continuing dialogue on the intersection between data protection and artificial intelligence, particularly in the domain of Generative AI and Data Protection.

Keywords: Data Protection, General Data Protection Regulation, Comparative Law, Technology Law, Singapore Law

Introduction

The proliferation of Artificial Intelligence ("**AI**") has arguably led to paradigm shifts in the context of innovation, with technologies such as generative AI, machine learning, and cloud computing becoming increasingly pivotal for businesses and organisations. Indeed, Klaus Schwab, Executive Chairman of the World Economic Forum, has persuasively argued that we are now in the fourth industrial revolution, where *"fusion of technologies"* have blurred the lines between the real, digital, and living worlds (Schwab, 2016). With the first industrial revolution having been mainly powered by water and steam power,[10] the second with electricity (Schwab, 2016), and the third with computers and gadgets,[11] it has now emerged, as famously predicted by mathematician Clive Humbly, that "data" – this broad catch-all description for all types of information capable of being stored – is the oil driving the fourth industrial revolution (Charles, 2013).

Historically, the common law has not regarded information as property (*Phipps v Boardman*, 1967). While this position may with time shift, it is fair to say that data is now seen and accepted as having tremendous value. With rapid advancement in technology over the last twenty to thirty years, large swathes of data were being generated, collected, and used. It was quickly recognised that this affected all facets of society, and that rules and regulations were urgently required to prevent the unfettered flow and (mis)use of data. One key legislation which emerged was the General Data Protection Regulation ("**GDPR**"), an overarching data legislation governing the European Union (EU) which sought to provide a comprehensive reform of the existing rules, which was adopted at a time when the internet was still in its infancy (EDPS, *n.d.*). According to the European Data Protection Supervisor, this legislation was needed given that over the last 25 years, technology has transformed our lives in ways nobody could have imagined, and hence a

[10] *Ibid.*

[11] Sakhapov & Absalymova, *Fourth Industrial Revolution and the Paradigm Change in Engineering Education*, MATEC Web of Conferences, 245, 12003 (2018).

review of the rules was needed (*Phipps v Boardman,* 1967). Simply put, the primary purpose of the GDPR was to grant individuals substantive rights in relation to and over their personal data. This was critical at a time when corporations were increasingly unlocking the value of personal data with little or no regulation. Hence, as Hoofnagle, Sloot & Borgesius rightly note, the GDPR *"attempts to put privacy on par with the laws that companies take seriously"*. Indeed, prior to this regulation, it was highlighted that large data companies faced low fines, with there being almost no deterrent effect for the unfettered use of personal data, thereby leading to an imbalance in power (Hoofnagle, Van der Sloot, and Borgesius, 2019). This finally changed following the GDPR's enactment.

Singapore was no different. It recognised that personal data about an individual stood on a different footing from other types of data. The Personal Data Protection Act 2012 ("**PDPA**") was thus enacted to provide a baseline standard of protection for personal data in Singapore. This is the central legislation in Singapore that governs the collection, use, and disclosure of individuals' personal data by organisations (Personal Data Protection Commission, *n.d.*). Chik rightly highlights that this legislation is timely, as *"the digital era poses increasingly greater challenges to the integrity of informational privacy for many reasons"* (Chik, 2013). Following the enactment of this legislation, non-complying organisations risk facing regulatory sanction as well as private civil action should they not handle personal data properly, with the due care that is required as set out in the PDPA.

Just over 10 years after the PDPA's enactment, another paradigm shift is on the horizon. Artificial intelligence and generative intelligence are radically transforming how data can be interpreted, used, and presented. As has been pointed out elsewhere, even within the field of artificial intelligence, the shift has been explicitly evident, with the function of AI having moved from simply analysing expansive datasets to actively generating innovative content (Du *et al.*, 2023). Indeed, consider AlphaGo's victory over the Go world champion in 2016 (Vincent, 2019). to ChatGPT's advanced conversational capabilities,[12] to the creative limits of Midjourney, whose artwork "Théâtre D'opéra Spatial" won the Colorado State Fair (Metz, 2022). These trends will only continue as the full limits of AI are explored. It has validly been pointed out that such generative artificial intelligence could bring forth a new epoch of data synthesis and augmentation, predictive analysis and management, and personalised user interaction (Metz, 2022), which brings in new unique opportunities and challenges for the world ahead.

Generally speaking, generative AI works by using neural networks to identify the patterns and structures within existing data to generate new and original content (NVIDIA, *n.d.*). Through learning the patterns and the structure of their input training data, generative AI tools are able to generate "new data" with similar characteristics AI (Verify, 2024). It is not the purpose of this article to explore the nuances in such training models or the future of generative AI. Instead, the purpose is a more modest one in examining whether existing data protection law is fit for purpose.

It is immediately evident that the use of data to train AI may engender potentially thorny legal issues. Take, for example, a situation where input training data is used to generate defamatory content, which then causes emotional distress. What duties do such AI companies owe to data subjects, if any? When does input data cease to become "personal data" (and consequently fall outside the remit of the PDPA?) Ye, Yan, Li, & Jiang (2024) opine that the rapid development of generative AI has arguably increased personal data risks, particularly in the context of AI pre-training. This is because generative AI consumes vast amounts of personal data while operating in a "black box." Yet, personal data is needed to complete the deep learning procedures that are required for the AI to gain its full potential. The use of such personal data, therefore, attracts scrutiny under the GDPR and PDPA, which were not specifically enacted with Generative AI in mind.

[12] ChatGPT, OpenAI, GPT-4 is OpenAI's most advanced system, producing safer and more useful responses.

In this regard, this article seeks to address some of these questions by examining existing GDPR regulations as well as provisions in the Singapore Personal Data Protection Act in an attempt to identify the possible lacunas in the evolving world of data protection law. The authors hope that this will further the continuing dialogue on the intersection between data protection and artificial intelligence, particularly Generative AI.

1. AI in the context of the GDPR and Singapore's PDPA

Ye *et al.* highlight that OpenAI uses three primary classes of data to train ChatGPT: data that is publicly available on the internet, data that it licenses from third parties, and data from its users or its human trainers. Although conversations with generative AI may not "overtly include direct identifiers like real name or phone numbers," they touch upon user's life experiences, work status, as well as recent thoughts, which can potentially reveal one's identity (Ye, *et al.,* 2024). These issues are best illustrated with the following hypothetical. Take John, an individual who enters his personal data as input into the ChatGPT system, and information relating to his own personal life, such as his hobbies, this being the fact that he likes to play the trumpet, and he then uploads a photo of himself onto the AI input system. John thinks that only he can access the data about himself. But what he does not know is that his data enters the open pool of training data. Thereafter, another anonymous user prompts ChatGPT to generate a funny drawing of a man playing the trumpet – leading to ChatGPT generating a relatively playful take on John's own image and going to the extent of naming this image John when prompted by another user. This is then published online, leading the real John to suffer emotional distress. Who, in this case, should be liable, if at all? What exactly is the personal data that is involved? Is generated data that is "inferred" by the AI considered personal data as well?

2. Is Inferred Personal Data Personal Data?

To answer this question, it would be appropriate to begin with the definition of personal data in the GDPR before examining specific provisions in the PDPA. Article 4(1) forms the definition of personal data, which reads that – (European Parliament and Council, 2016).

"Personal data" means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier to one or more factors specific to the physical, psychological genetic, mental, economic, cultural, or social identity of that natural persons".

From the definition, a key plank of personal data involves the concept of *identifiability*. Stated briefly, identifiability is about the conditions in which a set of data – even if ***not*** linked to a person – is still considered as personal data because it is possible to identify a person from existing data. In this regard, Recital (26) provides further guidance, highlighting that the objective factors which one should consider would be the costs and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments (European Parliament and Council, 2016).

Similarly, in the context of the PDPA, *s* 2 defines personal data as data, whether true or not, about an individual who can be identified – (a) from that data or (b) from that data and other information which the organisation has or is likely to have access to (Government of Singapore, 2020).

This concept of identifiability may be difficult to apply in the context of Generative AI. Given that generative AI systems are often trained by large data sets, including the input data that can be personal data, the issue that arises is whether inferred data (i.e. output data) **<u>is</u>** personal data which is then governed by the GDPR or PDPA. Consider this situation: Assume that a person's physical or mental health information can be inferred from input entered by a person with regard to his daily routine or his food consumption information by the AI. The question then turns towards whether this physical or mental health information is personal data. Indeed, the "inference" by the AI might ultimately not be valid, for example, if it is simply a probabilistic guess by the algorithm. The answer to this question can have far-reaching consequences, particularly as deeming such information as personal data triggers all the data protection obligations, be it under the GDPR or PDPA.

To answer this question, inspiration might possibly be drawn from how past cases in the European Union have been decided.

The European Court of Justice ("**ECJ**") was presented with two requests in two sets of proceedings. In this joint case, individuals sought to obtain a copy of various administrative documents that was drafted with regard to their residence permits. The officials, at first instance, refused these requests (CJEU, 2014). The officials argued in this case that although it was true that information provided could constitute personal data, information which required an abstract legal interpretation cannot be deemed to be personal data (CJEU, 2014). The ECJ held that the input data (such as the applicant's name, date of birth, and the like), as well as the holding by the Minister (that the residence permit was to be denied), were personal data (CJEU, 2014). What was not personal data, however, was the **legal analysis** by the Minister in reaching his decision. This is because the legal analysis was simply information "about the assessment and application by the competent authority of that law to the applicant's situation, that situation being established *inter alia* by means of the personal data relating to him which that authority has available to it" (CJEU, 2014).

However, the decisions do not all speak with one voice. In a different case heard by the ECJ, involving a Data Protection Commissioner's refusal to give an individual access to the corrected script of his examination, somewhat surprisingly, the ECJ held that the examiner's comments, which included the examiner's reasoning, were regarded as personal data. The ECJ held that the content of an examinee's answers was personal data – in addition to information as it related to his handwriting (*Peter Nowak v Data Protection Commissioner*, 2017). The ECJ went even further, holding that the information in the comments of an examiner with respect to the candidate's answers is information relating to the candidate (*Peter Nowak v Data Protection Commissioner*, 2017). Perhaps recognising the far-reaching consequences of its decision and the potential absurd results that might arise if taken too far, the ECJ then held that although such comments constituted personal data, the right to rectification (one such right as provided to data subjects under DPR) did not extend to the correction of an examinee's answers or the examiner's comments (*Peter Nowak v Data Protection Commissioner*, 2017). It reasoned that the assessment of whether personal data is accurate or complete must be made in light of the purpose for which that data was collected. That purpose exists, as far as the answers submitted by an examination candidate are concerned, in being able to evaluate the level of knowledge and the competence of that candidate at the time of the examination. Such errors in any answers do not represent inaccuracy, the existence of which would give rise to a right of rectification. Indeed, such a holding applied to the examiner's comments as well (*Peter Nowak v Data Protection Commissioner*, 2017). Notwithstanding this, the right of access continues to subsist, given that it was personal data about the candidate (*Peter Nowak v Data Protection Commissioner*, 2017).

How, then, does one deal with inferred data about someone created by generative AI? In answering this question, the Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679 states that "profiling can create special category data by inference from data which is not special category data in its own right but becomes so when combined with other data. For example, it may be possible to infer someone's state of health from the records of their food shopping combined with the data on the quality and energy contents of their food" (European Commission, 2018). In this context, profiling has been defined by the Guidelines to be "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular, to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements" (European Commission, 2018). Relying on this guidance, it is argued that in the context of profiling, such inferences (or instances of inferred data) ought reasonably to be considered as personal data. However, going further, should all inferred data created by AI from personal data about an individual be itself regarded as personal data? This article argues that there is good reason to consider such inferred data as personal data. Indeed, this conclusion is supported by the conclusion of the working party, which stated that in the case of automated profiling, a data subject ought to have the right to access both the input data and the conclusions which could be inferred from such data (Article 29 of the Data Protection Working Party, 2016). Such a conclusion would have the effect of requiring AI developers to set clear boundaries and policies in the context of generative AI as to what output can come out of the AI system.

How then may these principles be extended in the context of Singapore's PDPA? According to the PDPA, personal data is data, whether true or not, about an individual who can be identified – (a) from that data or (b) from that data and other information which the organisation has or is likely to have access to (Government of Singapore, 2020). In particular, the Advisory Guidelines to the PDPA states that there are two principal considerations to determining whether something constitutes personal data. The first consideration would be the **purpose** of the information, and the second would be whether a subject would be **identifiable** from that data (Personal Data Protection Act Advisory Guidelines, 2022). The PDPA advisory guidelines do not go further to address inferred data – making this issue a novel one for Singapore. We might find further guidance in the Personal Data Protection Commission's Guide to Basic Data Anonymisation Techniques, which states that it is possible for "certain information" to be inferred from de-identified data but admits the "problem of inference is not limited to a single attribute, but may also apply across attributes, even if all have had anonymisation techniques applied" (Personal Data Protection Commission, 2024). Although a useful starting point, it does not entirely address the questions surrounding inferred data. In this regard, it is suggested that Singapore ought to follow in the EU's footsteps and deem those inferences created by generative AI in the context of profiling to be considered as personal data.

Drawing from this, we consider that conclusions or inferences generated by generative AI could properly be considered as personal data, hence covering the vexed issue of profiling as well.

With this, we turn now to examine the issue of profiling in more detail and its struggles with existing Data Protection law, particularly what rights data subjects should have over data in which they are profiled.

3. The Quandary of Profiling in Data Protection Law – What rights might a Data Subject have?

As Du and others posit, the exponential growth of generative adversarial networks ("**GAN**") has been a foundational technique of generative AI (Du *et al.*, 2023). Brownlee explains GAN in simple terms, highlighting that it is a way of using "generative modelling", which "is an unsupervised learning task in machine learning that involves automatically discovering and learning the regularities or patterns in input data in such a way that the model can be used to generate or output new examples that plausibly could be drawn from the original dataset". According to Brownlee, this is a clever way of training generative AI. The way GAN works involves two key components – the first being the generator model, and the second the discriminator model. The generator model creates new examples using the data set that is provided, whilst the discriminator model performs the function of discriminating the real from the fake. This process is repeated many times, at least until the discriminator can be tricked only half the time, following which the generative AI would then be at an adequate level to generate inferences for the user (Brownlee, 2019).

The problem that data protection law has with this development is as follows. There are huge swatches of input data, some of which are personal data, which could at any one time be sent into the generative AI to be processed by the GAN. Inferences can then be drawn from such data – hence the term “profiling”.[13]

We return now to the example of John and the trumpet. Assume further that a generative AI model has been developed to identify the correlation between educational qualifications and number of instruments played. For every individual in such a case, there would be a whole host of personal data, such as educational level, music preferences, as well as instruments played. When ran through the GAN network, the algorithm will generate examples using the input data, whilst the discriminator will then draw conclusions until it reaches a reasonable level of accuracy. How the GAN would handle these data would be through the drawing of correlations, for example, how higher education might be linked to an increased number of instruments played. This correlation and the algorithm developed is likely not personal data. Instead, this is group data and does not fall within the definition of personal, data whether presented in the GDPR or PDPA.

In this case, assume that John then inputs his data into the generative AI model, which then makes a prediction as to the number of instruments that he plays. Here, one might properly argue that the data provided by John is personal data, whilst the inference as to the number of instruments he played is inferred data, which belongs to him as well.

The implications of such a conclusion can indeed be far-reaching. If it is John’s personal data, should it then be within John’s rights to require the generative AI company to accord him obligations *vis-à-vis* his inferred personal data? As Ye *et al*. (2024) correctly point out, this does not appear to accord with the common understanding of Generative AI .[14] Quach has posited that the output of GPT-2 included at least 0.1% of personal information, including names, addresses, and the like (Quach, 2021). Indeed, the CEO of OpenAI himself admitted that some ChatGPT users could access other’s conversation histories as a result of problems with the GPT open-source database (Haughey, 2023).

Wachter and Mittelstadt have argued towards a right over inferred data, which, according to them, would be an *ex-ante* justification to be given by a data controller (*i.e*., the AI company) as to whether an inference is reasonable, albeit such rights should only apply to “high-risk inferences” drawn through big-data analytics which are privacy-invasive or damaging or have low verifiability. The reasons why such a right is required, according to them, is because “such data draw on highly diverse and feature-rich data of unpredictable value, and create new opportunities for discriminatory, biased, and invasive decision-making” (Wachter and Mittelstadt, 2019). These scholars argue that presently under the GDPR, such individuals are granted little to no oversight of how their personal data has been used to draw inferences about them, in effect according “economy class” status to such data. This is particularly the case as it relates to the data subject’s right to know (Articles 13-15), rectify (Article 16), delete (Article 17), object to (Article 21), or portability (Article 20) (Wachter and Mittelstadt, 2019).

