Documentation as a Tool for Algorithmic Accountability

by

Taylor Lynn Curtis

B S Eng.
McGill, Faculty of Engineering, 2022

Submitted to the Institute for Data, Systems, and Society
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN TECHNOLOGY POLICY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by: Taylor Lynn Curtis
Institute for Data, Systems and Society
May 21, 2024

Certified by: Dylan Hadfield-Menell
Assistant Professor, Department of EECS, Thesis Supervisor

Accepted by: Frank R. Field, III
Senior Research Engineer, Institute for Data, Systems, and Society
Interim Director, Technology and Policy Program

Documentation as a Tool for Algorithmic Accountability

by

Taylor Lynn Curtis

Submitted to the Institute for Data, Systems, and Society
on May 21, 2024 in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN TECHNOLOGY POLICY

ABSTRACT

This thesis argues that civil liability should rest on the deployer's understanding of system behavior, and that documentation is the necessary tool to accomplish this goal. This work begins by establishing the "hole" in current approaches to AI risk regulation, the lack of a civil liability regime. It also highlights that civil liability is an already existing and effective regulatory tool that can be applied to AI. The rest of this thesis develops this argument by looking at what is necessary for such a framework to exist. It argues that an understanding of system behaviour is essential and achievable through documentation. It is divided into two substantive chapters. Firstly, Chapter 2 outlines how system behaviour can inform policy through documentation, linking the necessity of documentation to liability and proposing a concrete liability scheme based on documenting system understanding. Secondly, Chapter 3 discusses how documentation can alter a person's understanding of system behaviour, presenting a user study that demonstrates how system understanding can be achieved through documentation and structured data interaction. It argues that testing and system understanding are not insurmountable challenges and that by engaging in a relatively simple process, AI deployers can better understand the behaviour of their models. Overall, this thesis provides a methodical guide to understanding AI system behaviour and the establishment of a new pathway for effective regulation, arguing for the understanding of system behaviour and documentation at deployment as the path forward to achieve civil liability in AI.

Thesis supervisor: Dylan Hadfield-Menell
Title: Assistant Professor, Department of EECS

Acknowledgments

First off, thank you to my family. Specifically, my mom for always being there for me and believing in me when I couldn't, even when she was going through one of the hardest points of her life. My dad, for his amazing support and thoughtfulness, and his insights into how research can and should be conducted. Thanks to my brother, Aaron, who has also made me feel intelligent, capable, and maybe sometimes just a little too much like the younger sister.

Thank you to Anssam Ghezala, not only did you help power me through the last couple of months (and hosted me in Montréal when I needed a break), but you also helped me debug react.js past the point when most people would've given up. Even after completing a couple of react-based projects, I can confirm I still have terrible react style even though there is no agreed-upon style. Additionally, thank you to Emilie Ladouceur. We have had too many voice messages in the past few months, that is all I can say. But, thank you for giving me valuable insight into the psychology literature.

More generally, thank you to all my friends for your support. Both from before MIT and friends that I've met here (y'all know who you are). This also includes my friends' pets (thanks Sia and Beau)! Thanks to MIT Woman's Ice Hockey team, y'all are great and I'm so glad I got to play with you (special shoutout to Mal for their leadership)!

Thank you to the current and former members of the Algorithmic Alignment group for your time and feedback. In particular, thank you Rui-Jie for being my introduction to the AI policy world and always listening to my half-baked ideas (I appreciate you my 5'2" and delulu friend). Thank you to Deepika and Prajna, you both always gave me the confidence and power to produce a thesis (thank you Deepika for the texts to make sure I was on schedule and Prajna for your kindness and check-ins). Also, thanks to Stewy & Tim for putting up with my request to generate word embeddings on very short notice.

Last but not least, thank you to my advisor Dylan. I appreciate your guidance and insight into how research is done. As I once told you, I've learned a lot from you, and ever more so in ways I did not expect.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Artificial intelligence (AI) harm exists. While this is an unsurprising statement, as every new technology is not value-neutral and carries with it societal implications (see Stone [1] and Winner [2]), the AI community has a unique approach to managing and mitigating this harm. Specifically, AI harm has been thought to produce an AI risk [3]. This social construction is what frames the modern discussion on the effects of AI. Furthermore, the explosion of generative, and "general-purpose" models intensified this regulatory framing. This is especially true because general-purpose models often lack a specific use and exhibit novel and surprising failure modes [4].

Indeed, in a growing body of work, several AI researchers have argued that sufficiently general systems carry risks that emerge from training, either as it grows awareness of the wider world it operates in or the power it can exert over others in its environment [5]–[7]. As AI progresses, there are more general, non-specific use cases that are difficult to understand and can pose a risk for discriminatory, unsafe, or incorrect content [8]–[12]. Therefore, the current AI landscape has served to entrench the importance and prevalence of AI risk as a discussion and investigation.

Societal framing matters. In choosing to portray AI harm as risk, AI risk becomes what is frequently referenced **as the motivation for policy** and is becoming central to legal discussion [13]. Particularly, risk is becoming the primary language surrounding *algorithmic accountability* when discussing legal frameworks. This has served to intertwine the risk discussion with the accountability discussion. However, AI risk as policy is simply one aspect of the AI accountability or algorithmic accountability discussion. In traditional terms, according to Bovens [14], Lindberg [15], Mulgan [16], and Thynne and Goldring [17] as cited in Novelli, Taddeo, and Floridi [18], accountability is simply the obligation to inform and justify behaviour to authority. It has four dimensions in a policy space: compliance, reporting, oversight, and enforcement [18]. Translating this discussion to AI or algorithmic accountability, it becomes *how developers, deployers, and designers can ensure the proper functioning of an AI algorithm throughout its lifecycle* [19].

However, there is no one, specific approach to algorithmic accountability [18]. For example, Kroll [20] argues for traceability as a foundation of accountability, which is contrasted by a step-by-step identification and evaluation process outlined in the March 27, 2024, AI Accountability Policy Report from the Nation Telecommunications and Information Administration [21]. This thesis does not attempt to propose a new accountability system. Instead, it accepts the centricity of the AI risk discussion to the AI accountability discussion. This will ultimately contribute to what is known as the accountability ecosystem for AI, a series of tools and regulatory approaches all working together to achieve the goal of algorithmic accountability [22]. It aims to contribute, in part, to the discussion of AI accountability by examining AI risk management approaches. Through this examination, this thesis will establish what is "missing" from current risk approaches that can translate a pure risk framework to an accountability methodology. The rest of this introduction will examine risk management frameworks, identify a common "hole" in this kind of legislation, and then establish a model deployment documentation strategy and civil liability as the essential tools to overcome this "hole."

1.1 AI Risk and Risk Management Frameworks

To understand AI risk management frameworks, the concept of AI risk must first be addressed. There is no one agreed-upon definition for the specifics of this risk. For example, Curtis, Gillespie, and Lockey [23] describes AI risk as a broader harm to society. Others have suggested that AI risk is the potential for harm at either the individual or community level as a negative consequence of using an AI model [24], [25]. Further, other scholars have gone deeper and claimed that AI risk is systemic and immediate with foreseeable dimensions [26], [27]. However, regardless of specificity, scholars recognize that AI risk can refer to harms either current or potential [3].

This has produced a policy agenda in AI that is largely shaped by the discussion of AI *as a risk*. Thus, while the specifics of the kind of risk are up for debate, the discussion is inevitably intertwined with the discussion of current and potential harm. This has led to a widespread agreement that some kind of regulatory framework must be established to mitigate and avoid potential harms [28]. The dominant solutions have become risk management frameworks. For example, the soft law standard of the National Institute of Standards and Technology (NIST) Risk Management Framework is at the top of the United States policy discussion [29]. In the European Union (EU), the approach of the EU AI Act is similar, taking a risk management approach that relies upon an intertwining of product safety and fundamental rights regimes [30]. This is not to suggest that all risk-based approaches are equal. In AI this is clearly not the case, as highlighted by Kaminski [3], and others that have noted differences between the NIST framework, the EU AI Act, and other risk-based evaluations [31]–[33].

Despite the differences between these key approaches, there are some generalizations that the adoption of the risk narrative provides. Namely, no risk management framework will prevent all harm. Two kinds of harm happen within risk management frameworks: (1) harm that is incurred due to risks that fall outside the framework and (2) harm that occurs despite evaluation under the framework ("acceptable" harm or "irreducible" harm). To better understand each of these situations, consider the two examples below.

Firstly, there is the harm that can occur outside the confines of any risk framework. To understand *how* this harm emerges, consider the specific example of the EU AI Act and its risk structure. This risk framework focuses on the use case of AI to determine the level of risk, for example, the use of AI in healthcare is considered high risk and is thus subject to a higher level of protection [34].

However, there is a gap in this regulatory framing. On April 13, 2023, the AI Now Institute released a policy brief, raising concerns about the risks of general-purpose artificial intelligence (AI) and the inability of the proposed European Union (EU) AI Act to target generality as a separate risk category [35]. As suggested by AI Now, the inherent risks arising from generality are not being properly accounted for in the current EU framework. Other scholars have observed this about the proposed EU framework. For example, De Cooman [36], in analyzing the EU AI Act *ratione materiae* (the definition of features of AI). While general-purpose AI is not explicitly mentioned, it does state that just because the EU "exhaustively enumerates high-risk AI systems does not mean the residual category displays non-high-risk" [36, p. 50]. Ultimately, the risk frameworks that are currently in existence (such as the EU risk framework) that focus on the effects in each use case are not *entirely* sufficient [13], [36]. However, despite these limitations, it is possible to leverage the benefits of the EU AI Act, and specifically the risk categories [37].

Secondly, there is the notion of "acceptable" harm. "Risk-based frameworks look principally to control relevant risks, not to secure compliance with sets of rules" [38, p. 281]. As such, risk regulation is not a perfect system of control, and there will be leakages. These "leakages" occur due to uncertainty. The world is uncertain, and not entirely predictable. Science, in modelling this world, encounters what is known as irreducible uncertainty [39]. AI is not exempted from this trend and will fall into the same pitfalls. Thus, when dealing with risk regulation you *must* accept that a risk framework can not cover everything.

This lack of sufficiency of risk regulation is where this thesis begins. The next section of this chapter will outline this "hole" and introduce a regulatory strategy (liability) that can address it.

1.2 The "Hole" in Risk Management Frameworks

Risk-based regulation of AI algorithms is not going to be enough. This thesis does not aim to critically evaluate the use of AI risk management frameworks. Instead, it accepts the trend of risk management as the emerging regulation for AI. However, a risk management framework *is a choice*, and it is a choice that has consequences. Thus, when using such a framework, we are establishing a preventive or preemptive regime that attempts to **foresee** rather than **react** to harm occurring. As established above, this kind of regulation not only allows harm, but *expects* it. This means there needs to be a complementary set of regulations that offer a path for redress. There needs to be a supporting retro-active policy arrangement.

In fact, the need for a retro-active approach is not a new concept in policy and law. Risk management regulation and risk regulation have historically been backed by a regime of tort liability [3], [13], [40]. Thus, this thesis proposes the use of **tort or civil liability** as a way to support this risk based approach.

Civil liability is an effective regulatory tool, that is not only useful as a support to risk regulatory frameworks. Furthermore, it is an already existing regulatory framework in the United States. This means that there is an *efficient* and *effective* pathway to regulation that can be adapted to a new area. This would ease the political burden and difficulty of creating new regulations.

So, if civil liability exists, and is a well-established regulatory solution, then the answer may seem simple: we should simply apply tort liability to AI systems. However, this is not straightforward [41], [42]. A core challenge is that scholars do not agree on the appropriate analogy or even considerations - for example, Zech [43] highlights the policy consideration of which agent exercises control. The other core challenge is the lack of predictability or foreseeable of AI systems, particularly in the context of generative models [42]. This can even counter the development of safer systems, as deployers may face more liability for systems that they understand, creating misaligned or even perverse incentives. The rest of this section will outline current solutions to algorithmic liability, and suggest how documentation at deployment could be used to solve it.

1.3 A Potential Solution

The next important point to highlight is *why* algorithmic liability is challenging. To establish this point, this work begins by highlighting two current trends of algorithmic liability. These two areas are:

- The actions of AI systems and the type of harm they can cause, which is subject to only strict liability [41], [44]
- The oversight of decision-making processes by a human [45]

Thus, liability is either strictly imposed, or, in the context of algorithmic decision-making in particular, relies on a human to make the "final" judgement, and simply use AI as a tool to inform *their* decision-making.

These approaches fail to capture (1) the consideration of the actions taken on behalf of those who choose to use AI systems, and (2) that involving humans at random points in the design process can intensify instead of solving issues with AI systems.

In Chapter 2, this discussion of liability, current approaches, and a liability scheme will be further elaborated upon, however, this base of inadequacy of liability is what this thesis claims to solve. We propose that the issues with current liability regimes can be overcome by incorporating *specific* and *intentional* documentation elements into the algorithmic design process. The next sub-section will briefly introduce documentation as a solution, and then present the unifying thesis statement of the rest of this work.

1.3.1 Documentation

To reduce harm, many of these model accountability frameworks call for documentation throughout the entire process [20], [46], [47]. This is supplemented by pieces that call for and provide AI-model-specific guidelines for documentation [48]–[50].

In contrast to the general calls, this thesis contributes a concrete approach to model documentation that records the system deployer's understanding of the model behaviour and justification of that behaviour. This would form the basis of proof for **system understanding**. This thesis argues that civil liability should rest on this understanding of system behaviour, and that documentation is the necessary tool to accomplish this goal.

1.3.2 Roadmap

To make this argument, the rest of this thesis will be divided into two substantive chapters. Firstly, Chapter Two will outline how system behaviour can be considered in policy through documentation. This theoretical argument will link the necessity of documentation to liability. It will then offer a concrete liability scheme based on documentation of system understanding.

Secondly, Chapter Three will outline what documentation of system understanding could look like. By employing a user study, it will argue that system understanding can be achieved through documentation and structured data interaction. It will show that testing and system understanding are not impossible to determine, and by engaging in a relatively simple process, AI developers and deployers can understand the behaviour of their models.

This introduction has established that algorithmic accountability, in the current regulatory landscape, relies upon civil liability. Furthermore, it has shown that civil liability is not just a necessity of the regulatory reality, but an optimal option for quick adoption, as many civil liability decisions already exist. Finally, the remaining two substantive chapters will demonstrate the overarching thesis of this work. Algorithmic civil liability should be based on the understanding of a system before deployment, as this is a feasible policy strategy, and we can design systems to help invdividuals understand the models they create.

Chapter 2

Understanding as a Policy Strategy

This chapter examines liability in AI and demonstrates how those who choose to deploy AI models can be incentivized to inform policymakers. As is argued by Hammond [39], the role of scientists is to reduce uncertainty as much as they can and give policymakers the tools to adapt to that. By applying this adage to the field of AI, this chapter aims to demonstrate that AI, while a new technology, poses the same problems that have always existed in technology policy. It builds upon the work of Buiten [51], in that it does not rely on post-hoc explainability and transparency, it instead looks at what kind of information programmers can provide as a way of managing AI risk/harm.

There are several ways AI risk/harm can be managed. For example, the EU has just passed the EU AI Act, a risk-based framework. While the National Institute of Standards and Technology (NIST) has launched its own risk-based framework, the horizon for the official adoption of such an approach in the U.S. context seems unlikely. Furthermore, what these approaches do not cover is what the standard or best practice is.

There is no clear answer right now as to what a best practice or standards should be for AI use and deployment [23], [52], [53]. There needs to be an experimental stage, as it is only through different approaches that we can select the 'right' one. Indeed, this experimentalist approach and recognition of more solutions rather than 'the' solution is a new kind of legal and institutional thought [54]. A simple step to enable varied practice, and also see the results, is to look at the documentation that is generated by deployers.

This is why this thesis defines documentation at the time of AI model deployment. Specifically, this thesis recommend documentation of the deployment decision, including the intended use (or use cases) as well as its behaviour within those use cases. This thesis is calling for deployers to commit to a deployment context and document it. Documentation is the only artifact of the decisions of the deployer. It is the only way (from an external perspective) to see what practices deployers decide to follow.

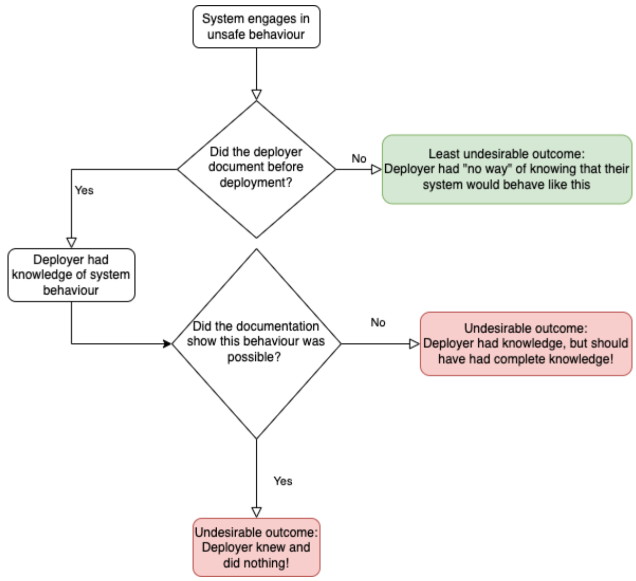

Figure 2.1: Overview of the incentive structure

However, current incentive structures can make this an unattainable goal. There is no incentive structure in the U.S. to encourage companies to engage in this kind of documentation [55]. In fact, the incentives currently encourage less-than-safe practices. Consider a system that performs unsafe/undesirable behaviour. If they documented the system, they would have knowledge about the system. If they find nothing, they have the undesirable outcome of their documentation being incomplete. If they find something, then they also experience an undesirable outcome of having the knowledge but not acting on it. Now, consider if they didn't document their system. They could claim (either truthfully or not) that they could not foresee the system's behaviour. This becomes the best state, as they are no longer perceived as responsible for the system's behaviour. Overall, documentation is disincentivized as it is easier to claim no knowledge of system behaviour *than* knowledge and failure to correct.

As such, we need to change the incentive structures. One tool could be civil liability. However, there is a problem in holding AI models liable [41], [42]. This problem is only intensified in the case of generative models. A model can not be held liable the way a person can be, as it is an agent of whoever chooses to deploy it. This creates a problem as there is no clear path to holding those who choose to deploy AI models liable for the processes they set in motion.

A potential solution could be through strict liability. Per Zech [43, p. 152], "strict liability assigns the economic risk to the injurer regardless of whether the injurer behaves in accordance with existing duties or not." However, strict liability does not promote the adoption of safe behaviour and can have a limited use [44]. Thus, we need a different kind of liability to enable and incentivize this practice.