The scholars go further to suggest that for an inference to be deemed reasonable, the inference should fulfil the three criteria of (a) acceptability, (b) relevance, and (c) reliability. Limb (a) requires the input data to be normatively acceptable (*i.e*., race or sexual orientation should be excluded); limb (b) requires the inferred data to be relevant for the chosen processing purpose or type of automated decision (*i.e*., this requires the data to have a link to the processing purpose); and limb (c) requires the data used must be accurate and reliable (and not from dubious sources) (Wachter and Mittelstadt, 2019).

While, in one view novel, this can be seen as a way forward for Data Protection Law. As examined in Joint Cases C-141 and 372/12, as well as Case C-436/16 above, it is reasonably clear that inferred personal data does constitute personal data when construed in the broad sense and that, therefore, the rights to know, delete, object to, or port are available, albeit the right to rectify has been limited to a certain

[13] This has been defined above.

[14] Ye, Yan, Li, Jiang, *Privacy and Personal Data Risk Governance for Generative Artificial Intelligence: A Chinese Perspective*.

extent, as suggested by Case C-436/16. It is not that far of a stretch to then provide for a right to a reasonable inference. This would avoid any ambiguity in these concepts and provide clarity for generative AI companies when developing their ethical and operational policies to operate within certain pre-defined limits. The authors do not consider that requiring responsible use would hamper innovation. On the contrary, it would foster innovation with the right ideals and boundaries.

This suggestion follows closely to the Singapore Personal Data Protection Commission's ("**PDPC**") Model AI Governance Framework, which is largely undergirded by the principles that the decisions made by AI ought to be explainable, transparent, and fair, as well as the fact that AI systems should be human-centric (Personal Data Protection Commission Singapore, 2020). In this regard, a possible step forward for Singapore's PDPA would be for this right over inferred data to be reasonable to also apply in the context of inferred personal data in Singapore, which would be particularly appropriate in light of the Model AI Governance Framework.

4. The Question of Damages in the Context of Generative AI

We turn now to address one final and perhaps the most practical question in this context – damages. In the hypothetical given earlier, an image created by generative AI has led to John suffering emotional distress. The nub of the issue, therefore, concerns whether John, as a private party, has any cause of action against the AI company for a remedy from emotional distress. This article will, therefore, walk the reader through a two-part analysis. First, it will be considered whether a claim under the relevant data protection statute even arises. Second, it will be considered what exact obligation is typically breached, in the context of generative AI.

For an AI company to *owe* an obligation towards John, it must first owe rights to John as a data controller or data processor. We shall first examine this framework under the GDPR before moving on to our analysis of the PDPA. Pursuant to the GDPR, Article 82(1) entitles "any person who has suffered material or non-material damage…shall have a right to receive compensation from the controller or processor for the damage suffered".[15] A right to sue may, therefore, be created upon breach of the "Rights of the data subject" undergirded by Chapter 3 of the GDPR, such as the right of access, the right to erasure, the right to restriction of processing, or even an omission to provide information where data is collected from the data subject.

In examining this thorny issue, a case decided by the Court of Justice of the European Union ("**CJEU**") might once again prove instructive. In *UI v Österreichische Post AG*, the Court of Justice of the European Union ("**CJEU**") held that Article 82 of the GDPR does not provide for compensation to be payable for the mere infringement of a data subject's rights. In this regard, the CJEU held that the mere infringement of the provisions is not sufficient to confer a right to compensation (CJEU, 2022). By relying on the plain statutory language of the provision, the CJEU held it is clear from the wording of Article 82 that the existence of "**damage**" which has been "**suffered**" constitutes one of the conditions for the right of compensation, as does the existence of the infringement and of a causal link between that damage and that infringement, with the three conditions being cumulative (CJEU, 2022).

Hence, a mere breach of a GDPR obligation is insufficient. What this means would be that in John's hypothetical, he would have to show that the use of the generative AI company had indeed breached one of his rights, and therefore, he would have to prove damage.

In this regard, the Singapore Court of Appeal judgment of *Reed, Michael v Bellingham (Attorney-General, intervener*) is helpful (*Reed, Michael v Bellingham*, 2022). In that case, the Court of Appeal held that emotional distress was sufficient to constitute the "loss or damage" limb under *s* 32(1) of the PDPA. Applying the principles of statutory construction to *s* 32(1), the Court adopted a wide interpretation of the section, noting that there was nothing found in the plain language of the PDPA which expressly

[15] Article 82, General Data Protection Regulations

excluded emotional distress as a type of damage that was covered by *s* 32(1) (*Reed, Michael v Bellingham*, 2022). In doing so, the Court looked towards the statutory rationale of the PDPA, considering the "vast and ever-increasing volume of personal data being collected and processed increases the risk of misuse of personal data", and that *s* 32 "must have been intended to be effective in guarding the right of individuals to protect their personal data" (*Reed, Michael v Bellingham*, 2022). As such, adopting a wide interpretation would serve to further the statutory purpose of the PDPA, allowing the PDPA to provide "robust protection for individual's personal data" (*Reed, Michael v Bellingham*, 2022). As such, the Singapore Courts held that emotional distress was actionable under the PDPA. This is, of course, still subject to a "strict causal link" *vis-à-vis* a breach of the PDPA and the loss or damage suffered, and no legal recourse will be permitted for minimal loss (*Reed, Michael v Bellingham*, 2022).

Given that emotional distress is claimable under the relevant data protection statutes, it would appear critical to identify the obligation that might be breached in the context of generative AI. In other words, the key is to point to what data subject right would be breached in most cases?

The European Commission has highlighted that a data controller is defined as any company or organisation which determines the purpose for which and how personal data is processed. Deloitte provides a good example of the nuances involved in the context of how a generative AI company operating an app like ChatGPT might function. According to Deloitte, a Generative AI system provider (such as OpenAI), would likely operate as a ***data controller*** as it relates to the first layers of training and input data. At the same time, the provider will also likely act as an independent data controller for all data as well. In this regard, it may also play a dual role of being a ***data processor*** – particularly in the case where the AI company simply licenses this "AI engine" to enterprise customers without any embedded data. Hence, a generative AI system provider can clearly be brought under the governance of the relevant data protection statutes.

The above discussion is likely to be critical as AI further develops. In 2024, Italy's data protection authority had informed OpenAI that ChatGPT clearly violated data protection rules. In this regard, the Italian Data Protection Authority had stated that they suspected ChatGPT to have breached Articles 5 (principles relating to the processing of personal data), Article 6 (lawfulness of processing), Article 8 (conditions applicable to child's consent in relation to information society services), Article 13 (information to be provided where personal data are collected from the data subject), and Article 25 (data protection by design and by default) (Lomas, 2024).

Looking at the list above, it would seem the main obligation that OpenAI has failed to comply with would be the obligation to provide certain information where personal data is collected from the data subject. As Lomas explains, ChatGPT was developed using "masses of data scrapped off the public internet", this being information which "includes the personal data of individuals". Amongst the six legal bases to use such information, Lomas highlights that only two possibilities remain – these being that of consent or legitimate interest (given that OpenAI was told by the Italian Data Protection Authority to remove references to "performance of a contract" as a legal basis).

It is unlikely consent can apply as a legal basis, given that consent (other than the privacy policy) is difficult to obtain from millions of users absent a mandated information notice. As far as OpenAI's privacy policy is concerned, it states that data subjects can "*withdraw their consent – where [OpenAI] rely on consent as the legal basis for processing*" (Open AI, 2024). It is, therefore, likely that should OpenAI not provide such a notice to users upon a user operating ChatGPT, it would likely be in breach of Article 13 of the GDPR.[16]

All the same, it is likely that absent consent, the only other legal basis that remains would be legitimate interests – which requires that the

[16] Article 13(1)(d) GDPR.

processing is necessary for the purposes of the legitimate interest pursued by the controller or a third party, except where such interests are overridden by the interests of fundamental rights and freedoms of the data subject which requires protection of personal data, in particular where the data subject is a child.[17] Whether the collection of such personal data to advance generative AI is a legitimate interest has not been decided yet by the CJEU, and this does remain an open question.

A similar position applies in Singapore as well – *s* 13 requires an organisation not to collect, use, or disclose personal data about an individual unless (a) the individual gives his consent, or (b) the collection, use, or disclosure without the individual's consent is required or authorised under the PDPA (Government of Singapore, 2020). Legitimate interests do exist as a valid legal basis to collect, use or disclose personal data in the PDPA as well – though it remains a question as to whether the legitimate interests of the generative AI organisation do outweigh any adverse effects on data subjects. This question will remain an open question to be decided by the Singapore courts.[18]

Returning to the hypothetical of John and the trumpet, it is, therefore, likely that a data controller, such as OpenAI, would be liable for damages for emotional distress. In any event, generative AI companies should ensure that the decisions made by their proprietary AI are explainable, transparent, and fair. This can be done through privacy by design principles, ensuring an appropriate degree of human involvement occasionally, as well as ensuring that the black box of decision making does not become too opaque at times. Such principles are undergirded by the Model AI Governance Framework by the PDPC and would likely serve as a useful roadmap for generative AI organisations to follow.

Conclusion

A few decades ago, none of us would have imagined the capabilities of AI to develop to such an extent, and it is likely that AI will be – and possibly already is – the mantra of the fourth industrial revolution. This article has explored three key issues, (a) whether inferred personal data by generative Artificial Intelligence can be considered as personal data, (b) the rights which data subjects have over such data, and (c) remedies that can be claimed because of a mishap by a generative AI company. This article has then suggested that the Singapore Model AI Governance Framework is a right step forward, particularly as jurisdictions around the world begin to frame their privacy legislation to handle the new epoch of AI generated data. The future is exciting, and the potential risks of generative AI should not be hidden by its immense potential – with strong privacy laws and adequate guidance, it is likely AI can chart an explainable, transparent, and fair path ahead as we move into the future of tomorrow.

References

AI Verify. (2024). P*roposed Model AI Governance Framework for Generative AI – Fostering a Trusted Ecosystem.* Retrieved from https://aiverifyfoundation.sg/downloads/Proposed_MGF_Gen_AI_2024.pdf

Brownlee, J. (2019). A Gentle Introduction to Generative Adversarial Networks (GANs). Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

Charles, A. (2013). Tech Giants may be huge, but nothing matches big data. *The Guardian.* Retrieved from https://www.theguardian.com/technology/2013/aug/23/tech-giants-data

Chik, W.B. (2013). The Singapore Personal Data Protection Act and an assessment of future trends in data privacy reform. *Computer Law & Security Review.* 29 (5), pp. 554–575, doi: 10.1016/j.clsr.2013.07.010.

Court of Justice of the European Union (CJEU). (2014). *Joined Cases C-141/12 and C-372/12, YS v Minister voor Immigratie, Integratie en Asiel and Minister voor Immigratie, Integratie en Asiel v M and S.* Retrieved from https://fra.europa.eu/en/caselaw-reference/cjeu-joined-cases-c-14112-and-c-37212c-judgment Metz, R. (2022). AI won an art

[17] Article 6(1) GDPR.

[18] First Schedule, Personal Data Protection Act.

contest, and artists are furious. *CNN*. Retrieved from https://edition.cnn.com/2022/09/03/tech/ai-art-fair-winner-controversy/index.html

Court of Justice of the European Union (CJEU), (2022). Opinion of Advocate General in Case C-300/21, Bundesrepublik Deutschland v XT. Retrieved from https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:62021CC0300

Du, H., Niyato, D., Kang, J., Xiong, Z., Zhang, P., Cui, S., Shen, X., *et al.* (2023). The Age of Generative AI and AI-Generated Everything. *ArXiv.* doi: 10.48550/arXiv.2311.00947.

European Commission. (2018). Article 29 Working Party - Newsroom. Retrieved from https://ec.europa.eu/newsroom/article29/items/612053

European Data Protection Supervisor (EDPS). (n.d.). *The History of the General Data Protection Regulation.* Retrieved from https://www.edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en

European Parliament and Council. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. Official Journal of the European Union, L119, pp. 1–88. Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj

Government of Singapore. (2020). *Personal Data Protection Act 2012 (2020 Rev Ed)*, s 2. Retrieved from https://sso.agc.gov.sg/Act/PDPA2012

Haughey, L. (2023) Are your conversations safe?, *Daily Mail.* Retrieved from https://www.dailymail.co.uk/sciencetech/article-11893689/ChatGPT-creator-confirms-bug-allowed-users-snoop-chat-histories.html

Hoofnagle, C. J., van der Sloot, B. and Borgesius, F. Z. (2019), The European Union general data protection regulation: what it is and what it means, *Information & Communications Technology Law,* 28(1), pp. 65–98. doi: 10.1080/13600834.2019.1573501.

Lomas, N. (2024). ChatGPT is violating Europe's Data Privacy Laws, Italian DPA tells OpenAI. *Tech Crunch.* Retrieved from https://techcrunch.com/2024/01/29/chatgpt-italy-gdpr-notification/

NVIDIA. (n.d.) *What is Generative AI.* Retrieved from https://www.nvidia.com/en-us/glossary/generative-ai/

Open AI. (2024). *Open AI Privacy Policy.* Retrieved from https://openai.com/policies/privacy-policy

Personal Data Protection Act Advisory Guidelines. (2022). Retrieved from https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/ag-on-key-concepts/advisory-guidelines-on-key-concepts-in-the-pdpa-17-may-2022.pdf

Personal Data Protection Commission. (n.d.) PDPA Overview. Retrieved from https://www.pdpc.gov.sg/overview-of-pdpa/the-legislation/personal-data-protection-act#:~:text=What%20is%20the%20PDPA%3F,Banking%20Act%20and%20Insurance%20Act.

Personal Data Protection Commission. (2024). Guide to Basic Anonymization Techniques. Paragraph 4.1(c). Available at: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/guide-to-basic-anonymisation-%28updated-24-july-2024%29.pdf

Personal Data Protection Commission Singapore. (2020). *Singapore's Approach to AI Governance.* Retrieved from https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework,

Peter Nowak v Data Protection Commissioner. (2017). CJEU Case C-434/16. Retrieved from https://curia.europa.eu

Phipps v Boardman. (1967). 2 AC 46.

Quach, K. (2021). How to curb GPT-3's tongue. *The Register.* Retrieved from https://www.theregister.com/2021/03/18/openai_gpt3_data/

Reed, Michael v Bellingham (Attorney-General, intervener). (2022). SGCA 60. Retrieved from https://www.elitigation.sg/gd/s/2022_SGCA_60

Schwab, K. (2016) The Fourth Industrial Revolution: What it means and how to respond. *World Economic Forum.* Retrieved from https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

Vincent, J. (2019). Former Go Champion beaten by DeepMind retires after declaring AI invincible. *The Verge.* Retrieved from https://www.theverge.com/2019/11/27/20985260/ai-go-alphago-lee-se-dol-retired-deepmind-defeat

Wachter, S. and Mittelstadt, B. (2019). A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review.* Issue 2.

Ye, X., Yan, Y., Li, J. and Jiang, B. (2024). Privacy and personal data risk governance for generative artificial intelligence: A Chinese perspective. *Telecommunications Policy.* 48 (10), p. 102851, doi: 10.1016/j.telpol.2024.102851.

Are Tech Companies Responsible for Solving the Global AI Divide?

A Practical Exploration of Libertarian, Rawlsian and Utilitarian Points of View

Laura Rachevsky
Lucy Cavendish, University of Cambridge

The global AI divide, marked by the unequal distribution of AI benefits between developed and developing countries, is a pressing ethical concern. This paper examines the moral responsibility of tech companies in addressing this divide, analysing it through the lenses of libertarianism, Rawlsianism, and utilitarianism. It delves into the nuances of each perspective, particularly highlighting their limitations in a global context, and contrasts the current focus on productivity-enhancing AI applications in developed countries with the potential of life-saving AI applications in developing countries. The paper explores empirical examples of tech companies' investments in developing countries, revealing that libertarian and Rawlsian perspectives, despite initial differences, converge in their practical implications on a global scale. Ultimately, it argues that utilitarianism, although not without its challenges, provides the most actionable framework for addressing the global AI divide due to its emphasis on measurable outcomes and its ability to transcend national boundaries. It further performs a simplistic redistribution calculation as a proof of concept to demonstrate how incorporating life-saving AI applications into the benefits calculation can result in different investment recommendations.