Our solution is to base liability on the documentation of deployment. More precisely, this thesis argues that it is essential and useful to document at the deployment stage and we need incentives utilize existing civil liability frameworks to make this happen

2.1 Related Work

2.1.1 Ethical AI Principles

Recently, multiple organizations have been instituting principles or standards that they view as important for producing ethical or responsible AI. While ethical or responsible AI is generally agreed upon as a priority for AI systems, the exact definition of what makes AI ethical is still an open discussion [52]. Nonetheless, there is an emerging coherence around certain principles that AI must possess to be deemed ethical. This is observed by Jobin, Ienca, and Vayena [52] in their survey of the global landscape of AI guidelines, where they note a trend towards 5 clear principles: transparency, justice and fairness, non-maleficence, responsibility, and privacy. In another approach, Fjeld, Achten, Hilligoss, *et al.* [19] identified privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values, while also mentioning human rights. Finally, the human-centred AI literature is also contributing a substantial number of guidelines. Shneiderman [56] proposes 15 different principles, many of which overlap with those already mentioned.

In addition to these formal presentations of AI ethical principles, multiple organizations are contributing to the landscape of ethical deployment. For example, OpenAI (in conjunction with Cohere and AI21 Labs), a research and development company, released a company blog post that describes principles that must be followed to safely deploy large language models (LLMs) [57]. While there is some consistency, each AI system is ultimately subjected to different priorities and ethical concerns. Thus, a principle-based approach, while centralizing the ultimate goal of ethical AI, will fall short of standardizing expectations for AI systems. Furthermore, even if we could establish agreed-upon principles, they are often abstract and difficult to translate directly into an AI system. Hence, the creation of ethical principles is not sufficient, work must be done to operationalize or translate these principles into AI system development practice.

2.1.2 Beyond AI Principles and into Implementation

While ethical principles and guidelines may be effective for establishing priorities, there remains a challenge in operationalizing these principles. If the goal is responsible or ethical AI, then it becomes important to consider how these principles are being adopted. Kloker, Fleiß, Koeth, *et al.* [58] propose a potential approach to evaluating the level of explainability (a common ethical AI principle) that is expected of an AI system. In their proposal, they acknowledge the need to configure the balance between trust and caution separately for each use case of an AI system, subject to the level of risk. This concept of trust is expanded upon by Zhu, Xu, Lu, *et al.* [59], where they differentiate between the concepts of trust and trustworthiness to operationalize AI ethics. They argue that trustworthiness is technical whereas trust is dependent on stakeholder perception. First, trustworthiness must be established through different product and process mechanisms before being presented to stakeholders to establish trust.

Taking a different approach, Percy, Dragicevic, Sarkar, *et al.* [22] discusses how transparency (and eventually an accreditation system) could emerge from establishing clear explainability and documentation standards. In a more general survey of implementation, Boza and Evgeniou [60] discusses the necessity of AI principles entering an operationalization phase. They note that this is starting to happen and that four different methods are emerging to support this process: AI toolkits, auditing, documentation and organizational processes, and standards and certification. Of particular interest to our proposal is the inclusion of documentation as a way of operationalizing or implementing ethical AI. Boza and Evgeniou [60] emphasize model reporting, model cards, and factsheets as documentation procedures at the model level. The next section will explore the concept of AI documentation standards and the effect this can have on producing ethical or responsible AI.

AI Accountability Frameworks

There has been a push from systems engineering to formalize these processes and concepts. One such approach is taken by Kroll [20] to define requirements for traceability. This all-encompassing software development lifecycle outlines how accountability can be operationalized into the whole process of creating and deploying a system. This systems approach is supplemented by other accountability mechanisms. For example, Hutchinson, Smart, Hanna, *et al.* [61] discusses the importance of documenting datasets, to achieve accountability through dataset transparency. Another methodology is presented by Naja, Markovic, Edwards, *et al.* [47], which calls for the definition of specific accountability requirements and provides a high-level framework to incorporate these requirements into practice. Another approach is taken by Raji, Smart, White, *et al.* [46], who propose an end-to-end framework for internal auditing that aims to ensure that the model is subject to scrutiny at every stage. Further, they advocate for documentation at each stage [46]. Additionally, the concept of AI harm and mitigation is beginning to be incorporated into system approaches. For example, Khlaaf, Mishkin, Achiam, *et al.* [62] outlines requirements for a hazard-aware approach to LLMs.

Our work does not strictly build upon these concepts, as it does not take a systems approach. Instead, it proposes how one element of the development process, documentation of model testing, could be configured to properly ensure models are tested and documented prior to deployment. It is worth noting that this differs from previous work by providing a *specific type of documentation.* It is not as extensive as some of the frameworks presented here but contributes through its process specificity.

2.1.3 Documentation and Internal Organizational Procedures

Documentation and internal organizational procedures are one way of operationalizing responsible AI. However, given the unique nature of AI systems, specific guidelines to produce this documentation are required. Königstorfer and Thalmann establish 5 key goals that AI documentation must have that go beyond the requirements of standard documentation practice in software engineering: description of the application domain, description of the training data set, descriptions of design decisions, understandable documentation for knowledgeable third parties, and achieving a balance between the benefits and efforts of documenting artificial intelligence [50]. However, it is beyond the scope of their work to address the concrete steps needed to materialize this type of documentation. In another piece, Königstorfer and Thalmann establish that the level of documentation required should be correlated with the perceived level of risk of a particular use case [48].

Brundage, Avin, Wang, *et al.* [49], provide a series of institutional, software, and hardware recommendations for making verifiable claims and establishing trustworthy AI development, one such mechanism being auditing trails. Specifically, they illustrate how these audit trails can be established to demonstrate the system's properties and impacts [49]. While an overview of algorithmic auditing is beyond the scope of this thesis, Brundage, Avin, Wang, *et al.* [49] suggest that "standards setting bodies should work with academia and industry to develop audit trail requirements for safety-critical applications of AI systems" [49, p. 24]. These documentation standards provide some guidance on how AI documentation might differ from the standard documentation practices of other software systems. Moving beyond these documentation guidelines, another key element is what is covered by these procedures – specifically in the space of testing and verification/validation.

2.1.4 Verification and Testing

Perhaps one of the most common documentation procedures in software engineering is based upon testing and verification, hence the type of testing that is being conducted is relevant to the documentation that will be produced. In addition to their documentation recommendations, Königstorfer and Thalmann establish the importance of testing based on the risk level of the model's regulated use case in order to establish accountable systems [48]. Beyond this recommendation, testing can be very broad and mean many different things. Looking specifically at the capacity of testing to capture the alignment of AI with certain values, Brown et al. have theoretically shown that this can be achieved through verification [63]. They prove that it is possible to "verify exact and approximate alignment across an infinite set of test environments" [63, p.1105].

In addition to this theoretical work, several other authors have developed innovative testing mechanisms that can be used, each with its own relevant use case. Perez et al. apply a method known as 'red-teaming' to NLP models, specifically using one language model to generate test cases for another language model with the explicit purpose of detecting vulnerabilities. This serves as a testing tool to help limit the problematic and undesirable behaviour of models [64]. Wotawa takes a different approach, suggesting that traditional testing mechanisms can be adapted to AI models, specifically focusing on 'black-box' testing mechanisms [65]. Finally, Borg et al. (in the context of AI safety) discuss the high-risk use case of the automotive industry and safety-conscious testing practices. They propose cage architectures (a form of design that continuously monitors sensor input) and simulated system test cases [66]. These procedures and mechanisms are only a sample of the testing practices that are emerging; there are a variety of ways to test and verify an AI system. The type of testing used will depend on the AI system and what behaviour the developers wish to see; hence no testing strategy will be optimal in every case.

2.1.5 AI Governance

Current Situation

The landscape of AI is largely self-regulated. Notably, in the context of AI development, self-governance primarily relies upon public relations (PR) as an enforcement mechanism. Specifically, media exposure or 'bad'/negative PR or publicity is what encourages companies and firms to follow AI self-governance mechanisms [55]. While this forms a relatively weak incentive [67], it can be shown that AI self-governance can be effective. This is discussed by Roski et al. in [68], which suggests that by allowing and encouraging self-governance, practices will emerge that could lead to a more effective certification regime. This is coupled with the survey of AI in industry completed by Rakova et al., which demonstrates that individuals in companies will often go above and beyond their job title to institute responsible AI in practice [55]. Further, they also document that emerging practices are beginning to adapt the current structure of the AI workflow to include responsible AI practices throughout the development and deployment chain [55]. This signifies that the current status of AI is open to the adoption and integration of responsible AI practice.

AI and Legal Liability

The discussion surrounding legal liability and accountability of AI models is not new, but many of the documented issues within the AI liability space still exist today. One of the major hurdles to establishing liability for AI is the fact that it lacks legal personality: it is not a legal subject and thus lacks rights and obligations/responsibilities in the eyes of the law. Instead, AI is currently treated as an object of law, meaning it can feel the effects of the law but is not able to be held directly accountable [69], [70]. However, this does not mean that AI is completely free of liability, as the actors that surround the AI system can still be liable.

In terms of criminal liability, Kingston considers the use case of autonomous vehicles or self-driving cars [70]. In this example, Kingston differentiates between actus reus (the criminal action) and mens rea (the intent to commit this action) in order to distinguish between liability that is incurred when intent is considered and strict liability that only relies on action. Furthermore, in the case where criminal liability requires mens rea, there are two different standards: knowing an action would cause harm, and situations where 'a reasonable person' should have known harm could occur. While it may be impossible to prove that the developers knew that a criminal act did or would occur, it is entirely possible that the developers could have known about the possibility of the criminal act. Thus, criminal liability (in some cases) can be located at the level of the system developer [70]. Further, according to Mykytyn et al., as quoted in [70] by Kingston, a certification or expert qualification is important to ensure this accountability is translated to practice. Kingston and Chaudhary both describe this as a form of a natural-probable-consequence model, where AI system developers or end users can be held liable for what the system does [69], [70]. This is contrasted by the direct liability regime, which assigns both mens rea and actus reus to the model, and the perpetrator-via-another regime, which views the model as an innocent actor being 'controlled' by the development team [69], [70]. While both frameworks are interesting, direct liability can struggle, as we must determine the mens rea (intent) of the model directly.

Furthermore, the perpetrator-via-another framework assumes that the developer made the choice to cause harm instead of harm arising as a foreseeable consequence of their actions, which while useful will not address the type of harms we wish to regulate [69], [70].

In civil liability, per Tuthill, quoted by Chaudhary in [69], when software is defective, legislation tends to go through the tort of negligence. This provides a model for accessing liability through the lens of the civil system. As documented by Gerstner, there are 3 conditions that need to be shown for a negligence claim to be granted: the defendant had a duty of care, there was a breach of that duty of care, and there was an injury to the plaintiff [71]. In AI, the duty of care is placed at the software developer, however, the level of care is debatable. Additionally, per Gerstner, quoted by Chaudhary in [69], it is suggested that the breach in the duty of care can arise for multiple reasons, many of which are based upon the malpractice of the developer (for example, a developer's failure to detect errors in program features and functions). Finally, it has to be shown that the system decided something and did not simply recommend something [69]. Ultimately, while there are multiple ways of positioning liability within both the civil and criminal framework, in most proposed frameworks, culpability lies with a decidedly human component. Still, the question remains on how this human element can be defined or determined.

Explainability and Transparency to Establish Liability

At its core, liability is a way of determining accountability. Giuffrida [42] suggests that an AI system could result in harm to individuals or damage to property without a clear case for responsibility. To address these issues, Giuffrida [42] suggests that it is important to ensure that AI systems are designed to comply with legal requirements and ethical principles. This includes ensuring that AI systems are transparent and explainable so that their decision-making processes can be understood and scrutinized. This has caused a push for explainability of AI models and suggests that if a particular decision made by an AI model is understood, then it is possible to hold the model to account through legal scruitny [42], [72].

Collaborative Governance and Penalty Defaults

Collaborative governance is a form of governance that is particularly relevant to complex AI systems. This regulatory mechanism is neither entirely direct (a top-down command and control regime) nor self-governed and is instead founded upon private-public partnerships [73]. Formally, collaborative governance (a series of strategies, often referred to as a 'toolkit') "deploys private-public partnerships towards public governance goals [and] should not be confused with self-regulation, though it may include or even rely in substantial part on private governance" [73, p. 1564].

One particularly powerful collaborative governance tool is a penalty default regime. Penalty defaults, in the context of AI, are described by Yew and Hadfield-Menell as rules/-mandates that only occur within the context of an incomplete contract, only activating when certain conditions are not met. As such, when organizations fall into a 'default' state, they are subject to a penalty. The authors continue to argue that this regulatory tactic enables private research into how the harm that AI systems do can be mitigated [67]. Furthermore, penalty defaults, when established around AI systems, do not force the adoption of one particular methodology or standard, meaning they remain adaptable to uncertain and evolving systems. Overall, this kind of tool could provide a potential path to regulating complex models.

2.2 Documentation as System Understanding

This solution's core idea is to use alignment with human judgement as a grounding for liability. This thesis suggests that liability should be centred on the principal (the deployer) when the AI model (the agent) causes harm. This allows us to hold the model liable, based on the judgement by deployers at design time. By their judgement, this thesis refers to the actions taken by the deployer to understand the system at the testing and evaluation phase *before* they release the model. Therefore, in liability, we can use alignment to determine if the principal enacted 'adequate' precautions, exercised 'good' judgment, and took 'reasonable' steps to understand system behaviour. Note that the definitions of adequate, good, and reasonable are all left vague under this scheme.

This represents a shift in the location of liability. By shifting from an algorithm to a human, we do not need specialized legal rules. When a person is responsible for harm, we currently have **existing** and **effective** practices that serve to establish liability. Think of the hypothetical example of a person driving a car, the driver (or principal) gets into an accident. We have a legal regime to handle this. For example, if a person was behaving in an unsafe manner (e.g., driving over the speed limit by a significant amount) they would face more consequences than if they simply hit a patch of black ice while driving responsibly. Using this example, the question for liability in AI is whether or not the deployer was justified and made an understandable decision to release the model.

Thus, we need to understand *why* and *how* a deployer made their decision to release. Once this initial understanding is achieved, the many robust legal frameworks that exist can be used, as there are many laws that regulate how people make decisions. This paper argues that this information can be achieved by using **documentation at deployment as a proxy for the principle**. This approach aims to hold AI models liable by designing an incentive structure based on the alignment justification found in the documentation. This type of liability is a penalty default and operates by levying a penalty against the deployer when certain conditions are *not* met.

2.2.1 What is Documentation?

The type of documentation recommended here is not traditional software documentation. Traditional software documentation will not be enough [50].

At its core documentation "is intended to provide stakeholders with useful knowledge about the system and related processes" [74, p. 1199]. Documentation need not be of one specific format and can instead be a tool for stakeholder communication. Thus, this thesis is not requesting the traditional software documentation, but rather a new kind of written communication that can speak to deployment practices and choices. As will be discussed below, this work presents a series of minimum requirements or recommendations for what this documentation must have.

2.2.2 Who is a System Deployer?

This thesis chooses to locate our documentation at the system deployer level. It is beyond our scope to engage in the discussion of what exactly constitutes a deployer. Instead, this thesis chooses the term 'deployer' to refer to the individual or organization within the design chain who bears primary responsibility for model use.

I acknowledge that this is not a complete description. However, such a definition is beyond the scope of this work. I would instead refer the conversation to the distinction of 'frontend' and 'backend' operators in the European Parliament Proposal for a Regulation on AI Liability [75]. Furthermore, the discussion on which operator is responsible (for example, Wendehorst [44] argues that it is dependent on system control and autonomy) in any context is beyond this work. This thesis chooses 'deployer' to refer to where the locus of liability will rest once it emerges.

2.2.3 What Documentation Gap Does This Fill In?

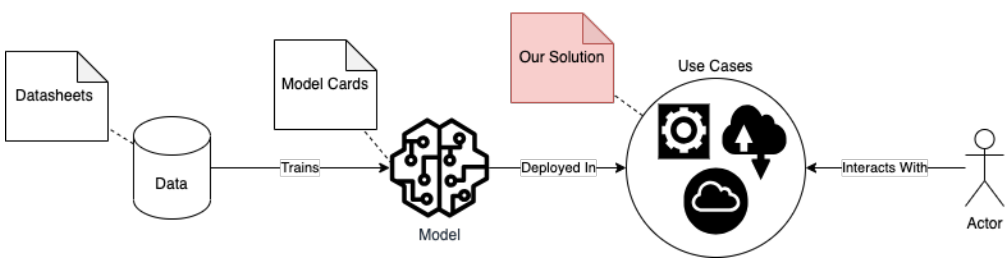

Figure 2.2: Where documentation could serve to found liability

There are several stages in development where documentation and understanding of the model is relevant.

This thesis makes note of two particular documentation 'touchpoints' that occur at different stages in the development lifecycle:

(1) the **data** level. Gebru, Morgenstern, Vecchione, *et al.* [76] outline datasheets for datasets, as a way of promoting accountability and managing the relationship between data producers and consumers.

(2) the **model** level. Mitchell, Wu, Zaldivar, *et al.* [77] develop model cards that are designed to promote transparency on the model's abilities and intended use. This form of documentation is designed to provide insight into the use and qualities of a particular model.

Our approach complements these proposals and emphasizes accountability with respect to deployment decisions. Instead of documenting the model itself, this approach deals with the intent, process, and decisions of those who choose to release AI models for use or the 'deployer'. The motivation behind this approach is to facilitate the gradual development of a documentation standard at the point of deployment, where deployers:

1. commit to a particular deployment context;
2. explicitly state claims about the performance of their systems;
3. compare these with the actual performance of the model;
4. justify why such claims and performance are sufficient to allow deployment.

This thesis suggests that this documentation will help us understand two key things about the system: (1) the intent of system deployers to reduce harm and (2) the conditions for a safe and effective deployment. This documentation, centred on the deployment point, can then be used to support algorithmic accountability through liability.

2.2.4 Why Document Use Cases?

This is a new kind of documentation call, one that is centred on the deployment context and use. I am specifically proposing documenting a particular use case and deployment decision for two reasons: (1) it will incentivize deployers to be aware of the societal use of their models, and (2) it will allow proportional testing to the deployment context.

Firstly, by forcing deployers to commit to a use case, and document it, we force deployers to think about the societal impact of their model. There is no longer a divorce between an AI model and its use. In fact, this would address calls for consideration of different domains, particularly, forcing a different process in domains that are classified as high-risk [78].

Secondly, this kind of approach recognizes that not all deployment contexts are inherently the same. For example, Solaiman [79] points out that "since a system cannot be fully safe or unbiased for all groups of peoples and there is no clear standard for when a system is safe for broad public release, further discourse across all affected parties is needed" [p. 118]. This strategy addresses this call, as it is sensitive to these differences. It also draws upon the same kind of logic common to Gebru, Morgenstern, Vecchione, *et al.* [76] and Mitchell, Wu, Zaldivar, *et al.* [77] in the form of model cards and datasheets, a strategy deemed successful by multiple authors [79]–[81].

The following section will differentiate our method from audits and other post-hoc explainability methods. Next, this thesis will explain how deployment documentation that meets our criteria could support liability regulation.