Keywords: Global AI Divide, Philosophy, Corporate Responsibility, Global Justice

Introduction

It is frequently noted that one of the issues in contemporary AI ethics is the "AI divide," or the unequal distribution of benefits produced through use of AI technology between the developed and developing countries. Multiple authors point out the fact that "the economic and social benefits of AI remain geographically concentrated, primarily in the Global North" (World Economic Forum, 2023) and some, such as Yuval Harari even go as far as questioning "will the rest of the world just become algorithmic data colonies for AI-dominating countries?" (LSE, 2023).

This divide also manifests itself empirically. A recent report by PwC, "Sizing the prize," seeks to size the benefits most prominent applications of AI would bring as measured by GDP gains. Use cases they consider mainly include productivity enhancing AI use cases, such as driverless cars and trucks, and scaled financial advice and customisation. Based on this their estimate of $15.7 trillion of GDP gains by 2030 is currently split by 79% of the benefit going to developed countries and 21% to developing countries - a stark contrast with population distribution as illustrated in Table 1 below.

Table 1: *Population distribution vs AI Benefits distribution in 2030*

	Population	**GDP gains under current applications ($, bn)**	**Population distribution**	**Benefits distribution**
Developed countries*	1368m	12,500	17%	57%
Developing countries	6632m	9,300	83%	43%
	8000m	**21,800**	**100%**	**100%**

Note: the GDP gains column of the table is sourced from the "Pwc-Ai-Analysis-Sizing-the-Prize-Report", 2017; UN definition of developing and developed countries used (UNCTDA, 2022)

It is rarely debated who is responsible for alleviating the global AI divide. Rare solutions offered emphasise the need for collaboration among governments, multilateral agencies and technology providers without explicitly attributing responsibility to either party.

Given that the vast majority of investment in AI is currently done by private companies, in my paper I explore whether it is the responsibility of companies developing AI to ensure that the benefits of their technology extend to people in the developing countries. I examine this issue from libertarian, Rawlsian and utilitarian points of view in turn. I first explain each stance's likely approach to the debate and delve into nuances within each that highlight the ambiguity of potential conclusions. For each school of thought I choose an example of tech companies' existing investments in developing countries where the ethos behind the investment seems to have been informed by this particular school. I then evaluate the pros and cons of each theory along two dimensions: (1) their usefulness in addressing the issue in a global context as opposed to that of a nation state as this transcendence of national borders is paramount in addressing issue of global AI divide and (2) the plausibility of real implementation of the practical recommendations offered by each theory as apart from providing the most useful methodological framework the best approach should also be judged on its impact.

I choose the above three schools of thought to analyse the issue as they present seemingly radically different answers to the debate in question from denying responsibility of tech companies (libertarianism) to advocating for it (Rawlsian) with utilitarianism falling somewhere in the middle depending on the calculations involved. However, when delving into nuance, it becomes apparent that libertarian and Rawlsian points of view despite initially offering contrarian points of view actually conflate when it relates to handling the issue from a global point of view. In both, responsibility could be assigned to the global tech companies and the degree to which is inconclusive in either. It then becomes a matter of usefulness in a global context, where I argue utilitarianism prevails based on its better accountability for the global nature of AI production and deployment and its grounding in methods similar to those used by tech corporations themselves which increases its likelihood to drive actionable change. I acknowledge that GDP is a metric that originated in the Global North and has since been challenged for undervaluing non-monetary values and explore the potential positive implications of including alternative measures such as Amartya Sen's capabilities systems and community wellbeing under the ubuntu philosophy in later sections. However, I base my argument mainly on GDP-based measures of benefits based on its current prevalence in global decision making and its importance to developing countries for whom a minimum level of GDP is often a prerequisite for achieving other development goals.

Definitions

I am fully aware of the complexities in grouping countries into developed and developing. This is done for conciseness of expression and largely aligns my definition with the UN definition based on scores on various Sustainable Development Goal dimensions (UNSTATS, 2024). My intention is to distinguish between populations of countries that produce AI and/or have economies advanced enough to benefit from productivity-enhancing use-cases of AI and populations of countries where human development and infrastructure indicators are on the lower end and hence where life-saving applications of AI are most needed.

Therefore, I define "AI divide" as inequality in AI benefits distribution as it relates to individuals' effective access to and benefits from AI on a scale comparable across countries.

This is roughly aligned with Beitz's definition of global inequality: "when I speak of inequalities among societies or states, unless otherwise noted, I shall mean this as shorthand for inequalities among the persons who inhabit them taken as a single group." (Beitz, 2001).

Finally, I define responsibility as a fundamental moral obligation and legal accountability stemming from ownership or control over something, in this case of AI technology. This is to contrast the current situation where corporate ESG efforts aimed towards benefiting the developing countries are seen as a "good thing to do" rather than a moral obligation.

1. Review

1.1. Libertarian

I begin examining the question on responsibility over equal spread of AI benefits from the point of view of libertarianism, given its dominance in the legislative landscape of the countries where AI is created. Libertarian thinking is often evoked by tech CEOs when advocating for government non-intervention, such as Tim Cook's statement on US government requests to decrypt iPhone OS that "would undermine the very freedoms and liberty our government is meant to protect" (Cook, 2016).

Libertarian stance on responsibilities of tech companies towards citizens of developing countries would emphasise non-intervention on the grounds of (1) violation of intellectual private property rights, (2) free markets' superiority in addressing issues.

Classic Libertarians such as Locke in his "Two Treatises on Civil Government" established natural rights, which included the right to property, protecting which he saw as one of the key functions of governments: "The great and chief end…of men…putting themselves under government is the preservation of their property" (Locke, 1884). It could be argued that strong IP laws of Western democracies and libertarian-like culture of Silicon Value was what contributed to AI breakthroughs in the first place and therefore companies and their shareholders are entitled to the full benefits of their innovation without an obligation to share it with others.

Libertarian economists would argue that free markets would achieve the goal of bridging the "global AI divide" better than any forced redistribution. Milton Friedman famously said: "The great virtue of a free-market system is that it does not care what colour people are; it does not care what their religion is; it is the most effective system we have discovered to enable people who hate one another to deal with one another and help one another." (Friedman, 1993). Among examples of market efficiency given by libertarians is healthcare provision, where government intervention can lead to price inflation, decreased quality of care due to reduced competition. Drawing analogies with AI, removing obstacles to data access and technology deployment in developing countries will be incentive enough for tech companies to provide most efficient entrepreneurial solutions to developing countries' needs.

1.2. Nuance

According to some libertarian thinkers, resource redistribution could be justified based on the following arguments: (1) rectifying past injustices and (2) protecting positive rights. Nozick, for example, mentions just acquisition and rectifying past injustices where it is plausible to do so. His principle of rectification of injustices in holdings requires that parties be returned to the situation they would have been in had the injustice not occurred. (Nozick, 1974). If a corporation has come to possess technology and profits from it in an unjust way, then they should be redistributed to its original holders. In the context of AI, data ownership comes up often and redistribution on the grounds of data collection practices could be made given not only the ownership of data collected from the developing countries but also an outsized role of developing countries in database creation through data annotation.

The idea of positive rights also allows for some redistribution in cases where severe inequality is preventing citizens from exercising control over their lives and therefore limiting their freedom. For example Vallentyne argues on egalitarian grounds that profits based on natural resource exploitation should be redistributed among global citizens through a "global fund" in an egalitarian manner (2000). It could be argued then that there are some minimum entitlements that each individual has and in order to protect those, some redistribution from corporates to individuals is warranted. Companies in other industries such as construction and healthcare are often mandated by governments to protect such positive rights of their citizens through compulsory licensing and social housing projects.

1.3. Accounting for Global Context

Libertarian thinking often tends to focus on legislation within nation states and tends to downplay the role of global natural rights. Strict interpretations of classical Libertarianism only mandate individual states to protect the natural right to life within its borders. So, Locke viewed protecting rights to life, health and liberty as within the state's mandate (Locke, 1884). There seems to be a dissonance, however, between the idea of the right to life that is "natural" and the fact that in a lot of the developing world this right is routinely violated through deaths from preventable causes. While during the time of Locke's writing the focus on a nation state might have been justified given the context of the emerging US independence movement, in the current globalised world, it is hard to confine the idea of natural rights to a single state. Resource redistribution arguments within libertarianism, such as that of Nozick as referenced above also tend to focus within the boundaries of nation states and while they could be extrapolated to international contexts, they do not explicitly address global injustices. It could be argued that current Western digital technological dominance is based on the history of colonialism and resource extraction that enabled select elites in the developed world to enjoy living and educational standards that allowed them to reach the levels of current technological innovation. "Europe is literally the creation of the Third World. The wealth which smothers her is that which was stolen from the underdeveloped peoples." (Fanon, 2001). This context, however, is largely absent from Nozick's thinking as he focuses on the justification of property rights within a society.

Vallentyne's idea of a "global fund" is a rare example of left-libertarians addressing the issue of global injustice. While the focus on natural resources which are viewed as a commonly owned good is not directly transferable to the issue in question, extrapolating this line of thought could serve as an argument for distributing the benefits of AI broader. As AI is trained on user data and hence could be seen as a public good, the distribution of its benefits should be more equal globally.

1.4. Example

Libertarian thinking is behind some of the current ESG efforts of tech companies. Open-source models are often lauded as prioritising social good over profits. Connectivity-focused projects, such as Meta's Free Basics and Express Wi-Fi aimed at providing affordable Wi-Fi to emerging markets through hotspots are underpinned by the idea that providing opportunities and removing barriers would allow the free market to alleviate economic disparity (Meta, 2024). Mark Zuckerberg in his argument for connectivity mentions "The richest 500 million have way more money than the next 6 billion combined. You solve that by getting everyone online, and into the knowledge economy." (Wired, 2013). One issue with this is that providing connectivity and source models alone rarely leads to progress in developing countries. Developed world software developers benefited most from open sources systems, while applications for developing countries are much harder to come across online. There is a long temporal lag between providing connectivity and benefits of technology being felt economically within communities. Interestingly, the META Express Wi-Fi project has now been scaled down and no tangible results were reported, which is perhaps telling of the efficiency of the approach based solely on providing access without further distributive assistance (TechRadar, 2024).

1.5. Evaluation

Overall, libertarianism is not very instrumental in evaluating whether or not the responsibility over AI benefits redistribution lies with tech companies. The school's coverage of global interdependencies is limited and its strong focus on a single nation state makes it difficult to apply to the globalised nature of AI production and consumption. Additionally, when informing ESG efforts in the real world, the school's recommendations fall short of delivering meaningful results to developing markets.

2. Rawlsian

A classical Rawlsian stance would posit that it is indeed the moral responsibility of companies that create transformational technologies to ensure a more equal global distribution of the benefits of such technologies. Based on the idea of the "veil of ignorance" if a neutral objective person would be deciding which use cases to deploy AI towards, she or he would direct it towards solving the most pressing global issues such as climate change, food security, illiteracy etc (Rawls, 1999). A lot of these use cases are relevant to developing countries, unlike the current productivity focused use cases.

A general criticism of Rawlsian thinking is its impracticality in a world where existing resource distribution is far from the original state. It is implausible that tech companies would agree to develop AI systems without commercial interest at stake and it is likely that in such a case the pace of AI development would be impeded.

Recognising the impracticality of "veil of ignorance", Rawls also argued under his difference principle that inequality could be justified as long as it makes the worst off in society better off: "While the distribution of wealth and income need not be equal, it must be to everyone's advantage, and at the same time, positions of authority and offices of command must be accessible to all." (Rawls, 2005). Every policy and investment decision then needs to consider its impact on the worst-off in society, something which current tech companies' investment principles do not and something that is quite different from libertarian thinking where concern for the worst-off is not a given.

2.1. Nuance

While the difference principle is powerful in putting a condition on inequality-producing decisions, it is difficult to measure what "to everyone's advantage" means. Rawls does not offer a single definition to subjective notions of "worst-off" and "improvement". This could hence be interpreted in a number of ways as it relates to the global AI divide (Rawls, 2005). One interpretation could be that as long as AI development for commercial purposes also funds some socially positive applications no further redistribution is needed. So, if people in the developing countries are slightly better off than what they would have been without any AI development, this is sufficient. In this instance, Rawlsian thinking could potentially paradoxically recommend a similar or lower redistribution of resources than that warranted by Nozick's redistribution principle discussed earlier. While this does not seem to be the intention of the theory, there is the danger that this vagueness could be used to "green wash" corporate ESG efforts.

2.2. Accounting for Global Context

Similarly to libertarianism, classical Rawlsian theory focuses on justice within the domestic nation state. While in his *Law of the Peoples* Rawls does state that "Peoples have a duty to assist other people's living under unfavourable conditions that prevent their having a just or decent political and social regime," he mainly places assistance responsibility with the developed states rather than individuals or corporates (Rawls, 1999). Under this constraint, tech corporations would be responsible for addressing inequalities within their own countries and communities as a priority to those of other nations, which given the fact that developing countries lack AI development capabilities will likely exacerbate rather than alleviate the global AI divide.

Ideas of Charles Beitz extended Rawlsian ideas in a domestic society to our duties as global citizens. He argues that the differences between the domestic and global realms have been overestimated and a lot of the arguments in favor of equality as domestic justice could be applied in a similar way to equality as global justice. "There is a dispute about whether we should understand global justice, so to speak, as an enlarged image of justice in one society – and correspondingly demanding – or rather as a distinct construction, suited to a world that cannot be described as a single society, and therefore as demanding less than its domestic analogy." (Beitz, 2001). He argues that for reasons of shared humanity and interdependence our duties to citizens of other nations are the same as to those in our own countries. Beitz's idea of the "Global Resource" dividend is surprisingly similar to that of left-libertarian Vallentyne and once again could be extended to be a "Global AI" dividend through the notion that AI is a common global good trained on global data and therefore that its proceeds could be distributed among developing nations.

2.3. Example

Some organisations such as OpenAI's original mission was to "ensure that artificial general intelligence benefits all of humanity."(OpenAI, 2024). which appears close to the Rawlsian ethos. While conceding the need for a commercial arrangement to reach scale, OpenAI argues that it "continued to advance our mission by building widely-available beneficial tools" in its recent blog. (OpenAI, 2024). The example given by OpenAI as it relates to developing countries is the Digital Green collaboration in Kenya aimed at improving agricultural knowledge in the current climate change affected environment. From reading the customer success story on OpenAI's website – it is not clear what role OpenAI itself played beyond providing the model. It would seem that the bulk of the effort fell with the Digital Green organisation itself – an NGO with diverse funding sources. Furthermore, this is the only developing country case study listed on OpenAI's website with other examples covering developed country applications. Additionally, and similarly to other tech firms, OpenAI was criticised for its working standards used for human labellers in Kenya in terms of wage levels (between around $1.32 and $2 per hour) and working conditions (Time, 2023).

OpenAI's example is illustrative of the idealism of the Rawlsian school of thought that appeared to have been "reality checked" in this case.

2.4. Evaluation

Overall, although later Rawlsian thinkers do explore the context of justice in the globalised world, the solutions they offer are surprisingly similar to those offered by libertarians and rather impractical and hence unlikely to have impact in the real world.

3. Act Utilitarianism

Act utilitarianism would approach the question of how a company should invest its resources, based on what would produce the best ultimate outcome for the majority of people (Bentham, 2012). I will proceed with illustrating a hypothetical approach to such evaluation below, with the purpose of illustrating the utilitarian approach rather than reaching a conclusion on the recommended benefit reallocation amount.

Going back to the PWC report referenced earlier – AI applications included in this report are those related to increased efficiency and accuracy in applications most relevant to developed countries. They measure gains in productivity and extrapolate this to resultant economic benefits. Based on this, applications in developing countries are relatively limited considering the smaller sizes of their economies.

AI use cases not considered in the report are those related to death prevention and improvements in basic quality of life, such as alleviating malnutrition, increasing literacy and preventing death and displacement through natural disasters. It could be argued that such applications will improve outcomes in emerging markets by the product of the lives saved and the current GDP per capita (or a significant proportion of it) as lives saved will create additional economic benefits proportionate to their number.

I attempt to make a calculation below where including such AI applications into the equation will suggest an optimum redirection of investment from the development of current commercial AI applications into life-saving AI applications more relevant to developing countries. Before doing so, I would like to reiterate that this is purely to illustrate the benefits and pitfalls of a utilitarian approach to this issue. Through performing this simplistic calculation I illustrate the possibility of doing so with an alternative objective in mind – calculating the economic benefits of life-saving AI applications. To my knowledge no comprehensive attempt to do so has been done globally and therefore there is no existing body of expertise. Through my simplistic demonstration I call on this viewpoint to be taken into account in similar future evaluations.

Following this approach, redistributing 12% of the benefits from current commercial uses to those designed to save lives in the developing world (aimed at preventing hunger, natural disaster, and treatable diseases) is the equilibrium point. This is illustrated in Table 2 below with corresponding assumptions and caveats.