2.3 Post-Hoc Artifacts: A Note on Explainability and Audits

2.3.1 Avoiding the Explainability Problem

This work is decidedly different than other ways of determining liability within the design process of AI models. While other methodologies recognize that "the ex-post explanation is essential when we are drawing an analogy between the conduct of an AI and a person" [72, p.141], they center this explanation on the results of explainability tools. Taking a different approach, Percy, Dragicevic, Sarkar, *et al.* [22] discusses how transparency (and eventually an accreditation system) could emerge from establishing clear explainability and documentation standards. However, explainable AI is not monolithic, with no one agreed upon methodology [82], and some even argue that explainable AI is dead [83]. Thus, regardless we are a long way from a concrete, definite approach to explaining AI decision-making. Instead, our proposal relies upon the generation of an artifact at *design time*, that can provide the ex-post context to create a conduct analogy between a person and an AI model.

2.3.2 This is Not a Compete Audit Framework

Raji, Smart, White, *et al.* [46] propose an end-to-end internal auditing framework. This process is designed as an accountability mechanism that an organization can adopt to ensure effective, fair, and safe systems. This kind of framework provides a systems approach to performing a post-hoc analysis or process for a system.

To address the creation of such an audit, Brundage, Avin, Wang, *et al.* [49] suggest that "standards setting bodies should work with academia and industry to develop audit trail requirements for safety-critical applications of AI systems" [49, p. 24].

Our work is not an audit, as it does not take a systems approach. It is not a post-hoc process (an audit) but rather an artifact generated at a design time to *be used as a part of a post-hoc process*. Instead, it proposes how one aspect of the lifecycle, could be configured to properly ensure models are tested and documented prior to deployment. It is worth noting that this differs from previous work by providing a *specific type of documentation*. It is not as extensive as some of the frameworks presented here but contributes through its process specificity. However, this thesis does suggest that a similar collaboration, as suggested by Brundage, Avin, Wang, *et al.* [49], should be encouraged to allow the emergence of a documentation standard. In addition, I propose the additional legal element of liability. Further, I could see this documentation standard being incorporated into an audit framework like what Raji, Smart, White, *et al.* [46] propose.

2.4 From Documentation to Liability

2.4.1 How does Algorithmic Liability Work?

At its core, liability is a way of determining accountability. Giuffrida [42] suggests that an AI system could result in harm to individuals or damage to property without a clear case for responsibility. To address these issues, Giuffrida [42] suggests that it is important to ensure that AI systems are designed to comply with legal requirements and ethical principles. This includes ensuring that AI systems are transparent and explainable so that their decision-making processes can be understood and scrutinized. This has caused a push for the explainability of AI models and suggests that if a particular decision made by an AI model is understood, then it is possible to hold the model to account through legal scrutiny [42], [72].

2.4.2 Documentation as a Tool For Liability

However, I suggest that explainability tools are not the answer, due to the explainability problem highlighted above. Instead, this work proposes that another artifact can serve to provide this required explanation: the documentation at the time of deployment that meets our proposed criteria.

2.4.3 What Could Liability Based on Documentation Look Like?

This approach aims to hold AI models accountable through liability by designing an incentive structure based on penalties. Essentially, penalties would be enacted when certain conditions are *not* met.

This approach to liability is broken into two steps or 'tiers'. Firstly, liability will depend on whether there is *sufficient* documentation. In other words, the *amount and quality* of documentation must meet a minimum baseline, secondly, whether or not the decision to deploy the system is *justifiable*. For this approach to work, failure to meet the first step must result in a 'higher' level of liability or penalty than failure to meet the second. This approach forms a 'tiered' penalty default or liability structure.

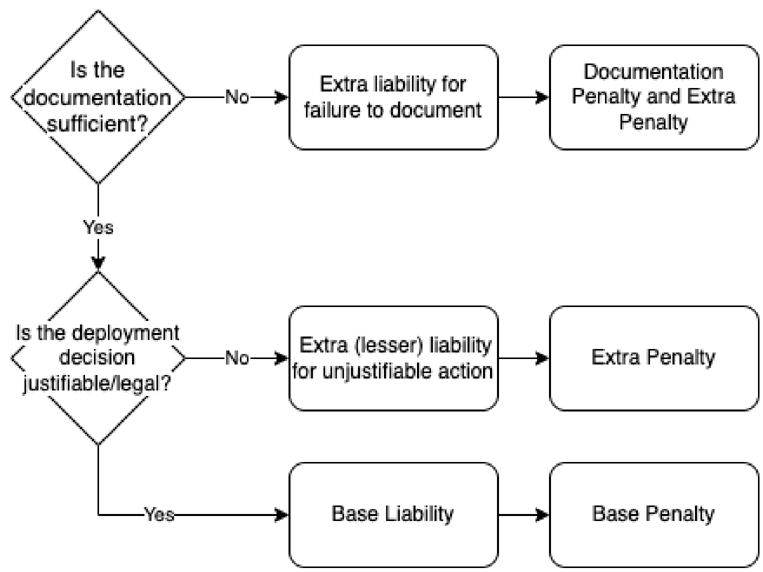

Figure 2.3: Liability explained: The tiered default structure

This tiered default structure would center liability on deployers and their intent by using the documentation they produce as a foundation. Essentially, this framework is composed of two default rules, of which the second is only relevant if the first default rule is not activated. In the first rule, we put the onus on organizations to produce sufficient documentation. If they fail to produce this documentation, they are in a default state with the highest liability and penalty. If the documentation they produce is sufficient, the second default rule is considered. This rule goes beyond the mere fact of documentation and looks at its content to determine if the deployment decision was justified from the perspective of intent. If it can be shown that the developer could not reasonably foresee the AI system's decision based upon the documentation, then the organization is in a base state of liability and penalty. Otherwise, they enter a default state, where they are subject to extra liability and penalty, but do not receive a penalty for failure to complete sufficient documentation. The combination of these two rules serve to create a framework where failure to participate in this software engineering practice is punished more severely than making an error in deployment. Hence, this framework is effective to first incentivize compliance, and then encourage proper practice.

Note that the exact meaning of "sufficient" and "justifiable/legal" is beyond the scope of this proposal and that these definitions are deliberately left vague. Our proposal relies upon the existence of a strong, independent judiciary system that would, through practice, develop these definitions via a series of decisions. Note that I have only considered adopting this system within a US context, but, future research should address *how* this could be instituted in the US and globally. However, I propose the foundations of such a documentation procedure and expect the legal system to reference other existing documentation standards, as a way of slowly developing this accountability mechanism.

2.4.4 Major Implications of Adopting This System

In practice, this system may encounter a few challenges that will require careful mitigation. Firstly, this system relies upon detecting the harm that an AI model causes and on reporting this harm. Hence, this model cannot exist in isolation and must be surrounded by monitoring mechanisms to achieve its true potential. A simple monitoring mechanism could be the public, meaning that the public needs to be informed of this system and know how to report harm. Secondly, this system supposes an independent and strong judiciary. However, even with a strong judiciary, judges and legal professionals may not fully understand the nature of the evaluation we are proposing in this framework. Furthermore, in a common law system (one based on precedent), a single ruling could overturn the effectiveness of the model. A potential mitigation for this concern could be the deliberate incorporation of AI specialists as expert witnesses into the legal system. Just as we have forensic specialists in physical matter today, we could eventually have similar experts in the field of AI harm.

This system has several positive implications for the development of more responsible AI. Firstly, this system will generate internal pressure to comply by forcing organizations to include this proposed documentation procedure in their development practice. Furthermore, this system will not mandate a specific type of analysis, allowing organizations to choose the method that works best for them. This will prioritize the desired outcome of this process, namely ensuring that an organization is taking concrete steps to produce more ethical AI, without specifying definitive methods. Secondly, this framework is based upon intention that drives practice, meaning that the actions organizations take will not merely be to ensure compliance with a particular standard. Thirdly, this framework takes advantage of strategic ambiguity, selectively remaining undefined or vague, to remain adaptable and allow conceptual definitions to emerge over time. In effect, this proposal is not tied to a particular technical requirement, meaning that it will be able to adapt to rapid technical evolution.

2.5 Summary of Chapter Findings

Overall this chapter has argued that documentation at the level of deployment is an effective and possible policy strategy. We can use documentation as a proxy for system understanding and form the basis of policy through it. Thus, an understanding of system behaviour can serve as a valid option to address the holes in a risk framework. This chapter contributes to the overall argument of this thesis. Specifically, it demonstrates the feasibility of proxying system understanding from both a civil liability perspective, and as an internal process (documentation).

Chapter 3

Understanding as a Procedure

This second section examines what kind of effect on understanding that documenting a system could have. I examine if documentation, or indeed any kind of post-hoc evaluation is truly a proxy for AI system understanding. Furthermore, I conduct a user study that examines how we can influence a human's ability to understand an AI system. I argue that system understanding as a basis for policy (as described in the previous chapter) only makes sense if there is some kind of intervention that can be employed to modify a person's ability to understand system behaviour.

AI is becoming more agent-like every day [84]. This is particularly true as we enter the age of generative AI. With these new uses of AI, we also live in a time of not knowing what the potential of these technologies is and that these technologies carry with them inherent risks [81], [85], [86].

Even before the explosion of generative AI, it was not practical or necessary to fully understand every possible input/response of AI, as AI is inherently unpredictable [87]. However, that does not mean that we can not understand the overall system behaviour. In fact, this is has been attempted both through auditing and explanations by Zhang, Cho, and Vasarhelyi [88] and Agarwal, Krishna, Saxena, *et al.* [89]. In other words, understanding system behaviour is a task with multiple different solutions. Furthermore, there is a rich literature on human-AI interaction, which contributes to this understanding of how humans interact with AI. Several different dimensions exist, there are those that look at individual effects in decision making [90], [91], human-AI systems or human-in-the-loop [45], [92] and also emerging trends [93], [94].

This creates what could be a paradox on the surface: the unpredictable nature of individual inputs and responses with the need to understand the overarching system behaviour. However, we are capable of understanding system behaviour. In AI, the existence of post-hoc explanations can enhance our understanding of system behaviour or at the very least help us make sense of behaviour that "makes no sense" [95]. The user study proposed in this section offers an alternative to understanding system behaviour. This proposal is based on the emerging agency of AI models, and their ability seem more "human-like" [96]

I explore the dynamics of human-AI interaction and the role of the human's ability to predict behaviour. This work does not accept that humans have an unmovable baseline when it comes to understanding the system. I argue that through repeated and structured interaction with model inputs and responses, it is possible to shift this baseline forward.

I propose a version of a structured interaction with model inputs and responses. This thesis does not comment on whether this is “the best” kind of interaction, but it does assert that it is *a way to move the baseline of system understanding forward*. As a methodology, I use a modified version of Open Coding for Machine Learning that was designed and implemented by Price [97]. This software is an application of Grounded Theory [98] and allows for the structured analysis of model input/response pairs that show behaviour.

Using this approach to structured analysis, I conduct a user study. In this thesis, I compare the performance of those who complete this structured analysis to those who do not complete it (and are instead simply shown the input/response pairs in no particular order). The user study analyzes the generative task of sentence simplification.

Following this process differentiation, participants were subsequently asked to predict the AI model behaviour. There was no general trend in the performance of one group over another. However, in cases where the simplification of the sentence made a large change to the sentence, those who completed the structured interaction were able to better predict the model response.

Ultimately, there is evidence that by engaging in this structured interaction of input and response pairs, when there is a higher complexity to the task (or more ‘options’ for task completion) humans will have a better ability to predict the AI system behaviour. This evidence is the initial indication of such a trend and indicates that it is possible to improve human understanding of system behaviour through a post-hoc intervention.

3.1 Related Works

3.1.1 Human-AI Interaction

There is a large body of research that investigates human-AI interaction [99]–[101], as well as human-in-the-loop systems [45], [91], [92] and human-centered AI [102]. Of particular relevance to this study is the literature on how users can be trained to better understand a model. This is seen in some of the work done by Amirizaniani, Yao, Lavergne, *et al.* [103] and Shankar, Zamfirescu-Pereira, Hartmann, *et al.* [104] in their human-in-the-loop frameworks that specifically look at how to incorporate humans in the age of LLMs. Their work goes beyond the work of Bandi, Adapa, and Kuchi [105] that highlights evaluation metrics in the age of generative AI and instead transitions to the incorporation of humans in meaningful ways.

In addition to the relevance of generative AI evaluations and LLMs, there is also substantial work that investigates human-robot interaction, particularly the relevance of cross-training. Nikolaidis and Shah [106] show the effectiveness of cross-training, which is a kind of structured human-robot interaction that has the human and robot switch roles to learn how to collaborate. This kind of structured learning had better performance rates than standard reinforcement learning, which has also been re-established by Nikolaidis, Lasota, Ramakrishnan, *et al.* [107] and Pan, Zhao, Pu, *et al.* [108].

Furthermore, Xu, Dainoff, Ge, *et al.* [94] calls for the use of behavioural science methodologies as a way of understanding and testing AI models. Traditional software testing will not lead to an understanding of the system’s behaviour. In addition, there are specific human considerations and hesitancies that humans inherently bring to the workflow. For example, Berkel, Opie, Ahmad, *et al.* [90] found in their study of AI-assisted medical imaging, that humans will pause longer to consider a false-positive recommendation rather than a true-positive recommendation.

Our work builds upon the works here by looking at how system understanding and developing an interaction paradigm could influence human understanding of the model performance. This is different from the previous work that has focused on improving the human-AI team and provides an answer to the calls for system understanding.

3.1.2 Predictable AI

Several authors have also touched upon the concept of predictable AI. Zhou, Moreno-Casares, Martínez-Plumed, *et al.* [109] speaks to the establishment of the field of predictable AI, a new approach to understanding AI. They argue that achieving predictable AI is crucial to trust, safety, and alignment, and they summarize this field through emerging calls for predictable AI. This differentiates from other work by Ganguli, Hernandez, Lovitt, *et al.* [110] that suggests that generative AI is inherently unpredictable in its capabilities. However, this claim comes from the property of AI observed by Rahwan, Cebrian, Obradovich, *et al.* [87] that it is neither practical nor necessary to observe all AI behaviour. Instead, predictable AI applies a tried and tested software engineering (outlined by Fink and Bishop [111]) technique to AI (as described by Ribeiro, Wu, Guestrin, *et al.* [112]). These methodologies could serve to make AI more predictable in a landscape that has also had uncertain capabilities. On the instance level, human prediction of AI behaviour could be useful Middleton, Letouzé, Hossaini, *et al.* [113]. Kandul, Micheli, Beck, *et al.* [114], Bansal, Nushi, Kamar, *et al.* [115], and Nushi, Kamar, and Horvitz [116] have all examined the ability of humans to predict AI behaviour in various contexts. Furthermore, Hase and Bansal [117] shows that post-hoc interaction with models (in the form of explainability tools) can impact how well a person can predict AI model behaviour.

Our work builds upon the ability of humans to predict AI. Instead of looking at the abilities of humans to predict AI responses, I analyze what kind of process/learning humans must go through in order to predict AI systems.

3.1.3 Human Mimicry and Learning Through Practice

This research draws upon the human ability to learn through observation and mimicry. Artino [118] summarizes key authors (including Ross and Bandura) in psychology for their work in observational learning. One experiment referenced is the famous Bobo Doll experiment, which demonstrated that if children saw an adult treating a doll violently, they would copy this behaviour. Modern research has further described and attempted to quantify this ability, for example, it occurs without intent Chartrand and Baaren [119] or as a way of deepening social interactions Gueguen, Jacob, and Martin [120]. Our research builds upon the human capability to learn through observation and subsequently copy behaviour by applying this principle to AI models. I want to establish that by observing the behaviour of AI models, humans can then copy this behaviour.

This mimicry is perhaps even more prevalent given advances in generative AI. In particular, humans currently personify AI [85], and the degree to which they feel empathy towards an AI influences their interaction with it [121]. This anthropomorphization also plays a key role in human acceptance of an AI system [96]. Thus, as the capabilities of generative AI evolve, the application of human behaviour methodologies will become increasingly more relevant.

3.2 User Study Design

In order to establish our proposed hypothesis, I conducted a user study. This section will discuss the study design, and the task setup before the results and analysis are shown.

As outlined at the start of this chapter, I am interested in examining the interaction of people with AI model outputs. I hope to contribute to the understanding of how people build a mental model of AI models. I choose to look at how individuals approach examining a series of input/response pairs of a model, given both the model goals and a particular use case. Thus, I am interested in whether the format in which a person interacts with a series of input/response pairs influences their ability to understand, in a broader context, how the system will behave. First, I present our user survey. Then, I formally state the hypotheses.

As a general overview, I constructed a user survey to follow three steps. In the first step, individuals were told they would be evaluating an AI model based on certain deployment criteria. Then, individuals were shown a series of input/response pairs of that AI model. Individuals were randomly assigned to either the control or the experimental group at this stage. In the control group, participants underwent an unstructured examination of these pairs and in the experimental group they underwent a structured interaction. In the second step, they were shown a model input and asked to predict the model response. They also compared their response to the actual model response. In step three, the participants were shown an input and two different responses (one from the AI model under evaluation and one from another AI model). They were then asked to select the response from the AI model under evaluation. Before each step is discussed in detail, there are two overarching points of the study design.

3.2.1 Cross-Cutting Concerns in Study Design

The Task: Sentence Simplification

Firstly, the AI model under evaluation, as well as the model desiderata and the deployment context for the task. The framing that was selected was that of sentence simplification. This task was selected as (1) it is a well-documented and understood task, (2) it can use LLMs that are correctly prompt engineered, and (3) it is a generative task [122]. In particular, I am interested in how generative model behaviour can be understood and this task satisfied all requirements. It also uses LLMs, models that are becoming increasingly prevalent in society to AI tasks [123], [124].

In order to establish the AI model under examination (and the comparison model for step three), I examined models that were benchmarked on sentence simplification by Kew, Chi, Vásquez-Rodríguez, *et al.* [125]. From this paper, I selected the top-performing prompt, as well as the top-performing model on the ASSET dataset to be our AI model under analysis (GPT-3.5-Turbo). As a comparison, I selected the Flan-T5-large model, an open-source model that keeps pace with OpenAI's GPT-3.5-Turbo when run on the same prompt [125]. I selected examples from the ASSET dataset, a curated dataset published by Alva-Manchego, Martin, Bordes, *et al.* [126] that is designed for sentence simplification. Note that Kew, Chi, Vásquez-Rodríguez, *et al.* [125] published a code base, and from that base, I selected several examples that were used in the benchmarking paper, and specifically run on the specified prompt using the ASSET dataset. These examples are pre-generated by Kew, Chi, Vásquez-Rodríguez, *et al.* [125] and taken from the code base. This pre-generation process, as well as the prompt used, informed the model desiderata.

From this setup, I provided the following desiderata to participants:

1. Simpler sentences
2. Sentences that are understandable by non-native English speakers
3. Meaning preservation

The Experimental Condition: Unstructured/Structured Interaction

Secondly, the definition of an unstructured or structured interaction during the first step of the experiment. An unstructured interaction with the model input and response pairs is simply the presentation of each pair in a table (and a downloadable CSV). There is no guided interaction, the participant is only presented with the data and asked to read and analyze it any way they see fit.