Table 2: Illustration of hypothetical utilitarian approach to sizing benefits of life-saving AI applications in developing countries

	Population	**GDP gains: current applications ($, bn)**	**Lives saved: life-saving applications**	**Lives with significantly improved quality**	**GDP per Capita ($)**	**Benefits Redistribution ($)**	**Benefits Redistribution (%)**
Developed countries	1368m	12,500	18m	409m	6,770	-1,508	-12%
Developing	6632m	9,300				1,508	16%
	8000m	**21,800**					
	Population	GDP gains: current applications ($, bn)	Lives saved: life-saving applications*	Lives with significantly improved quality**	GDP per Capita ($)	Benefits Redistribution ($)	Benefits Redistribution (%)
Developed countries	1368m	12,500	18m	409m	6,770	-1,508	-12%
Developing	6632m	9,300				1,508	16%
	8000m	**21,800**					

The methodology of my simplistic exercise could have been greatly improved given time and access to experts in relevant domains. However, I want to acknowledge that even the most comprehensive methodology would pose a number of challenges.

1. **Uncertainty**: one of the main criticisms of utilitarianism is that the outcomes of actions are extremely hard to predict especially when complex concepts or new technologies are involved. Predicting the impact of AI on a broad range of applications such as healthcare and agriculture accurately is extremely difficult and might lead to misleading conclusions.

2. **One unit of measurement**: Quantifying the value of a human life, education, and food safety along in the same monetary units (GDP gains) as improvements in financial planning and business productivity is not only methodologically difficult but ethically hugely problematic. As the above exercise shows, the result of factoring in global suffering only results in a modest recommendation for redistribution which is the by-product of the assumption that productivity in the workplace could be compared to the value of a human life.

3.1. Nuance

Using another quantitative measure of investment such as the OECD Better Life index, or Amartya Sen's capabilities framework, would have likely resulted in an even more favourable recommendation distribution in favour of life saving applications of AI. While admittedly facing the same methodological issue of scales, such a measurable, visual approach could act as a meaningful call to action, such was the case with Peter Singer's famous work “The Life You Can Save”. Impact dimensions on “The Life You Can Save” websites such as “Health”, “Education” and “Living Standards” are well defined with corresponding indicators, definitions and measurement systems. This acts as a motivator for investors based on utilitarian grounds. This approach works within frameworks understood and accepted by major philanthropic investors and corporations and in the case of individual donations has significantly elevated the profile of global philanthropy among the general public. There are some reality checks however that need to be kept in mind despite the powerful call to action of this methodology. Singer has famously advocated for directing up to ⅓ of one's income to philanthropy; however most individuals are not close to this suggested amount (WSG, 2015). Similarly, in the case of tech companies' investment distribution, we are unlikely to reach this “north star”, however we have a chance of making some progress towards this goal.

The highly numerical and material nature of utilitarianism also opens it to criticism from non-western schools of ethical thought such as ubuntu for example, which emphasises the wellbeing of the entire community and values the wellbeing of all individuals in its own right. My proposed methodology implicitly justifies actions that benefit the majority in a zero-sum game with implicit need for sacrificing the well-being of one group to increase the well-being of another. Ubuntu, on the other hand, rejects the notion that the well-being of some can be sacrificed for the benefit of others. Exploring whether both goals can be achieved simultaneously is a valid direction of enquiry that would enrich the argument as part of further exploration.

3.2 Accounting for Global Context

Utilitarianism does provide a useful basis for addressing the issue through allowing for transcendence of national boundaries. Peter Singer in "The Most Good You Can Do" provides a compelling argument why philanthropists should invest in alleviating global poverty as opposed to poverty in the developed world (Singer, 2015) . While not denying the negative effects of poverty in developed countries, he points out the various security mechanisms available to citizens of the developed countries through taxation for example. He concludes that there is a wide gulf between poverty in the developed and developing world and the value of donations (in life outcomes) is far greater in the developing world: "their dollars go much further when used to aid those outside the affluent nations" (Singer, 2015). Singer's writing has been influential in alleviating global poverty and its explicit address of the global nature of inequality makes it relevant to evaluating the issue of the global AI divide.

Utilitarianism also has a visual and measurable quality to it, something that makes it compatible with driving change within corporations. The presentation of productivity and lifesaving use cases side by side, while problematic, does allow for visual and measurable accountability in a format familiar to the corporate world. Like Peter Singer's case for individual donations, it helps substantiate the claim in measurable terms. The investment disparity, if presented internally within corporations, could ignite employee activism towards influencing corporate investment decisions or towards institutionalising individual voluntary time donations in an arrangement similar to the legal profession's pro bono practice.

3.3. Example

As opposed to Meta's connectivity projects, Google's Build for Africa investment announced in 2021 does attempt to combine connectivity provision (through a subsea cable) with initiatives aimed at local talent development (through an AI research facility in Ghana) and product usability improvements (increased language inclusion and Maps coverage) along with start-up empowerment programmes. Google's announcement invokes utilitarian values through its explicit mention of benefiting the lives of most people: "benefits of the digital economy for more people by providing useful products, programmes and investments" (Gajria, 2022). Benefits are also quantified in GDP terms similarly to the PWC report and anchoring to $1Bn as the investment sum, shows how numerical grounding could be instrumental in motivating corporate action.

While the focus on talent development and product localisation is a move in the right direction – it is worth noting that the $1Bn investment (spread over 5 years) is equivalent to only 0.4% of Alphabet's 2023 R&D budget (Alphabet Annual Report, 2023). This is illustrative of the pitfall of utilitarian thinking in not recommending enough redistribution despite being a useful framework for motivating corporate action in general. The results of this programme are yet to be seen.

3.4. Evaluation

Overall, however, despite its methodological and ethical challenges, I would rate utilitarianism as the more useful theory among the three considered in evaluating the issue of the AI global divide. This is based on its rich recent body of thought on global justice and its measurable and visual basis that is likely to lead to actionable change both at the corporate level and at the level of individual employees.

Conclusion

In conclusion, the three schools of thought come to three different conclusions on the moral debate of whether or not it is the responsibility of tech companies to alleviate the global AI divide. Libertarianism would argue that it is not the responsibility of the tech companies, Utilitarianism that it is, to a certain quantifiable degree, while Rawlsian that it definitely is as part of a moral imperative. Despite these broad-based conclusions, when factoring in nuance, there is room for an increased benefits redistribution within each school of thought, based either on rectifying past injustices, reframing benefits in terms of increases in "capabilities" rather than GDP or fully extending the veil of ignorance principle to the global context.

Libertarian focus on the nation state makes it hard to make judgements in a globalised setting and due to its original non-intervention stance, those recommendations aimed at redistribution on the basis of past injustices or positive rights, such as a Natural resource tax are quite unrealistic. Rawlsian stance although going beyond the nation state eventually is fairly vague in defining what making those globally worse off better is. Actionability of its recommendations such as the Global Resource dividend is also fairly improbable given its highly conceptual nature. Traces of libertarian and Rawlsian thought can be seen in various empirical examples of existing tech companies' investments in developing countries such as ESG and connectivity directed efforts and partnerships with local NGOs where both can result in quite in "greenwashing" one in line with its intention and the other due to its impractical nature.

Utilitarianism scores highly as it relates to evaluating this moral debate for two reasons: its well-established body of work on global inequality and actionability of its recommendations. The visual nature of the results it demonstrates is likely to draw more realistic action such as employee activism and increased collaboration with unilaterals as a result. Therefore I argue that this is the most useful framework in this case albeit not problem-free especially as it relates to its methodological complexity and ethical challenges. While utilitarianism does set a good basis for empirical actionability, in order for resource redistribution to increase in a meaningful way we need to employ cognitive behavioural tools, such as for example drawing from the more successful examples of cooperation within climate change and nuclear regulation domains. This could be a good next step to consider to further enrich this argument further.

Incorporating the viewpoints of other schools of thought, such as duty ethics for example, while out of scope of this paper would enrich the analysis and I recommend these as next steps of this line of inquiry. So, exploring how duty to the company shareholders might negate the recommendations made under utilitarianism. This would also have practical implications, for example due to the need to change the governance structures of corporations to mitigate this obstacle. On the contrary, duties of individual employees to their communities of origin could further strengthen the plausibility of some courses of action such as employee activism.

References

Beitz, Charles R. (2001). Does global inequality matter? *Metaphilosophy*, 32 (1-2), 95–112.

Bentham, J. (2012). *An introduction to the principles of morals and legislation*. Dover Publications.

Fanon, Frantz. 2001. *The wretched of the earth*. Penguin Modern Classics.

"Friedman-Government-Problem-1993.pdf."
n.d. Retrieved from
https://www.hoover.org/sites/default/files/uploads/documents/friedman-government-problem-1993.pdf.

Gajria, Nitin. (2022). "Delivering on Our $1 Billion Commitment in Africa."
Retrieved from https://blog.google/around-the-globe/google-africa/delivering-on-our-1b-commitmen t-in-africa/

Levy, Steven. (2013). Zuckerberg explains Facebook's plan to get entire planet online. *Wired*. Retrieved from
https://www.wired.com/2013/08/mark-zuckerberg-internet-org/.

Locke, J. (1884). *Two treatises on civil government*. George Routledge and Sons.

Nozick, R. (1974). *Anarchy, state, and utopia*. Basic Books.
Pwc-Ai-Analysis-Sizing-the-Prize-Report.pdf..
(2017). Retrieved from
https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf.

Rawls, John. (1999a). *The law of peoples*. Harvard University Press.
Rawls, John. (2005b). *A theory of justice*. Belknap Press.

Singer, P. (2015). *The most good you can do: How effective altruism is changing ideas about living ethically*. Yale University Press.

Slater-Robins, Max. (2022). "*Meta Is Shutting down Its Express Wi-Fi Service.*" TechRadar Pro. February 1, 2022. https://www.techradar.com/news/meta-is-shutting-down-its-express-wi-fi-service.
United Nations Statistics Division. n.d. Methodology. Retrieved from https://unstats.un.org/unsd/methodology/m49/.

Vallor, Shannon, ed. (2000). *Left-libertarianism and its critics: The contemporary debate*. Palgrave Macmillan.

Yu, Danni, Hannah Rosenfeld, and Abhishek Gupta. (2023). The 'AI Divide' between the global north and global south. Retrieved from https://www.weforum.org/agenda/2023/01/davos23-ai-divide-global-north-global-sout h/

Wolfe, Alexandra. (2015). Peter Singer on the ethics of philanthropy. WSJ Online. April 3, 2015. Retrieved from https://www.wsj.com/articles/peter-singer-on-the-ethics-of-philanthropy-1428083293

Getty Images v Stability AI: Why Should UK Copyright Law Require Licences for Text and Data Mining Used to Train Commercial Generative AI Systems?

Zoya Yasmine
Somerville College, University of Oxford

In 2023, Getty Images commenced legal proceedings in the United Kingdom High Court against Stability AI. Getty Images claims that 7.3 million images from its database were unlawfully used to train Stability AI's generative Artificial Intelligence system. Drawing inspiration from Getty Images v Stability AI, this paper addresses the complexities surrounding copyright protection for text and data mining (TDM) in the UK. It argues that expanding Section 29(A) of the Copyright, Designs and Patents Act 1988 to exempt commercial AI developers from TDM licensing obligations would undermine the creative sector and hinder responsible innovation. This paper outlines the case's background and provides justifications for requiring TDM licences in the training of commercial generative AI systems. It argues that licensing requirements prevent the unjust appropriation of creators' work, foster valuable collaboration between creators and AI developers, and could even create new markets for existing works. The paper addresses practical challenges of TDM licensing, such as high costs, complexity, and the opacity of generative AI models. To address these issues, it proposes a set of reforms, including the adoption of standardised contracts for TDM, cross-licensing arrangements to facilitate fair data exchanges, and "nutrition labels" on AI-generated content to increase transparency and accountability. The paper concludes that these reforms, alongside the proposed court decision in *Getty Images*, could strengthen the UK's AI and art industries by promoting innovation within a fair legal framework that strikes an appropriate balance of rights between technology developers and creators.

Keywords: Copyright; AI; Licensing; Text And Data Mining; Generative AI

Introduction

In 2023, Getty Images commenced legal proceedings in the United Kingdom High Court against Stability AI (Getty Images, 2023). Getty Images claims that 7.3 million images were unlawfully scraped from its website by Stability AI to train its Generative Artificial Intelligent System (GAIS) without an appropriate licence (Getty Images, 2023). The Copyright, Designs and Patents Act 1988 (the Act) provides Getty Images with copyright protection over its visual asset database, so unless an exception applies, permission (through a licence) is required if other parties wish to use or copy these images. Section 29(A) of the Act provides an exception which permits copies of any copyright protected material for the purpose of Text and Data Mining (TDM) without a specific licence if this is for *non-commercial purposes.*

TDM is the automated technique used to extract and analyse vast amounts of online text or data to reveal relationships and patterns in data (Holland, 2021). TDM has become an increasingly valuable tool to train lucrative and beneficial GAIS on mass amounts of data scraped from the Internet. But as profitable technology companies are using this process to train their GAIS without a licence, the Intellectual Property (IP) rights attached to training data have been under scrutiny because it is unclear whether developers need a TDM licence to train their *commercial systems* on copyright-protected materials. There has been a flood of copyright infringement cases against AI companies who have chosen not to use TDM licences to train GAIS, but most of these are against American companies in the US Courts (Lutkevich, 2024). *Getty Images v Stability AI* is the first case of its kind in the UK.

In 2022, the UK Government's Intellectual Property Office (IPO) proposed to broaden the scope of Section 29(A) to provide commercial generative AI companies, like Stability AI, with unprecedented access to train its systems on copyright-protected materials without a TDM licence (the Proposal). The Proposal was designed to align with the Government's (2021) National AI Strategy to make the UK the most attractive landscape for AI development and investment. AI developers claimed that GAIS would not exist without wide exceptions to copyright law which permit the free use of TDM on copyright-protected materials (Milmo, 2024). However, a few months after, the IPO was forced to pause its Proposal due to backlash from the creative industry who argued that their works should not be used as free training data without compensation provided by a TDM licence (House of Commons, 2023; Orlowski, 2024). It is unclear whether the Government plans to re-introduce the Proposal, but the IPO is likely awaiting the outcome of *Getty* to set the UK's future approach.

In this paper, I use the facts of *Getty Images v Stability AI* as a platform to consider how the judge should resolve this case. This paper argues that the IPO's Proposal (2024) overlooked the innovative and collaborative value of licensing in relation to the AI copyright "input dilemma". I propose that in relation to the upcoming case, the Court should decide in favour of Getty Images. This judgment would affirm the current scope of Section 29(A) so only entities using copyright-protected materials for *non-commercial purposes* will be able to do so without a TDM licence. There is a perception that requiring licences will stifle AI development and frustrate the Government's pro-innovation approach to AI regulation (Milmo, 2024). Throughout this paper, I argue that licences can encourage AI innovation, but also allow the creative industry to flourish.

The original contributions of this paper can be seen as threefold. Firstly, there is limited academic literature that clearly outlines the UK's copyright landscape in relation to TDM and GAIS. Academic commentary has focused on jurisdictions where there are more cases being decided based on this dilemma and the scale of AI development is larger – for example, the US, EU, or Japan (Dermawan, 2023; Manteghi, 2023; Li, 2024). In addition, beyond offering a descriptive account of the law, this paper also focuses on the normative, more ambitious, question: how *ought* UK copyright law apply to the training of commercial GAIS using unlicensed materials? I ground my analysis in *Getty Images* as an opportunity to consider the real implications and practicalities of these cases.

The second contribution is based on the interdisciplinary analysis that I adopt to reason how *Getty Images* should be decided. To date, lawyers, AI developers, and creatives, have been responding to this question in isolation. Thus, this paper aims to unify discourses between these communities and recommends a solution which balances the interests of the law, advancement of technology, and preservation of creatives' rights. Finally, this paper is also committed to go beyond description, analysis, and critique, by providing policy recommendations about how TDM licences can be improved to satisfy the needs of our growing technology industry and safeguard artists from copyright infringement. This reform-oriented element is seen as necessary to ensure that the benefits of the "law in books" translates into the "law in action" (Hutchinson, 2015). Focussing on the "law in action", I include real case studies and examples to point to opportunities to improve our TDM licensing landscape.