This user study employs a specific kind of structured interaction with model input/response pairs known as Open Coding For Machine Learning. This evaluation task is based upon the code base developed by Price [97], which is an application of the grounded theory methodology developed by Charmaz [98] for machine learning datasets. This structured interaction is hosted online at http://petrichor.csail.mit.edu/ (note, you can use any username to access the site). It uses two stages from the approach taken by Price [97] and applies them to input/response pairs. In the first stage, users are asked to annotate the input/response pairs based on the model desiderata in any way they see fit. In the second stage, they then group these annotations, with limited guidance, into categories that make sense. Essentially, this stage allows for the aggregation of annotations into broader categories, by making individuals complete a second pass.

This process differentiation is the key experimental condition of this study. This study will track the effects of a structured model input/response pair interaction on a person's ability to predict model behaviour. Additionally, this study will evaluate how this process differentiation affects user trust.

3.2.2 The Study Workflow

This survey was administered through the Qualtrics platform. Study participants were recruited using the Prolific administration service. Subjects were recruited based on residing in the United States, being fluent in English (and having it as their primary language), as well as completing some post-secondary education. Subjects were compensated for their participation in the survey, and bonus payments were offered to top performers for each question answered correctly in step three. Table 3.1 summarizes the makeup of both the control and experimental groups.

Table 3.1: Age, gender, and number of responses for the control and experimental groups

	Control	Experimental
Average Age	40.0	37.2
% Women	52.94%	33.33%
% Men	47.06%	66.67%
Number of Responses	51	48

Overall, the participants were randomly recruited and represented the general population. Participants were selected to ensure at least a college reading level and primary use of English due to the nature of the sentence simplification task.

Step One

In this step, a different interface is shown to the control and the experimental groups in the survey. Specifically, the experimental group is shown the modified Open Coding for Machine Learning process that is described above. This process lists out 50 pairs, that they are then asked to annotate and group. Note that both the groupings and annotations are stored in a file and are accessible for analysis. Appendix B contains images of the Open Coding for Machine Learning structured analysis.

The control group is given a table of input/response pairs within the survey, as well as a link to download a CSV version of the table if they so choose. The only instruction given was to read through the pairs and “examine these pairs in any way you see fit.”

Both groups were given 50 examples from the model to either annotate or examine. In both cases, attention checks were conducted. In the control case, this required the participants to answer questions about the contents of the model input/response pairs. In the experimental group, the files generated by each Open Coding for Machine Learning completed workflow were examined for completeness and validity in the annotations (i.e. not all the same, not random, and relevant).

Step Two

In the second step, all groups were shown 20 model prompts. For the full list of input/response test prompts, please see Appendix C. Participants were asked to predict the model response for each input.

They were then shown their response and the model response. Then, they were asked two follow-up questions: (1) if *their* response met the desiderata of what the model was supposed to do and (2) if they considered their response "similar enough" to the model

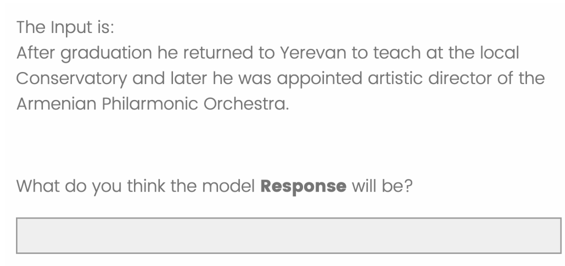

Figure 3.1: An example input that the participant must predict the model response to

response. I asked that they answer these questions in a binary Yes/No fashion and then provide a justification for each of their answers. Figure 3.2 shows this interface.

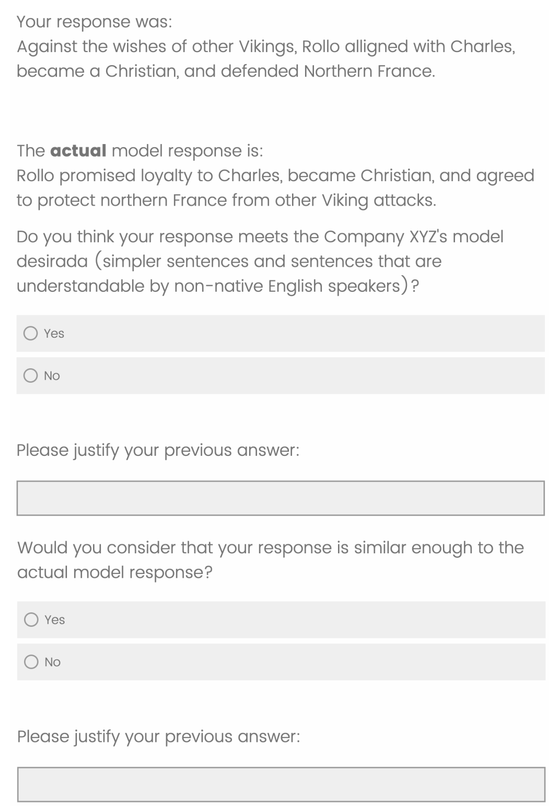

Figure 3.2: The follow-up questions that the participant is asked about the predictions

This step was designed to evaluate the participants' ability to predict how the model would behave in a free-form way.

Step Three

In step three, all groups were shown 20 model inputs in a random order. For each input, the user was shown a response from the model under analysis (GPT-3.5-Turbo) and the comparison model (Flan-T5-large) that were both run on the same prompts. For a full list of the inputs, model under analysis responses, and comparison model responses, please see Appendix C.

I asked the participants to identify which prompt was the response from the model under analysis. To incentivize performance, individuals received a bonus payment for each correct answer.

Final Questions

At the end of the survey, participants were asked summary questions about the task. Specifically, participants were asked about their general opinion of the survey results. I also chose to evaluate model trust at this time. I used the scale proposed by Hoffman, Mueller, Klein, *et al.* [127], which was an adaptation of the metrics proposed by Cahour and Forzy [128] and Adams, Bruyn, and Houde [129] to measure the trust in an AI model. This scale was selected for its application to autonomous systems.

This scale, as presented in Hoffman, Mueller, Klein, *et al.* [127] is displayed below:

1. What is your confidence in the AI model? Do you have a feeling of trust in it?
2. Are the actions of the AI model predictable?
3. Is the AI model reliable? Do you think it is safe?
4. Is the AI model efficient at what it does?

3.3 Results

The hypothesis behind this research is that people will be able to better predict the behaviour of an AI model if they engage in a structured interaction. I expected this to manifest in the better performance of participants who completed the Open Coding for Machine Learning interaction on steps two and three of the survey. Specifically, I expected that they (the experimental group) would have respond with a simplified sentence that was more similar to the model than those that completed the unstructured interaction (the control group). Furthermore, I expected the experimental group to be able to more accurately select the response from the model under analysis when shown it and a response from the comparison model. Below is a statement of all our hypotheses.

H_1: The participants in the experimental group will be able to better predict the model response than the control group.

H_2: The participants in the experimental group will be able to better match the level of simplicity of the model response than the control group.

H_3: The participants in the experimental group will have a better understanding of whether their response is similar to the model response than the control group.

H_4: The participants in the experimental group will be able to better identify the model response than the control group.

H_5: The participants in the experimental group will have more trust in the model than the control group.

Ultimately, there were no general trends across all questions. As is presented in the rest of this section, I saw some significant results across certain questions.

3.3.1 Data Cleaning and Statistical Methods

In order to process the results of this experiment, the textual responses of each individual were evaluated. Select responses were clear errors, and while the participants completed the survey in full, they made a mistake on a particular question. For example, when asked what the participant thought the model response would be, they responded with "None". Such answers were removed before the data was processed, and the analysis was conducted on all valid responses for each question. No more than 2 responses were excluded for any given question due to lack of validity.

To quantify the similarity between the textual response of the person in step two and the actual model response, I relied upon the cosine similarity between the vector embeddings of both the participant's response and the model's response. In order to calculate these embeddings, I used OpenAI's "text-embedding-3-small" model and calculated the cosine similarity via the "cosine_similarity" method in the "sklearn.metrics.pairwise" package in Python [130], [131].

Before running any further statistical testing, I decided to apply the Interquartile Range (IQR) methodology for outlier testing to the cosine similarity scores. In particular, I chose to classify values that were below $Q_1 - Q_{1.5} \times IQR$ or above $Q_3 + Q_{1.5} \times IQR$ (where Q_n refers to the nth quantile) as outliers, which is consistent with current best practices. These outliers were removed from further statistical tests.

Finally, I conducted normality tests on the cosine similarity for each question. In other words, for each specific prompt, I checked the distribution of cosine similarity scores for both the experimental group (those who completed the structured interaction) and the control group (those who were given only an unstructured interaction). I selected the Shapiro-Wilk test from the SciPy library in Python. This test was chosen due to the small sample size (approximately 50). In the case of 14 questions for the control group, and 13 for the experimental group, the data was found to not be normally distributed with a significance threshold of 0.05. Thus, I did not want to assume normality within the distribution of our results. As such, all comparisons for significance between the experimental and the control group were run using the The Mann-Whitney U test, and relied on the dissimilarity of the median rather than the mean. As a note, all other normality tests conducted are listed in their respective sections.

3.3.2 Cosine Similarity ($\mathbf{H_1}$)

This section presents the results of the hypothesis testing calculated on the cosine similarity scores. These scores represent the similarity of the model response to the participant's response.

I conducted a Mann-Whitney U test on the cosine similarity scores of each question with the null and alternative hypothesis (where $\tilde{x}$ represents the median) of:

$$H_{null}\colon\ \tilde{x}_{experimental} \leq \tilde{x}_{control} \tag{3.1}$$

$$H_{alt}\colon\ \tilde{x}_{experimental} > \tilde{x}_{control} \tag{3.2}$$

Table 2 presents a summary of the U statistic and the significance of each question. For a full list of the inputs that the participants were asked to predict and their responses, please see Appendix D. Remember that while the questions are numbered 1 through 20, they were presented in a random order to all participants.