In Section I, I outline *Getty Images* and the UK legal framework that applies to TDM. I will then briefly outline how the Court should decide in favour of Getty Images. Sections II and III will focus on the benefits and challenges of requiring AI developers to seek a licence to train their commercial GAIS on copyright-protected materials. Section II will explore three justifications for maintaining the scope of Section 29(A). For the first justification, I argue that a TDM licence is required so GAIS do not unfairly freeride off creator's content. The second justification argues that mandating TDM licensing will encourage creators and AI developers to unlock untapped value in materials and prevent obstacles that stifle innovation. For my final justification, I dispute claims that GAIS will erode the market for original works that serve as training data for GAIS – I suggest that TDM licensing could spur a new demand for existing works.

In Section III, I acknowledge that despite these justifications, issues with TDM remain, namely concerning the: (a) cost, (b) complexity, and (c) opaqueness of GAIS. Last year, the IPO announced that it was establishing a Code of Practice (COP) to improve the TDM licensing environment (IPO, 2023; Foerg, 2023). Just a few months after this announcement, the COP was abandoned as members of the committee could not agree on policies that balanced the rights of the creative industry and AI developers (Thomas and Criddle, 2024). In the final section of this paper, I respond to the challenges set out in Section III and provide some measures that could mitigate the shortfalls of TDM licences that should be implemented by the IPO. As this paper is mainly dedicated to the UK's *legal* response, the suggestions for the IPO are brief and provided as a platform for further research to supplement the proposed Court decision.

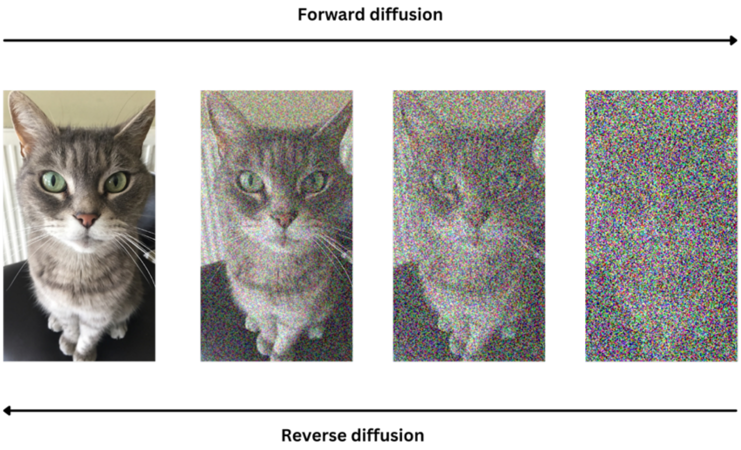
1. How Should the UK High Court Decide *Getty*?

In this Section, I outline the technical elements of *Getty* and map out the current UK copyright law in relation to TDM under Section 29(A) of the Act. I then argue that Stability AI infringed copyright when the company used Getty Images' protected materials to train its GAIS (called Stable Diffusion) without a TDM license.

1.1. Getty Images v Stability AI

To assess whether Stability AI violated Getty Images' copyright, it is important to understand how Stable Diffusion, an AI tool that turns text into images, is trained. This process involves utilising images from various online databases, including Getty Images, but these images are not stored directly. Instead, the AI developers utilise a specific training method, like a "diffusion model", to enable the model to learn patterns from the images.

The "diffusion model" training process works by adding random visual "noise" to each of the image present in the training dataset until the image is not recognisable – this process is understood as "forward diffusion" (Guadamuz, 2024). Once the images are "noised", the AI is trained to recognise and gradually remove that noise to reconstruct the original image in a process known as "reverse diffusion" (Guadamuz, 2024). Figure 1 illustrates the "diffusion model" training process using an image of a cat as an example. Through repeated training on thousands of images, the AI model learns to identify patterns, like what common objects and colours look like. As a result, the GAIS can start to generate new images based on these learned patterns.



Overlay image credit: Adrien Limousin / Better Images of AI / Non-image / CC-BY 4.0

Figure 1: *Illustrates the noising process during the diffusion model training process for an image of a cat*

Importantly, the images generated by the AI model in its output will not be exact copies of any original images used in the training process. Instead, the outputs are statistical approximations learned during the training process which inform the model's overall understanding of how objects are represented (Guadamuz, 2024). Getty Images' extensive library of over 12 million images served as a rich resource for training data for GAIS, contributing to Stable Diffusion's enhanced ability to generate vast, realistic outputs.

Copyright law becomes relevant in this training process when we focus on what this framework aims to protect. Copyright law determines that the protected element of works subsides in the creative expression – like the lighting, exposure, filter, or positioning of an image (*Temple Island Collections,* 2012). These are the parts of images that copyright protects because they require a creator's own thoughts and originality. However, what is significant about the GAIS training process for copyright law is that Stability AI does not use TDM to copy Getty Images' database for the protected elements of its materials (Lemley and Casey, 2020).

To train GAIS, it is often the *factual* elements of the work extracted through TDM which are the most valuable as opposed to the creative aspects (Lemley and Casey, 2020). The diffusion model training process relies on broad visual features of images, rather than specific artistic choices. For example, when training Stable Diffusion, TDM was not used to extract data about lighting techniques which were employed to make an image of a cat particularly appealing. Instead, the accessibility to a large collection of images which detailed the features that resemble a cat (fur, whiskers, big eyes, paws) were what Getty Images' database provided. The challenge for Stability AI is that it is unable to capture these unprotectable parts of the images that are essential for training Stable Diffusion, without making a copy of the protectable parts (Lemley and Casey, 2020).

1.2. The current UK copyright law framework in relation to TDM

In Section 29(A) of the Act, the UK currently permits TDM of copyrighted works for *non-commercial* purposes provided that the entity has lawful access to the work. Lawful access means that individuals do not require separate permission for TDM, they just require access to the works through a general licence or subscription (IPO, 2014). Section 29(A) is also mandatory, so even if contract terms to access materials might preclude TDM, these are unenforceable (IPO, 2014). Given that academics and researchers often have broad institutional access to materials, the UK Government has exercised a very facilitative approach to TDM for training non-commercial GAIS to drive scientific advancements (Flynn and Vyas, 2023).

The question of whether the training of Stability AI classifies as a *non-commercial* purpose is likely to be an unproblematic for the Court because: (i) it has already been decided that it is a commercial entity in the US case (*Getty*, 2023), (ii) Stable Diffusion was monetised (*ibid.*), and (iii) the Government intended for Section 29(A) to be used by universities and charities (IPO, 2014). I acknowledge that UK data laundering practices (where commercial technology companies outsource data collection and model training to academics) present a loophole in this framework that must be addressed (Baio, 2022), but consideration of this is beyond the scope of this paper given that this did not occur in *Getty.* Therefore, Stable Diffusion will likely fall outside the non-commercial exception in Section 29(A). It will be for the Court in *Getty* to decide whether to extend Section 29(A) to *commercial* use (as in the Proposal) to free Stability AI from copyright infringement.

1.3. How should the UK's High Court decide Getty?

I argue that the Court should decide in favour of Getty Images and refrain from expanding Section 29(A) to commercial GAIS. Therefore, Stability AI infringed copyright when it did not acquire a TDM licence to train its system on Getty Images' protected materials.

It is important to note here that since Getty Images' legal action in 2023, Stability AI has later filed a defence against its copyright infringement (Cooke, 2024). Stability AI is arguing that it cannot be held liable for copyright infringement in the UK because the training of its GAIS took place on servers in the US (*ibid*.). Stability AI originally tried to have the case struck out based on this jurisdictional fact. However, the judge overseeing the litigation decided that the case should go to trial so more evidence could be gathered about this matter (Davies and Dennis, 2024).

Thus, there remains a strong possibility that Stability AI's defence will not be upheld in court and the judge will have to determine how the scope of Section 29(A) applies to the case (*ibid.*). It is also possible that the judge will address this question of law in the case regardless of the jurisdiction in which the training took place. In a recent AI and patent case (*Emotional Perception*, 2023), the judge went beyond resolving the matters between the parties to answer wider questions of law relating to the patenting of artificial neural networks (*ibid.*). It is assumed that this is because of the long backlog of cases and the rapidly evolving development of AI which requires faster responses and legal certainty to protect creators and AI developers. Therefore, a decisive ruling in *Getty Images v Stability AI* is welcomed to provide much-needed legal guidance to the industry and to align UK copyright law with the rapid development of commercial GAIS.

2. Justifications For Requiring TDM Licences To Train Commercial GAIS

In this Section, I provide three justifications for my proposal aimed at balancing the IP rights of original creators with the goal of fostering AI innovation according to the Government's Strategy (2021). Firstly, I suggest that requiring TDM licences means that generative AI developers cannot freeride off creator's works. The freeriding argument claims that creatives will lack sufficient incentives to develop new works if their materials are leveraged by others without fair compensation (Lemley, 2005). Despite claims that the freeriding argument does not apply to GAIS, its relevance persists when considering how the value of existing works can be reimagined when used as training data. Secondly, TDM licences can enable a more collaborative innovation process, supporting AI developers in creating advanced GAIS more efficiently. Finally, contrary to the perception that GAIS will diminish market value for original works, I propose that mandating TDM licences might occasionally reinvigorate demand for creators' original works.

2.1. Freeriding and reimagined value

Protectionist IP theorists argue that copyright law should uphold a robust exclusionary right to prevent unauthorised use of protected works (Lemley, 2005). Getty Images (and its photographers) invest substantial resources in curating a high-quality image repository, with over $200 million invested between 2017 and 2020 alone (Getty Images, 2023). Photographers depend on the royalties received from Getty Images to sustain their livelihoods and continue producing content (Getty Images, 2023). Copyright protection thus enables Getty Images to maintain profitability by determining its competitors from using its images without bearing the associated costs of production. Without fair remuneration, Getty Images and its contributors would not have the resources, incentives, or time to invest in its database. To date, Stability AI has raised more than $100 million in financing (Getty Images, 2023). But without scraping images from Getty Images' database, Stability AI might not have had access to the extensive data needed to train its model effectively. The success of Stable Diffusion rests on the time and investment of Getty Images (and its photographers) into its database. Stability AI's reluctance to seek a licence amounts to freeriding on Getty Images' materials. Therefore, mandatory TDM licences will ensure that commercial GAIS cannot benefit from protected works without compensation to the creator to recognise how these materials are the foundation of GAIS.

The freeriding argument has been criticised in relation to the training of GAIS (Lemley and Casey, 2020). This is because, as explored in Section I, TDM does not extract the protected elements of copyright materials. According to this argument, Stable Diffusion does not freeride on Getty Images' photograph of a cat. The factual elements that compose a cat are not connected to a photographer's time and investment into the image – this is only directed at the expression of the cat (captured in the angle of a shot, exposure, or colour manipulation) which Stable Diffusion did not capitalise on (Lemley and Casey, 2020). However, the potential for TDM to "re-imagine" the value of such materials suggests that the freeriding argument may still apply.

An example of re-imaged value is demonstrated by Gmail's predictive email response algorithm which was trained on romance novels (Smith, 2016). Google leveraged the fact that these romance novels would provide convenient training data for its algorithm to learn varied language, phrasing, and grammar structures. The algorithm was not used to replicate specific story elements like the characters, settings, or descriptive tone. Instead, its sole use was for the purpose of understanding the English language (Smith, 2016). Nevertheless, these romance novels were still valuable (albeit in a reimagined way) to the success and effectiveness of Gmail's tool.

Similarly, Stability AI's use of Getty Images' database illustrates how re-imaged uses can result from TDM practices. It would have been significantly more difficult for Stability AI to train its GAIS without the convenience, existence, and volume of data extracted from Getty Images' vast database. Thus, even though this is not connected to the traditionally protected elements of images under copyright law per se, the underlying freeriding motive still stands. The use of TDM to train GAIS still freerides on the creator's materials by extracting valuable data from existing materials which would not exist without creators' significant time, resources, and efforts.

I do not suggest that the boundaries of copyright law should be extended to protect all materials that serve as the basis of profitable innovation. Copyright law maintains appropriate exceptions to protection for scientific formulas or symbols to ensure the necessary access to the basis of our scientific and creative developments. However, I do argue that copyright law should reassess what was traditionally deemed unprotectable in light of GAIS to ensure that the law still supports the appropriate balance of rights. In this context, TDM licences could ensure that AI companies appropriately compensate creators for their works which provide the foundations of profitable GAIS.

2.2. Collaboration to unlock untapped value

A central feature of the IP system is the licensing framework, which enables lawful access to copyright protected materials to progress innovation. Therefore, the fact that TDM can extract untapped value in existing materials to develop new and innovative AI systems is exactly what copyright law supports (Leval, 1990). Examples of untapped value include user interactions on social media being used to train virtual assistants (Meta, 2023) and international legislation texts used to train deep learning translation tools (DeepL, 2023). These uses illustrate how existing, protected content can contribute significantly to the development of further innovation, like GAIS. Copyright law stands to incentivise individuals to develop upon existing protected works using licences to unlock further creations which are socially beneficial.

Thus, the second reason that copyright law should encourage TDM is because it saves AI developers time and resources from training systems when resourceful data already exists. AI innovation efforts can then be directed at developing cutting-edge GAIS, as opposed to data creation and training. A legal framework that offers clarity on IP rights related to training data could encourage creators and AI developers to explore usually beneficial uses of existing content (Brook and Murray-Rust, 2014). TDM licences would allow creators to profit from such uses, while fostering a collaborative environment that strengthens the development of GAIS.

Stability AI had already started to re-imagine the use of existing materials by leveraging Getty Images' database which was originally designed for use by media and corporate companies. However, since Stability AI did not obtain a TDM licence, the materials scrapped from Getty Images' website were low-quality and distorted by watermarks (Getty Images, 2023). A formal licensing agreement would have enabled access to high-quality data, and might have also encouraged collaborative enhancements, including machine-readable metadata which would have streamlined and enhanced the training process. Getty Images have already worked with AI companies, so licensing negotiations could have also offered opportunities for Getty Images to further improve Stable Diffusion's development process with its valuable domain knowledge and experience (Getty Images, 2023). Thus, TDM licences facilitate collaboration between AI developers and creators which is necessary to better optimise training data to efficiently develop better GAIS.

Without adequate compensation measures provided by TDM licences, creators are stifling the innovation process (Shan *et al.*, 2023). Using data tags on their materials (like robots.txt which contain do-not-scrape directives to block web crawlers), creators are blocking and distorting the TDM processes to retain control over their works. Data tags, like Nightshade, can even "poison" the TDM process by sending back the incorrect images to distort the accuracy of GAIS's training process (Shan *et al.*, 2023). The use of data tags has been an act of resistance from creatives against AI companies freeriding on their materials. But this is not because creators are reluctant to have their works being used as training data *per se*; creators just want to control the use of their works and ensure that they are adequately compensated (Dean, 2023).

Data tags and other resistance efforts create a divergence between AI companies and creators, preventing any possibility of their works being used for remuneration and corrupting the training process for GAIS. Furthermore, as materials are being increasingly withheld from AI companies, this will ultimately lead to the self-demise of GAIS. New data is needed for GAIS to meet the evolving demands of consumers. Therefore, TDM licences offer a way to resolve the tensions between these two communities and support a more productive innovation process for GAIS whilst adequately compensating artists.

2.3. Re-invigorating value in original works

It is argued that the use of creator's materials as training data for GAIS will devalue the market for the original work (Sobel, 2017; Lucchi, 2023). An alternative hypothesis is that TDM could also hold the potential to occasionally *improve* the market for original works despite their inclusion in datasets for training GAIS. To explore this argument in a different context, Snapchat has made licensing agreements with minority artists to prompt users to use their music in videos. Snapchat has benefited from a cheaper method to obtain music on its platform and smaller musicians have benefited from increased exposure of their works on the popular app (Malik, 2022). While the original intention for these artists was not to produce works for this purpose, it provides an alternative avenue to attract audiences and generate additional market access.

In a similar way, using existing materials to train GAIS could actually prompt renewed appreciation for these works. Benn argues that AI art might increase the public's appreciation for human creativity, as human-centred works can carry emotional or aesthetic value that digital creations may not fully replicate (Aesthetics for Birds, 2022). Therefore, if a photographer or artist exclusively licences their unique database of images which are distinctive with respect to the style or skills needed to replicate the images, the licensing AI company will benefit from a significant competitive advantage. Greg Rutowski is a Polish digital artist who uses classical painting styles to create fantasy landscapes which are used in illustrations for games like Dungeons & Dragons. His images have become more popular since his images were used as training datasets for text-to-image AI generators. Rutkowski was optimistic that this could be a good way to reach new audiences who appreciate and value his fantastical and ethereal artistic style. However, the problem is that the GAIS did not disclose or acknowledge the artists or sources for which the training materials were derived from so it was impossible for users to find Rutowski's artworks. Therefore, it is acknowledged that the strength of the "reinvigoration" argument relies on GAIS being transparent about their training materials, but also only where datasets hold certain unique value. But it is maintained that in these instances, TDM licences could drive revenue and appreciation towards the original materials.

3. Problems with TDM Licensing and Mitigating Measures for the IPO

In this Section, I outline three drawbacks with TDM licensing: cost, complexity, and opacity. While these problems raise valid concerns, I detail mitigating measures which could be implemented by the IPO to improve TDM licensing through industry changes.

3.1. Cost: cross-licensing arrangements

The main problem with TDM licences is that they are very costly for AI developers. GAIS require vast amounts of data to produce good quality outputs – just training the first two versions of Stable Diffusion required around 12 million images (Getty Images, 2023). Collating smaller datasets from individual owners is usually a time-consuming and expensive task (Lemley and Casey, 2020). Alternatively, the possibility of acquiring large datasets from bigger companies is unlikely as these have significant commercial value so are priced highly or not licensed at all. The BBC has admitted that it relies on its own proprietary data as licensing third-party materials for their AI tools is too expensive (BBC, 2022). The BBC is in a fortunate position to at least have its own data, but for smaller companies the cost of TDM licensing creates barriers to enter the AI market. The cost of TDM licences creates monopolies in

AI development as only a few companies can afford to licence third-party datasets or have access to their own data to train GAIS (Lucchi, 2023).

While TDM licensing might be expensive, adapting existing cross-licensing mechanisms to copyright-protected data could be a useful mechanism to help smaller companies develop and train their own GAIS (Fershtman and Kamien, 1992). A cross-licensing agreement occurs where parties exchange licences (instead of money) for use of each other's IP. In this context, I suggest that companies with access to large (often homogenous) datasets could exchange their materials with smaller companies who may have more diverse datasets. Gaining richer data is important for companies to avoid their models' "overfitting" (creating outputs which replicate the training data) which could result in costly copyright claims in the output materials of GAIS (Carlini *et al.*, 2023). AI developers are also under increased pressure to limit the bias outputs of their GAIS – especially as new tools are being released to scrutinise unrepresentative models (Heikkilä, 2023).

An example of a cross-licensing opportunity could involve "Better Images of AI" licensing data about the accurate representation of AI in exchange for larger datasets from the BBC, giving each other the resources to generate valuable and representative GAIS. It is possible for terms in the cross-licensing agreement to stipulate that each party does not use the data for the same purpose, so they do not develop identical GAIS or saturate the market. I suggest that the IPO should raise awareness of TDM cross-licensing arrangements to reduce the monetary barriers required to enter the generative AI market and facilitate the creation of more diverse and cutting-edge AI tools.

3.2. Complexity: standardised licences

TDM licensing is also a time-consuming process if complex contracts are drafted which require legal assistance if parties want to have an informed understanding of the scope of data use for TDM (BBC, 2022; Vollmer, 2016). Big corporations can often leverage their powerful position to draft licences in an overly complex way to attain broad rights over creators' materials (Stevens, 2023; Sobel, 2017). To mitigate this problem, I suggest that the IPO creates standardised TDM (and even cross-licensing) contracts to be used between AI companies and creators. Standardised TDM licences will empower creators to licence their materials without the need to navigate the complex legal landscape to control the use of their data. Comprehensible licensing contracts will also streamline the innovation process for AI developers who can train GAIS faster without the need to spend time drafting contracts and negotiating TDM terms (Maffioli, 2023).

A global movement towards open-source standardised contracts for routine arrangements has already begun with Non-Disclosure Agreements (oneNDA, *n.d.*). While it is outside the scope of this paper to detail what standardised TDM contracts should include, Maffioli (oneNDA, *n.d.*) has proposed a standardised template that could be a good baseline for the IPO to develop. This proposed TDM contract includes terms relating to usage and access rights, risk allocations and liabilities, transparency provisions and compensation (oneNDA, *n.d.*). Therefore, the complexity of TDM licensing could be mitigated if the IPO designs standardised TDM contracts for use between AI companies and creators.

3.3. Opacity: nutrition labels

Due to the vast amounts of data that GAIS are trained on, and the number of parameters within models, GAIS often produce content that does not resemble its training data which makes it difficult for creators to know if their materials have been unlawfully used as training data (Guadamuz, 2024). It is also in the best interests of the company to ensure that there is no resemblance with creator's works to avoid copyright claims targeted at the output imagery (Guadamuz, 2024). The images used to train Stable Diffusion were watermarked, so Getty Images could identify its images in the output imagery. However, images will not always be watermarked, so creators will be unaware of the use of their works as training data for GAIS. This creates a loophole for AI developers who can avoid obtaining TDM licences (even if legally required) because the opacity of GAIS provides a shield against accountability for the infringement of protected materials. Thus, TDM licensing is only effective if AI developers are forthcoming about their use of protected materials, or creators are made aware of the use of their works as training data for GAIS.

To increase transparency, I propose that the IPO implements a requirement for AI developers to embed "nutrition labels" on content created by GAIS (Lucchi, 2023; Maffioli, 2023). Nutrition labels are already being used by leading AI companies to disclose information about what data was used to create AI-generated images (Swant, 2023). By integrating nutrition labels onto output imagery, transparency is instilled in GAIS development, empowering creators to better recognise potential copyright infringements by GAIS and encouraging AI developers to scrutinise the origins of their training materials (Maffioli, 2023).

I do acknowledge that there are limits to transparency, as AI companies should not be expected to publicly disclose their training datasets or open-source their models – such would undermine a company's competitive advantage. However, in light of the opaqueness of models, creators should be afforded with greater awareness of whether their works are being unlawfully used as training data. Additionally, the requirement for nutrition labels aligns with the argument made in Section II which recommends that transparency could increase demand for creators' original works. From a commercial standpoint, GAIS with accredited sources are also perceived as more reliable and responsible by users (Swant, 2023). Thus, the IPO should mandate developers to embed nutritional labels on AI content to strike an appropriate balance between AI developers and creators, while promoting the advancement of improved GAIS.

Conclusion

This paper has argued that without a TDM licence, training commercial GAIS on copyrighted materials should be considered as infringement. In *Getty,* the Court should refrain from expanding the scope of Section 29(A) so only entities using copyright-protected materials for *non-commercial* purposes can do so without a TDM licence. Three reasons have been presented to highlight how the Proposal overlooked the benefits of TDM licensing for the AI and creative communities.

A majority of literature attempting to resolve dilemmas in the intersection of copyright law and AI do not focus on the UK jurisdiction and are not interdisciplinary in their analysis. In this paper, I attempted to address this research gap by focusing on the upcoming *Getty* decision as well as exploring reasons for deciding the case which balances the interests of the law, AI developers, and creatives. While I have focused on the UK, the justifications, challenges and recommendations outlined in Sections II and III can be adapted to other jurisdictions – especially where the courts have already ruled that TDM licences are necessary. The paper's more novel and optimistic perception of the value of TDM licensing will be helpful to bridge innovation efforts between the AI industry and creators. I hope that the hypothetical examples and real examples included in this paper shed light on how these communities can work together in a responsible and mutually beneficial way. Within the art industry, original creators, AI start-ups, minority artists, and large AI companies can all bring something to the innovation ecosystem if they want to. I challenge these actors to take collaboration opportunities more seriously and think about how they can use the law to facilitate this process to ensure their respective needs, commitments, and rights are upheld.

The solution to the copyright problems in relation to training commercial GAIS is complex. In this paper, I have been a strong advocate for the use of TDM licences as their innovation effects have often been overlooked. The proposals outlined in the final section of this paper serve as a purpose to show that while TDM licences can resolve some problems relating to freeriding and creator resistance, they are not perfect and require shaping to meet the demands of the working industry. While I have pointed to some of the shortfalls and mitigating measures, including standardising licensing, prompting cross-licensing opportunities, and utilising nutrition labels, further research is required. I suggest that further research adopts a more empirical methodology to investigate the real challenges relating to "licensing in action" faced by AI developers and creators. In this paper, I used the Government's consultation on IP and AI which yielded responses from various actors with different interests, like the BBC, IBM, the Music Publishers Association, The Law Society, Siemens, and the Wellcome Trust to name a few (Intellectual Property Office, 2022). However, given that these all groups submitted to the consultation, the responses might not represent wider views in the ecosystem from underrepresented artists and smaller AI developers who might also be facing issues that have not been reported or raised. I hope that the recommendations provided in this paper can set out the first steps for other researchers to advocate for changes to our licensing and innovation frameworks to protect creators and improve clarity over the scope of rights in the face of GAIS.

Finally, in relation to the wider intersection between copyright and generative AI, this paper has exclusively focussed on the "input question". But questions remain to be answered in relation to whether the outputs of GAIS can infringe on creators' copyright. The TDM licensing approach suggested in this paper is one way to facilitate a better dynamic between the AI and creative industry which could limit the legal action necessary to monitor the output imagery by resolving issues in initial licensing negotiations. For instance, TDM licences could allow AI developers and creators to negotiate in advance to compensate artists if the outputs of GAIS are to the likeness or similarity of the artist's original work. Future research could understand how TDM licences, if at all, could benefit legal questions focussed on infringement of the output imagery. This would provide a more rounded and comprehensive understanding of the innovative value of TDM licences in relation to GAIS

References

Aesthetics for Birds, (2022). Eights Scholars on Art and Artificial Intelligence. *Aesthetics for Birds,* 2 November. Available at: https://aestheticsforbirds.com/2023/11/02/eight scholars-on-art-and-artificial-intelligence/. (Accessed: 30th November 2023).

Andersen v. Stability AI Ltd (2023). U.S. District Court for the Northern District of California, No. 3:23-cv-00201.

BBC (2022). *BBC Response to the UKIPO Consultation on AI and IP: Copyright and Patents.* Available at: https://www.gov.uk/government/consultations/artificial-intelligence and-ip-copyright-and-patents (Accessed: 30th November 2023).

Baio, A. (2022). AI Data Laundering: How Academic and Non-profit Researchers Shield Tech Companies from Accountability. 30 September. Available at: https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield tech-companies-from-accountability/. (Accessed: 14 January 2024).

Brook, M., Murray-Rust, P. & Oppenheim, C. (2014). The Social, Political and Legal Aspects of Text and Data Mining (TDM). *D-Lib Magazine* (November). Available at: https://openaccess.city.ac.uk/id/eprint/4784/1/D-Lib%20Magazine.pdf.

Carlini, N *et al.* (2023). Extracting Training Data from Diffusion Models. Available at: https://arxiv.org/abs/2301.13188.

Cooke, C. (2024). Stability files defence in important test case on AI and UK copyright law. *Complete Music Update.* Available at: https://completemusicupdate.com/stability-files defence-in-important-test-case-on-ai-and-uk-copyright-law/. (Accessed 4 March 2024)

Copyright, Designs and Patents Act 1988, Section 29(A). Available at: https://www.legislation.gov.uk/ukpga/1988/48/section/29A (Accessed: 30th November 2023).

Davies, C. Dennis, G. (2024). Getty Images v Stability AI: the implications for UK copyright law and licensing. Pinsent Masons, 29 April. Available at: https://www.pinsentmasons.com/out-law/analysis/getty-images-v-stability-ai-implications copyright-law-licensing. (Accessed 19 June 2024)

DeepL (2023). Why AI translation is a must-have for legal firms with global colleagues and clients. DeepL, 11 April. Available at*: https://www.deepl.com/en/blog/why-ai-translation-is a-must-have-for-legal-firms-with-global-colleagues-and-clients.* (Accessed: 11 April 2023).

Dermawan, A. (2023). Text and data mining exceptions in the development of generative AI models: What the EU member states could learn from the Japanese "nonenjoyment" purposes. *The Journal of World Intellectual Property* 27(1).

Emotional Perception AI Ltd v Comptroller-General of Patents, Designs and Trade Marks [2023] EWHC 2948 (Ch).

Fershtman, C Morton, I. (1992). Cross licensing of complementary technologies. International Journal of Industrial Organisation, 10(3).

Flynn, S. Vyas, L. (2023). Examples of Text and Data Mining Research Using Copyrighted Materials. *Kluwer Copyright Blog*, 6 March*.* Available at: https://copyrightblog.kluweriplaw.com/2023/03/06/examples-of-text-and-data-mining research-using-copyrighted-materials/. (Accessed: 20th January 2024).

Foerg, M. (2023). The UK government steps towards a code of practice on copyright and AI. *Kluewer Copyright Blog,* 27 September. Available at: https://copyrightblog.kluweriplaw.com/2023/09/27/the-uk-governments-steps-towards-a code-of-practice-on-copyright-and-ai/. (Accessed 19 June 2024).

Getty Images v Stability AI. (2023). United State District Court of Delaware. No. 1:23-cv 00135-UNA.

Getty Images. (2023). *Getty Images Statement.* Available at: https://newsroom.gettyimages.com/en/getty-images/getty-images-statement (Accessed: 30th November 2023).

Guadamuz, A (2024). A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs. *GRUR International,* 140(2).

Heikkilä, M. (2023). These new tools let you see for yourself how biased AI image models are. *Technology Review,* 22 March. Available at: https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for yourself-how-biased-ai-image-models-are/. (Accessed: 30th November 2023).

Holland, C. (2021). Copyright and Text & Data mining – what do I need to know?. *Open@UCL Blog,* 6 July. Available at: https://blogs.ucl.ac.uk/open access/2021/07/06/copyright-text-data-mining/ (Accessed: 30th November 2023).

House of Commons (2023). *Artificial Intelligence: Intellectual Property Rights.* Available at: https://hansard.parliament.uk/commons/2023-02-01/debates/7CD1D4F9-7805-4CF0-9698-E28ECEFB7177/ArtificialIntelligenceIntellectualPropertyRights (Accessed: 30th November 2023).

Hutchinson, T. (2015). The Doctrinal Method: Incorporating Interdisciplinary Methods in Reforming the Law. *Erasmus Law Review* 3.

Intellectual Property Office (2014). *Exceptions to Copyright.* Available: https://www.gov.uk/guidance/exceptions-to-copyright. (Accessed: 30th November 2023).

Intellectual Property Office (2022). *Artificial Intelligence and Intellectual Property: copyright and patents: Government response to consultation.* Available: https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents government-response-to-consultation. (Accessed: 30th November 2023).

Intellectual Property Office. (2023). *The government's code of practice on copyright and AI.* Available at: https://www.gov.uk/guidance/the-governments-code-of-practice-on-copyright and-ai (Accessed: 30th November 2023).

Leffer, L. (2023). Your Personal Information Is Probably Being Used to Train Generative AI Models. *Scientific American*, 19 October. Available at: https://www.scientificamerican.com/article/your-personal-information-is-probably-being used-to-train-generative-ai-models/. (Accessed 30th November 2023).

Lemley, M. (2005). Property, Intellectual Property, and Free Riding. *Texas Law Review* 83(291).

Lemley, M. Casey, B. (2020). Fair Learning. *Texas Law Review* 99(4).

Level, P. (1990). Towards a Fair Use Standard. *Harvard Law Review* 103(5).

Li, J. (2024). Managing Copyright Infringement Risks in Generative Artificial Intelligence Data Mining' *4th International Conference on Management Science and Industrial Economy Development* 39.

Lutkevich, B (2024). AI lawsuits explained: Who's getting sued?. *Tech Target,* 2 January. Available at: https://www.techtarget.com/whatis/feature/AI-lawsuits-explained-Whos-getting sued (Accessed: 13th January 2024).

Maffioli, D. (2023). Copyright in Generative AI training: Balancing Fair Use through Standardization and Transparency. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4579322.

Malik, A. (2022). Snap's new creator fund will award independent musicians up to $100,000 per month. *Tech Crunch,* 28 July. Available at: https://techcrunch.com/2022/07/28/snaps new-fund-award-independent-musicians-100000per month/guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAKgtyuQ5DB9lnJXeUMSLXttk4EGx_DRWjEGmlTNLk33nb7i8YYBi4lqX_Qg9_kE_naZi TObBUj4jfJwHnrEDC7eYADj3w96HAIETgOZ WjtlOs4Y2jsPnCciVoDD0reGgm gBsyroNS3trQvYsRNsn34FXLzOsE5-q2I9TvIEIYa. (Accessed: 14 January 2024).

Manteghi, M. (2024). Can text and data mining exceptions and synthetic data training mitigate copyright-related concerns in generative AI?. *Law, Innovation, and technology* 1.

Meta, (2023). Privacy Matters: Meta's Generative AI Features. *Meta Newsroom,* 27 September. Available at: https://about.fb.com/news/2023/09/privacy-matters-metas generative-ai-features/. (Accessed: 30th November 2023).

Milmo, D. (2024) ''Impossible to create AI tools like ChatGPT without copyrighted material, OpenAI says', *The Guardian,* 8 January. Available at: https://www.theguardian.com/technology/2024/jan/08/ai-tools-chatgpt-copyrighted-material openai (Accessed: 14 January 2024).

New York Times v Open AI (2024). United States District of Southern New York. No. 1:23-cv 11195.

OneNDa. (n.d.). *oneNDA started off as a bit of a whim which went global – overnight.* Available at: https://www.onenda.org/about (Accessed 30th November 2023).

Orlowski, A. (2024). How 'big tech' barons are plotting to steal Britain's creativity. *The Telegraph,* 28 October. Available at: https://www.telegraph.co.uk/business/2024/10/28/how-big-tech-barons-plotting-steal-british-creativity/ (Accessed 30 October 2024).

Rouse, M. (2024). Generative AI. *Techopedia,*15 January. Available at: https://www.techopedia.com/definition/34633/generative ai#:~:txt=Generative%20AI%20(genAI)%20is%20a,with%20the%20sam%20statistical%2 0properties (Accessed: 20th January 2024).

Shan, S *et al.* (2023). Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. *Cryptography and Security.* Available at: https://arxiv.org/abs/2310.13828.

Smith, L. (2016). Google's AI engine is reading 2,865 romance novels to be more conversational. *The Verge,* 5 May. Available *at:* https://www.theverge.com/2016/5/5/11599068/google-ai-engine-bot-romance-novels. (Accessed: 30th November 2023).

Snow, J. (2018). "We're in a diversity crisis": cofounder of Black in AI on what's poisoning algorithms in our lives. *Technology Review,* 14 February. Available at: https://www.technologyreview.com/2018/02/14/145462/were-in-a-diversity-crisis-black-in ais-founder-on-whats-poisoning-the-algorithms-in-our/ . (Accessed: 30th November 2023).

Sobel, B. (2017). Artificial Intelligence's Fair Use Crisis. *Columbia Journal of Law and the Arts* 41(45).

Stevens, M. (2023). If Big Brands Copied Their Work, What Are Artists to Do? *New York Times,* 2023. Available at: https://www.nytimes.com/2023/03/01/arts/design/digital-art copyright-marvel-panini-wizards.html. (Accessed: 30th November 2023).

Swant, M. (2023). Adobe debuts new icon as a 'nutrition label' for generative AI content'. *DigiDay,* 11 October. Available at: https://digiday.com/media-buying/adobe-debuts-new icon-as-a-nutrition-label-for-generative-ai-content/. (Accessed: 14th January 2024).

Temple Island Collections Ltd v New English Teas Ltd. (2012). EWPCC 1.

Thomas, D. Criddle, C. (2024). UK shelves proposed AI copyright code in blow to creative industries. *Financial Times,* 5 February. Available at: https://www.ft.com/content/a10866ec 130d-40a3-b62a-978f1202129e. (Accessed: 19 June 2024).

Vollmer, T. (2015). More licences are not the solution for text and data mining. *Communia,* 11 June. Available at: https://communia-association.org/2015/06/11/more-licenses-are-not the-solution-for-text-and-data-mining/. (Accessed: 30th November 2023).

How Metaphors Influence Ontology, Epistemology, and Methods in AI: Rethinking the Black Box

Dr. Angy Watson
Hughes Hall, University of Cambridge

In this response paper, I explore how metaphors influence ontology, epistemology, and methodology within AI. Using the example of the black box metaphor, I demonstrate that an over-reliance on one metaphor forecloses potential futures, limiting discourse, research and policy. I thus conclude that reflexivity about our use of metaphors is necessary and that we should strive to utilise a range of metaphors to capture the full scope of concepts we aim to express. To establish the foundation for my thesis I examine and critique two articles: "The Ethnographer and the Algorithm: Beyond the Black Box" (Christin, 2020) and "Prediction Promises: Towards a Metaphorology of Artificial Intelligence" (Möck, 2022).

Keywords: Metaphorology, Blackbox AI, Ethnography

Introduction

Metaphors are essential in how we create meaning in the world. They help us bridge the gap between complex concepts and our understanding, allowing us to work with these ideas and create new knowledge (Lakoff & Johnson, 2008; Möck, 2022). “Artificial Intelligence is a metaphor, and AI as a technoscientific discipline in between science and engineering, is a highly metaphorically loaded field of scientific inquiry” (Möck, 2022). In this paper, I explore how metaphors influence ontology (the nature of reality), epistemology (the nature of knowledge), and methodology (how one obtains knowledge) within AI (Killam, 2013; Rawnsley, 1998). I use the example of the black box metaphor to demonstrate that an over-reliance on one metaphor forecloses potential futures, limiting discourse, research and policy. Consequently, it is important to be more reflexive about our use of metaphors and strive to utilise a range of metaphors to capture the full scope of concepts we aim to express. I examine and critique two articles to establish the foundation for my thesis and then discuss my argument.

In section one of the paper, I examine the first article, "The Ethnographer and the Algorithm: Beyond the Black Box" (Christin, 2020). This article was selected for its relevance to the research question and the author's focus on the black box metaphor. Thus, in this section, I provide an overview of Christin's argument regarding the problematic opacity of algorithms, followed by her three strategies for conducting ethnography on algorithms. This section concludes with a short critique of Christin’s article and possible counter-arguments.

The second article, "Prediction Promises: Towards a Metaphorology of Artificial Intelligence" (Möck, 2022), is discussed in section two. This article provides essential counter-points to Christin's article and is a necessary scaffold for my thesis, outlined in the subsequent section. In order to ensure that my argument and discussion are narrowly focused, the review of Möck's article only includes aspects that relate to my research question. These aspects include how metaphors shape knowledge and the problem of the black box metaphor. I conclude this discussion by considering the strengths and potential missed opportunities in Möck’s article.

In section three, I explore how metaphors influence ontology and epistemology and thus inform methodology in AI. To illustrate my argument, I use Christin’s article as an example to suggest that her focus on the black box metaphor may prevent her from going "beyond the black box".

In the fourth and final section, I consider options for how we might address the problem of the black box metaphor. While I consider two documented alternative metaphors and suggest one of my own, this paper’s thesis indicates that no one metaphor should be relied upon but that instead, a more reflexive process of enquiry and a range of metaphors may serve us better as we seek to broaden our metaphorical landscape and thus future possible outcomes.

1. "The Ethnographer and the Algorithm: Beyond the Black Box"

Dr. Angèle Christin's article responds to Seaver's (2017) call for concrete "tactics" to study algorithms ethnographically (Christin, 2020). Christin uses the black box metaphor as a heuristic for algorithmic opacity, and as such, this metaphorical imagery pervades the article and framing of her ethnographic strategies. In this discussion, I outline Christin's arguments regarding the problem of algorithmic opacity and her proposed three ethnographic strategies for studying algorithms. Furthermore, I relate Christin's use of the black box metaphor and associated light imagery where relevant, as it is central to her article and my thesis.[1]

1.1. Algorithms are opaque...opacity is problematic

Christin asserts that "algorithms are profoundly opaque and function as inscrutable "black boxes" that can only be analysed in terms of their inputs and outputs" (Christin, 2020). From this position, Christin examines why algorithms are opaque and why this is important and then relates different methods for rendering them transparent or, at least, less opaque. Christin provides a robust discussion of the opacity of algorithms[2], citing Burrell's (2016) analysis that technical opacity has the following characteristics: (1) Algorithms are intentionally secret (companies that own the algorithms recognise their intrinsic value and thus guard them as intellectual property); (2) technical illiteracy may be unavoidable, even when the code is available (the code being too technical for most people to understand); (3) machine learning algorithms have become unintelligible to even highly trained engineers, and (4) the scale of these systems is so large that we cannot fathom which part is responsible for which outcome (Burrell, 2016; Christin, 2020). Christin suggests that these dimensions have resulted in "scholars refer[ring] to algorithms as "black boxes", or devices that can only be understood in terms of their inputs and outputs" (Christin, 2020). Drawing on the work of Pasquale (2015) and Eubanks (2018), Christin argues that the opacity of the algorithms, or black boxes, to use her terminology, is "particularly problematic" as "algorithms are increasingly making decisions hidden behind corporate walls and layers of code...since algorithms are often biased, [as] they draw on historical data...that end up "automating inequality" (Christin, 2020).

1.2. Christin's three ethnographic strategies

After establishing that algorithmic opacity can create harmful outcomes, Christin builds on Seaver's (2017) ethnographic work by offering three ethnographic strategies for studying computational systems: Algorithmic refraction, algorithmic comparison, and algorithmic triangulation. Christin refers to these as enrolment strategies, i.e. ways to use the algorithms as a central part of the ethnographic methodology.

Algorithmic refraction is "derived from physics, [and] refers to the changes in direction and strength that occur whenever a wave of light or sound passes from one medium to the next" (Christin, 2020). Christin applies the idea of refraction to algorithmic systems to invite the ethnographer to consider what changes occur in the presence, or sites, of algorithmic systems. Extending this metaphor allows us to see algorithms as "prisms" that can both "reflect and reconfigure social dynamics" (Christin, 2020). Thus, by studying their use, development and situatedness in social contexts, Christin suggests that ethnographers can begin to understand better (and see through) the "complex chains of human and non-human interventions that together make up algorithmic systems." (Christin, 2020). For example, suppose one was interested in how the algorithm for TikTok worked. In that case, one might study how the use of the platform changed the humans within its ecosystem (users), how the platform (algorithm) adapted based on their behaviour, and how users spoke about it as a result. These "outputs" would indicate changes due to the algorithm. Inferences could then be made about the algorithm and its operation on the human and non-human actors within that system.

Algorithmic comparison involves using multiple sites to examine algorithms through a similarities and differences approach (Christin, 2020). For instance, to study bias in algorithms, we might compare decision-making tools in Human Resources and finance (e.g. hiring algorithms and credit scoring tools), examining the similarities and differences of how they operate and impact users and applicants. Such a comparison would reveal "not only the uses of algorithmic systems but also their inner workings, regardless of how opaque" (Christin, 2020).

Christin proposes directly addressing the methodological requirements of ethnography – saturation, positionality, and disengagement – through *algorithmic triangulation* (2020)**.** To address *saturation* (how large a sample should be), she suggests using various social media platforms to recruit the theoretical sample (Christin, 2020). To understand positionality, the ethnographer can examine how they are perceived and interacted with on these platforms. For disengagement (the challenge of leaving the site and saying goodbye to informants), Christin suggests this can be facilitated by the algorithmic platform being studied (2020).

1.3. Critique – Beyond the Black Box

Assessing Christin's article by her own goal to "offer a toolkit of practical strategies" / "tactics" for conducting ethnographic studies on algorithms (2020), one must ask, are these three strategies helpful to ethnographers? As "tactics", one would expect these to be described in sufficient detail, allowing readers to replicate them in their research (Hennink *et al.*, 2020). I would argue that Christin's algorithmic comparison and algorithmic triangulation discussion does this well, as it incorporates examples from her fieldwork and concretely demonstrates how these strategies would be enacted and to what benefit.

However, algorithmic refraction seems comparatively less tangible (and thus less useful). In discussing this strategy, Christin uses the light metaphor (more than in the other two strategies). Christin refers to algorithmic tools as "prisms that both reflect and reconfigure social dynamics", providing "a useful strategy for ethnographers to bypass algorithmic opacity" (Christin, 2020). Algorithmic refraction may, thus, be more challenging to implement as the steps were less descriptive (despite an example); this causes one to question whether this is a tactic or a way to understand what is happening in the algorithmic system (a theory, perhaps).

While it is beyond the scope of this paper to provide a thorough evaluation of algorithmic triangulation, such an exploration would be valuable regarding whether this approach allows for sufficient reflexivity (Forberg & Schilt, 2023; Markham, 2020). Christin's tactics and discussion could arguably be strengthened by considering the scholarship on ethnography in digital contexts (Forberg & Schilt, 2023; Markham, 2020).

My second criticism of Christin's (2020) article is that while she bases her choice of the black box metaphor on Burrell's (2016) four characteristics of algorithmic opacity, she does not consider alternative metaphors. Christin does not justify why the black box is the best metaphor to use, nor does she recognise that this metaphor might foreclose alternate interpretations of the conceptual space. Thus, my argument is not that the black box is a poor metaphor but that, as "the black box has become the leading image to express opacity in AI" (Möck, 2022), it is crucial to understand what implications this has on research, policy and public discourse. In order to provide the necessary scaffold for my thesis, I will now examine my second article.

2. "Prediction Promises: Towards a Metaphorology of Artificial Intelligence"

Möck's article focuses on the "epistemic significance of metaphors" (2022, p. 121). It explores how philosophical theory can address and reframe the metaphorical images that "co-constitute and shape leading paradigms within socio-technical systems" (Möck, 2022). Möck discusses the "epistemic status of metaphor"; she draws heavily on Hans Blumenberg's work on phenomenology and suggests a methodological framework for a metaphorology of AI (Möck, 2022). To illustrate her argument, Möck provides two examples: the expert and the black box metaphor. Unfortunately, a full review of Möck's paper is beyond the scope of this paper; thus, I will discuss only the two specific aspects that provide the necessary scaffold for my thesis[3]: "the epistemic status of metaphors" and her critique of the black box metaphor.

2.1. How metaphors shape knowledge

Möck asserts that "metaphorical notions serve the communication purpose of making complex concepts graspable" and that metaphors not only reveal what technology is presently capable of (or at least perceived to be capable of), but importantly, metaphors foreshadow what technologies are "supposed to become" (Möck, 2022). However, while metaphors can serve as an "epistemic bridge", helping us to articulate concepts that would otherwise not have words, they also risk obscuring meaning (Möck, 2022). Consider the example of war as a metaphor for debate (*win the argument, shoot holes in the argument*) (Lakoff & Johnson, 2008). While on the one hand, this highlights the combative nature that often arises (the metaphor is a useful epistemic tool), on the other hand, it forecloses the possibility of a mutually beneficial outcome. In war, there is only one winner. Thus, when the war metaphor is invoked, this is the frame of reference through which we see the discussion. However, if we used a dance metaphor instead, we might expect a more mutually beneficial process and outcome (Lakoff & Johnson, 2008). Thus, Möck asserts that we need to analyse our metaphors to understand how we have made sense of our technologies, as this may foreshadow and foreclose future possibilities.

2.2. The problem of the black box metaphor

"...the black box has become the leading image to express opacity in AI" (Möck, 2022). Wiener and Ashby initially used the black box metaphor in cybernetics as both "metaphor and theory". As a theory, the black box model enabled cyberneticians to study the brain's response to its environment despite not understanding how it worked. Thus, the black box functioned as an epistemic tool, serving as a theoretical model and a metaphor for a closed system that was not understood (Möck, 2022). Latour further explored this concept, introducing the term "unboxing", the "process of not only making the inner technical operations of the algorithm transparent but situating the technology within its contextual materiality" (Möck, 2022). This history coalesces into the concept of the metaphor we use today, where writers use the black box to refer to the lack of explainability and interpretability of algorithms.

Möck raises two concerns about the black box metaphor: Firstly, that by focusing too narrowly on the black box, we risk simplifying the problem in AI to a problem only about the algorithm; we fail to see it as "a problem that emerges within socio-technical systems" (Möck, 2022). A broader understanding of the context in which the black box operates demonstrates the additional power and epistemic dynamic at play between the makers and users of black boxes, where not only is the black box's creator superior to those unable to see inside, but the black box itself ultimately becomes superior to all (including its creator), as it has "superhuman capabilities" (Möck, 2022). Secondly, Möck questions "if the constant reproduction of the image of the black box in research might help manifest this dynamic" (Möck, 2022) i.e. if by constantly referring to the black box, even with the positive intention of promoting an agenda of unboxing, we may unwittingly be causing a "closure of debate and strengthen an epistemology of non-understanding that sticks with us in the box's materiality" (Möck, 2022). I will explore these concerns in greater detail in the context of my argument in section three.

2.3. Critique – Towards a Metaphorology of Artificial Intelligence

Möck’s analysis of the epistemic value of metaphors is particularly useful for scholars, providing a foundation for research on metaphors in AI. Furthermore, she makes a novel contribution in her article by advancing Hans Blumenberg’s metaphorology to include political considerations critical for a framework in the AI context (Möck, 2022). Möck’s proposed metaphorology of AI, thus, recommends that we engage along four dimensions: (1) Examine the *history* of AI metaphors, (2) reveal *motivations* of AI researchers through the metaphors they use, (3) understand “what metaphors of AI can tell us about humans and their needs”, and (4) consider the “political aspects of the imaginaries and the material-political embeddedness of the dominant narratives” (2022, p. 126). This framework provides a tangible way for scholars to research AI metaphors, as Möck exemplifies in her article. However, while Möck uses two examples to demonstrate how this helps to surface and explore each of the four epistemic dimensions of the metaphor, the framework does not seem to encourage the search for alternative metaphors, nor an exploration of what these metaphors might be missing. Möck might argue that her frame of reference is philosophical[4] and that her goal is to frame the issue, not to solve it, i.e. not to provide alternate metaphors, but to elucidate the problems with those being used. That may be a defensible stance for Möck. However, to ensure that we widen the metaphorical landscape and increase the possibilities for future research, policy and discourse, one could argue that we need an approach that generates more metaphors, not only questions the ones we have. In section four, I consider alternative questions to stimulate the generation of more metaphors to address this critique (Maas, 2023).

3. Why New Metaphors for AI Might Support Different Futures

In this section, I explore how metaphors influence ontology and epistemology and thus inform methodology in AI. I use Christin's (2020) example of the black box metaphor to demonstrate that an over-reliance on one metaphor forecloses one's ontological framework, thus potentially limiting epistemological and methodological choices, with implications for future discourse, research and policy.

3.1. The black box metaphor – influences ontology and epistemology

At its core, this is primarily a critique of language, which I argue is valid for two reasons. Firstly, language matters (Lakoff & Johnson, 2008). The metaphors we use have a real-world impact: They shape innovation, spur or halt investment, inform the study of technologies, and help to set regulatory agendas (Ganesh, 2022; Maas, 2023). For example, the current narrative of the potential of Artificial General Intelligence to become a *superintelligence,* capable of solving the world's most intractable problems, has arguably contributed to driving significant investment and research. In terms of regulatory implications, conceiving this future technology as a *superintelligence* has implications for the nature of the regulation that is developed (Maas, 2023). If it is *intelligent,* what legal rights does it have? If it can solve all problems, should we risk over-regulation, thereby slowing it down?

Secondly, if we accept Christin's position that addressing the opacity of algorithms is necessary and that ethnography is a valuable method to do so, then language and metaphor are central, as they are core to the ethnographic method (Rabinowitz *et al.*, 2018; Geertz, 1973; Marcus, 2021; Gullion, 2021; Seaver, 2017). Thus, to understand Christin's ontological and epistemological frame of reference concerning algorithms, I have tried to adhere to Marcus' call to "follow the metaphor" and let Christin's language speak rather than our assumptions about what might typically be constructed by ethnographic research (Marcus, 2021).

3.2. The black box metaphor – locks us in

Christin's article is well-intentioned and arguably both necessary and helpful – her specific strategies respond to Seaver's call for concrete ethnographic strategies to study algorithms (Christin, 2020; Seaver, 2021). However, Möck might point out that Christin's use of the black box metaphor and associated language may reinforce the "box's materiality" with unintended consequences (Möck, 2022). We can see how this metaphor has influenced Christin's language throughout the article and her resulting approach to methodology.

By conceptualising the algorithm as a black box, Christin's ontological and epistemological frame of reference becomes scientific. The algorithm occupies one side of a light spectrum, representing the greatest opacity/darkness. As a result, the ethnographer works to "shed light on the complex intermingling of social, cultural, and technological aspects of computational systems in our daily lives", rendering the algorithms transparent (Christin, 2020). Christin develops her ethnographic strategies from this frame, as the "concept of refraction is derived from physics" (Christin, 2020).

While using metaphors in technology and science may be unavoidable, "the less familiar we feel with a technology, the greater our need for visual language as a set of epistemic crutches" (Sommerer, 2022). The use of the black box metaphor invites a particular framing of algorithms and the systems that create them, one that risks excluding humans by "obscure [ing] our view of the people behind the algorithmic systems and their value judgements…falsely suggest[ing] that algorithms are independent of human prejudices" (Sommerer, 2022). Ontologically, the black box occupies a materialist paradigm, where the algorithm's inner workings are sealed off, solidified and unknowable. Christin offers ways to render this black box knowable within this scientific paradigm through tools like refraction. The implication for epistemology is that by using the tools, the ethnographer can shed light on the algorithm or the system, thus creating knowledge that was not accessible before. While this may not be Christin's intention, one might argue that such language could suggest an empirical, if not post-positivist, epistemology (Lincoln & Guba, 2013; Malik & Malik, 2021; Omodan, 2022). Christin may disagree with this assessment. My intention is not to suggest that empiricism (nor post-positivism) is at odds with ethnography (Williams, 2020), but merely that as the metaphor creates ontological and epistemological paradigms, a more explicit explanation of one's epistemological framework becomes necessary.

To summarise – by framing algorithms as black boxes, Christin's ontological frame of reference positions algorithms as material, closed systems that are difficult to access. Consequently, knowledge of the system must be gained through direct means. Unsurprisingly, Christin's methodological strategies are thus empirical and inspired by the scientific paradigm (e.g. refraction). While this may be fruitful, offering new tangible strategies for ethnographers to study algorithms (Christin, 2020), there is a risk that by focusing on the "box's materiality", we might be "distract[ed] from the ethical or epistemic problems of these models" (Möck, 2022). If it is material, it cannot become non-material. Thus, once the black box metaphor has been asserted, it is difficult to move beyond it, even if that is the stated intention.

3.3. The black box metaphor – has real-world implications

The black box metaphor has implications for policymakers, who continue to see algorithms' opacity as intractable and a growing risk to humanity because we do not have control over them (Sommerer, 2022). However, the less control humans are perceived to have over these systems, the less responsibility they subsequently have, which increases the power of these systems and reduces the agency and accountability of the people involved in creating them (Maas, 2023; Sommerer, 2022). Thus, researchers are increasingly concerned about the continued use of this metaphor (Lehr & Ohm, 2017; Maas, 2023; Marcus, 2021; Möck, 2022; Sommerer, 2022).

Furthermore, some scholars suggest that this metaphor is neither technically accurate nor practically helpful (Murray-Rust *et al.*, 2022). Others argue that "the steps of playing with the data are actually quite articulable" and that the black box creates a "misimpression that machine-learning systems spring into being fully formed and are impenetrable" (Lehr & Ohm, 2017). To reconsider such a ubiquitous metaphor, however, alternative solutions will be required. Thus, in the final section, I explore possible solutions to prevent metaphorical foreclosure when discussing algorithms.

4. New metaphors, new frames, new futures

As researchers continue to address algorithmic opacity, one might consider a new metaphor to address the problems arising from an overreliance on the black box metaphor (Sommerer, 2022). Seaver's ethnographic work on recommender systems is one potential source of inspiration (Seaver, 2021). Seaver's metaphorical landscape includes human and other organic images, like the gardener whose "curation" "maintains[s] balance in the garden", that is, the algorithmic system (Seaver, 2021). Seaver's garden metaphor, relates the algorithm to something more organic and tangible, something that is curated and nurtured by the human gardeners who care for it and conscientiously prune it according to an intentional design. By following the metaphor, we understand Seaver's ontological frame as different to Christin's. In Seaver's worldview, humans have more agency; through their care, they can shape algorithms. However, even Seaver notes that some of his respondents refer to themselves not as gardeners but as data cleaners (Seaver, 2021), suggesting that the garden metaphor does not paint the complete picture.

Another alternative to the black box is the “algorithmic veil” articulated by Lucia Sommerer (2022), whose primary concern with the black box metaphor is that it "falsely suggests that the algorithms are independent of human prejudice" (2022). In her description, the algorithmic veil overcomes this issue as it is an item that, by definition, relates to the human form, inviting one to draw it back to see behind it. The veil is of a different nature to the black box, suggesting that the algorithm would inhabit a different ontology and epistemology, one that is less fixed, more translucent and something with which humans could interact (perhaps more co-constitutive) (Sommerer, 2022). By their very nature, veils allow one to see the subject beneath the veil (despite obscuring the image to the onlooker), it may be possible to both identify the obscured image and reveal the true image if the veil is lifted. Arguably, this could be likened to people trying to make sense of how an algorithm performs. Unfortunately, Möck did not fully develop this metaphor, so it is not clear how it should be interpreted. As a result, one might argue that aspects of algorithmic opacity are missing and that this image does not sufficiently demonstrate the vast complexity and interconnectedness of these systems.

Finding both metaphors unable to fully explain algorithmic opacity, I tried to develop a new metaphor. Consider the metaphor of a spider's web that spans a multi-dimensional space. The spider’s web metaphor may offer some benefits over the black box in that it is organic (thus allowing humans to act on and in the system); it inherently demonstrates high levels of interconnectivity; it is highly complex yet transparent, so it does not feel as intractable; and is sensitive to interdependencies (i.e. things that affect one part of the web impact other parts). If we used this metaphor to extend Christin’s article, one could imagine a title: “Beyond the black box – into the spider’s web”. By conceiving of algorithmic opacity as a spider’s web rather than a black box, the ethnographer might consider other techniques such as: Detangling (what concepts, narratives, and stakeholders are weaved together and enmeshed in the narratives and systems?), locating the source of attachments (as a spider’s web attaches to objects for structural integrity, the ethnographer might ask what socio-technical or political foundations underpin the narratives revealed through the ethnography); and looking for who/what is caught in the web (who are the algorithms acting on and to what effect?).

However, simply offering an alternate metaphor misses the overall point of this argument (Maas, 2023). If we rely only on one metaphor for our understanding and shaping of how we see algorithms (or any concept), then we narrow our frame of reference to only that particular image; like an aperture, it forecloses other ways of seeing algorithmic opacity, other research agendas and thus potential futures. Similarly, it foreshadows likely outcomes by framing algorithms in a particular way (Maas, 2023). Over-reliance on any metaphor (even a powerful one) can risk consumers and users of those metaphors becoming unreflective.

The solution is not to abandon metaphors; they are critical epistemic bridges between complex, inarticulable concepts and our current language framework. However, we must be more reflexive in using and consuming these metaphors (Möck, 2022). We need to interrogate our use of metaphors to make explicit the work they are doing. Möck’s metaphorology of AI has been discussed as a valuable approach to deepening our understanding of the metaphors we use and their histories, socio-political contexts, and implications. Maas (2023) provides an additional method by which we can examine our metaphors and metaphorical landscapes. This framework has an advantage over the more philosophical metaphorological framework, as it encourages the development of a broader range of metaphors.

Maas’ (2023) five-step process for evaluating metaphors invites one to ask a series of questions about the metaphor in question: (1) What foundational metaphors are being used?; (2) What other metaphors could describe the same features?; (3) What aspects does the metaphor capture well?; (4) What aspects does the metaphor not capture, and what are its consequences?; and (5) What are the regulatory implications of this metaphor? It is the second question that, I believe, brings the most significant opportunity for increasing the range of metaphors used to describe a conceptual space.

This approach could be used to create language that better captures the metaphorical landscape of concepts, like algorithmic opacity, ensuring a broader range of metaphors to describe the phenomena in AI. However, this approach may be practically challenging, as it can be hard to find powerful metaphors.

Conclusion

"Metaphor is pervasive in everyday life, not just in language but in thought and action…Our ordinary conceptual system…is fundamentally metaphorical in nature" (Lakoff & Johnson, 2008). The implications of this are critical in artificial intelligence, the term being a metaphor itself, where we must be intentional and reflexive about the language we use and consume. In this paper, I explored how metaphors influence ontology and epistemology, informing methodology, through a critical review of two articles.

While providing tangible strategies for ethnographic fieldwork, Christin's reliance on the black box metaphor provided a focusing example for my thesis. Drawing on Möck's discussion of the importance of metaphors and her critique of the black box, I argued that to move "beyond the black box," we must expand our metaphorical range when considering concepts like algorithmic opacity. Having argued that the black box is limited as a singular metaphor, I recognise that the solution is not simply to provide a single better metaphor, as this, too, would risk limiting other possible conceptions and possibilities of meaning.

The future of AI research and policy would benefit from greater reflexivity, where we examine the metaphors, we use and consume and introduce a broader range of metaphors. This approach will ensure that we expand the description of our concepts and thus our understanding of these technologies; in so doing, we increase the range of possible future outcomes.

References

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 205395171562251. https://doi.org/10.1177/2053951715622512

Christin, A. (2020). The ethnographer and the algorithm: Beyond the black box. *Theory and Society*, *49*(5), 897–918. https://doi.org/10.1007/s11186-020-09411-3

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group.

Forberg, P., & Schilt, K. (2023). What is ethnographic about digital ethnography? A sociological perspective. *Frontiers in Sociology*, *8*. https://doi.org/10.3389/fsoc.2023.1156776

Ganesh, M. I. (2022). Between metaphor and meaning: AI and being human. *Interactions*, *29*(5), 58–62. https://doi.org/10.1145/3551669

Geertz, C. (1973). *The interpretation of cultures* (Vol. 5019). Basic books.

Gullion, J. S. (2021). *Writing Ethnography (Second Edition)*. BRILL. Hennink, M.,

Hutter, I., & Bailey, A. (2020). *Qualitative Research Methods*. SAGE.

Killam, L. (2013). *Research terminology simplified: Paradigms, axiology, ontology, epistemology and methodology*.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press. https://books.google.com/books?hl=en&lr=&id=r6nOYYtxzUoC&oi=fnd&pg=PR7&dq=metaphors+we+live+by+lakoff&ots=L

ps3cm5t5W&sig=zALxRVad4upVfBtYfWCBJpJQNk0

Lehr, D., & Ohm, P. (n.d.). *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*. *51*.

Lincoln, Y. S., & Guba, E. G. (2013). *The Constructivist Credo*. Taylor & Francis Group. http://ebookcentral.proquest.com/lib/cam/detail.action?docID=1187038

Maas, M. M. (2023). *AI is Like… A Literature Review of AI Metaphors and Why They Matter for Policy* (SSRN Scholarly Paper 4612468). https://doi.org/10.2139/ssrn.4612468

Malik, M., & Malik, M. M. (2021). Critical Technical Awakenings. *Journal of Social Computing*, *2*(4), 365–384. Journal of Social Computing. https://doi.org/10.23919/JSC.2021.0035

Marcus, G. E. (2021). *Ethnography through thick and thin*. Princeton University Press. https://books.google.com/books?hl=en&lr=&id=rskkEAAAQBAJ&oi=fnd&pg=PP9&dq=2++BOOK+Ethnography+through+thick+and+thin+/+George+E.+Marcus.&ots=6y6aHYxlrD&sig=2JL1WqifD8QF3Gvi3CTP0zrzmSI

Markham, A. (2020). *Doing ethnographic research in the digital age*. OSF. https://doi.org/10.31235/osf.io/hqm4g

Möck, L. A. (2022). Prediction Promises: Towards a Metaphorology of Artificial Intelligence. *Journal of Aesthetics and Phenomenology*, *9*(2), 119–139. https://doi.org/10.1080/20539320.2022.2143654

Murray-Rust, D., Nicenboim, I., & Lockton, D. (2022). Metaphors for designers working with AI. *DRS Biennial Conference Series*. https://dl.designresearchsociety.org/drsconferencepapers/drs2022/researchpapers/237

Omodan, B. I. (2022). A Model for Selecting Theoretical Framework through Epistemology of Research Paradigms. *African Journal of Inter/Multidisciplinary Studies*, *4*(1), 275–285. https://doi.org/10.51415/ajims.v4i1.1022

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018). *Machine Theory of Mind* (arXiv:1802.07740). arXiv. https://doi.org/10.48550/arXiv.1802.07740

Rawnsley, M. M. (1998). Ontology, Epistemology, and Methodology: A Clarification. *Nursing Science Quarterly*, *11*(1), 2–4. https://doi.org/10.1177/089431849801100102

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2), 205395171773810. https://doi.org/10.1177/2053951717738104

Seaver, N. (2021). Care and Scale: Decorrelative Ethics in Algorithmic Recommendation. *Cultural Anthropology*, *36*(3), Article 3. https://doi.org/10.14506/ca36.3.11

Seaver, N. (2022). Computing Taste: Algorithms and the Makers of Music Recommendation. In *Computing Taste*. University of Chicago Press. https://doi.org/10.7208/chicago/9780

226822969

Sommerer, L. (2022, February 1). From Black Box to Algorithmic Veil: Why the image of the black box is harmful to the regulation of AI. *Better Images of AI Blog*. https://blog.betterimagesofai.org/from-black-box-to-algorithmic-veil-why-the image-of-the-black-box-is-harmful-to-the-regulation-of-ai

Williams, R. T. (2020). The Paradigm Wars: Is MMR Really a Solution? *American Journal of Trade and Policy*, *7*(3), 79–84. https://doi.org/10.18034/ajtp.v7i3.507

CJAI

CAMBRIDGE
JOURNAL OF
ARTIFICIAL
INTELLIGENCE