Question #	U-statistic	P-value
1	1349.0	0.1431
2	1039.0	0.6899
3	1222.0	0.2360
4	1150.0	0.6402
5	1300.5	0.1382
6	1164.0	0.2476
7	1090.5	0.7828
8	1206.5	0.1163
9	1144.0	0.6560
10	725.0	0.9853
11	964.0	0.0351
12	1311.5	0.1188
13	1407.0	0.0695
14	842.5	0.9763
15	824.0	0.6133
16	1223.5	0.2962
17	1112.0	0.4774
18	1142.0	0.1909
19	1062.0	0.5565
20	830.0	0.6798

Table 3.2: Per question statistical significance with the specified null and alternative hypotheses

Overall, at a significance level of 0.05, I was only able to reject the null in the case of question 11. At a level of 0.1, I was able to reject the null hypothesis for both questions 11 and 13. In all other cases, I failed to reject the null. Furthermore, in a test for aggregate significance (i.e. the mean of the median across all 20 questions for both the control and experimental groups) there were no significant results.

3.3.3 Sentence Simplicity (H_2)

I chose to look at the "goodness" or "effectiveness" of the simplification across both groups. I used the SARI metric, which is a simplicity metric designed by Xu, Napoles, Pavlick, *et* *al.* [132] and used to benchmark sentence simplification for the models presented by Kew, Chi, Vásquez-Rodríguez, *et al.* [125]. I evaluated our examples on the SARI metric using the easse library, which is presented by Alva-Manchego, Martin, Scarton, *et al.* [133].

SARI Score

I decided to test for simplicity. I calculated the SARI score for each person's response to a predicted output. In other words, I generated a metric for the simplicity of what they thought the model response would be. Since the task was sentence simplification, I hypothesized that the experimental group would generate "better" simplifications (have a higher SARI score) than the control group. I conducted a Mann-Whitney U test for each question with the null and alternative hypothesis (where $\tilde{x}$ represents the median SARI score) of:

$$H_{null}\text{: } \tilde{x}_{experimental} \leq \tilde{x}_{control} \tag{3.3}$$

$$H_{alt}\text{: } \tilde{x}_{experimental} > \tilde{x}_{control} \tag{3.4}$$

There were no results (i.e. none of the 20 questions) that had a significance of less than 0.05. Thus, I failed to reject the null hypothesis to show that the experimental group was "better" at simplifying.

Difference in Sari Score

Subsequently, I decided to evaluate the difference in the simplicity of the participant's response to the model response. I used the SARI score for each person's response to a predicted output and subtracted the SARI score of the model response. I then took the absolute value. I hypothesized that the experimental group would better match the degree of simplification of the model (have a difference in participant-model SARI score closer to zero). I conducted a Mann-Whitney U test for each question with the null and alternative hypothesis (where $\tilde{x}$ represents the median SARI score and s_{model} represents the SARI score of the model response for that particular question) of:

$$H_{null}\text{: } |\tilde{x}_{experimental} - s_{model}| \geq |\tilde{x}_{control} - s_{model}| \tag{3.5}$$

$$H_{alt}\text{: } |\tilde{x}_{experimental} - s_{model}| < |\tilde{x}_{control} - s_{model}| \tag{3.6}$$

There were no results (i.e. none of the 20 questions) that had a significance of less than 0.05. Thus, I failed to reject the null hypothesis to show that the experimental group was able to "better" match the model simplification.

3.3.4 Participant Perception of Similarity (H_3)

In addition to being able to predict the model outputs, I also made a point to ask if the participants considered their response "similar enough" to the model response once they were shown the actual model response. On aggregate, the experimental group did not have a significantly higher percentage of people who believed that their response was "similar enough" to the model response.

I decided to investigate the correlation between the participant's similarity score and if they considered their response "similar enough". A high correlation would indicate that what individuals consider "similar enough" scales with increased cosine similarity. A negative correlation would show that lower cosine similarity scores are more likely to be considered "similar enough". If the p-value is too high, I fail to have a significant result and can not show a definite relationship between cosine similarity scoring and the participant's binary evaluation. Due to the nature of the data (one binary variable and one continuous variable, and non-normality in the continuous variable), I used the point-biserial statistical test. Table 3 has the results, per question, for both the experimental and control group.

Question	Control corr. coef.	Control p-value	Exper. corr. coef.	Exper. p-value
1	0.1788	0.2293	0.1202	0.4009
2	-0.0368	0.8080	0.1886	0.1992
3	0.3853	0.0089	0.0807	0.5773
4	0.0217	0.8835	0.2903	0.0409
5	0.1103	0.4606	0.0820	0.5753
6	0.2844	0.0645	0.2831	0.0464
7	0.1371	0.3526	0.2470	0.0837
8	0.0722	0.6455	0.3161	0.0269
9	0.2458	0.0922	0.1877	0.1917
10	0.1410	0.3556	0.1510	0.3278
11	-0.0567	0.7352	0.2303	0.1474
12	0.2611	0.0797	0.4579	0.0008
13	0.4995	0.0004	0.2637	0.0615
14	-0.0917	0.5446	0.2718	0.0616
15	0.1603	0.3364	0.3481	0.0191
16	0.3709	0.0112	0.0796	0.5827
17	0.0311	0.8376	0.1326	0.3691
18	0.1966	0.2064	0.3953	0.0054
19	0.1724	0.2574	0.1404	0.3411
20	-0.0576	0.7204	-0.0351	0.8234

Table 3.3: Correlation coefficients and statistical significance values for cosine similarity scoring and participant "similar enough" evaluation

Overall, there are no consistent trends across either the experimental or control group or across the majority of questions. While sometimes a weak positive correlation is observed, as the highlighted rows in Table 3 point out, there is never a case where both the experimental and control groups have a positive correlation with a significance less than 0.05.

3.3.5 Participant Ability to Detect Model Outputs (H_4)

In step three of the survey, participants were shown 20 outputs and asked what they believed the correct response to be. They were then rated on their ability to predict the correct model response. I hypothesized that the experimental would get more correct answers than the control group. The data failed the Shapiro normality test (i.e. the number of correct answers). I again employed the Mann-Whitney U test. However, there was no significant difference in medians across both groups.

Interestingly, the mean of both groups was very similar (note that the possible values range from 0 to 20):

- Experimental group mean: 13.125
- Control group mean: 13.0784

These results are worth noting as they are significantly higher than a purely random score of 10. Conducting a Wilcoxon signed-rank test (due to lack of normality in the scores) the *combination* of both groups with the following hypotheses (where $\tilde{x}$ is the median of the number correct answers in step three of every participant) :

$$H_{null}\colon \tilde{x} \leq 10 \tag{3.7}$$

$$H_{alt}\colon \tilde{x} > 10 \tag{3.8}$$

This results in the following values:

- A median across both groups is 13
- A p-value of 2.0986×10^{-14} for the statistical test

I can conclude that the performance in this section, while not statistically significant between groups, is statistically significant in that it remains higher than random performance.

3.3.6 Trust Evaluation (H_5)

At the end of the survey, I asked participants to complete the Trust evaluation. Notably, all answers were evaluated on a like scale (values range from 1-7). For each question on the trust scale, I ran the following Mann-Whitney U test with the following null and alternative hypotheses for each question (where $\tilde{x}_{experimental}$ is the median of the 1-7 scale in the experimental group and $\tilde{x}_{control}$ is the median of the 1-7 scale in the control group):

$$H_{null}\colon \tilde{x}_{experimental} \leq \tilde{x}_{control} \tag{3.9}$$

$$H_{alt}\colon \tilde{x}_{experimental} > \tilde{x}_{control} \tag{3.10}$$

For each question, the results are displayed in Table 4.

Trust Question	U Statistic	P-value
1	1152.0	0.7040
2	1405.5	0.0945
3	1115.0	0.7864
4	1337.5	0.2026

Table 3.4: Trust evaluation statistical testing results

While there is no strong evidence for a higher trust metric across any of the four questions in the Cahour and Forzy [128] scale, question 2 does have some relative significance. This relative significance combined with the weak evidence (at a significance level of 0.1) on question 2 is interesting to note. In particular, this question evaluates the participant's perception of the model **predictability**.

3.4 Analysis

3.4.1 Overall Performance

The first thing I noted after conducting this study, is that there is no significant aggregate trend among all the questions in part two of the survey (where the participant was asked to predict the model response). Specifically, there is no statistical metric that I found, either in cosine similarity of the participant's response to the model response, simplicity of the participant's response, or comparative simplicity of the participant's response to the model response that is significantly higher across all or most questions for the experimental group. This indicates, that going through a structured interaction process, may not be useful in the general case.

Secondly, I noted the lack of a consistent correlation between the participants perceiving their response as "similar enough" to the model response and the cosine similarity of their response and the model response. This indicates that, in general, individuals may not be able to determine what "similar enough" means in a quantitative sense and could instead rely on their intuition, which is often wrong. In fact, this could be a manifestation of what Klayman [134] describes as confirmation bias. In this survey, humans wish to believe that their prediction is accurate enough, making them more likely to classify something as accurate regardless of its similarity. However, even though only 6 out of 40 display a significant (< 0.05) correlation, it is worth noting that 5 of these correlations emerge from the experimental group. In other words, on 25% of the questions, there is a positive correlation observed in the experimental group, this is especially notable when you consider that this only happens in 5% of the control group. Thus, I find evidence that the structured analysis better equips participants to understand if their response is similar to the model response.

Finally, there was not a significant difference between the ability of the control and experimental groups to detect the outputs from the model under analysis. However, when considered together, the median of both groups was substantially and significantly higher than the pure random score of 10 out of 20. While this does not contribute to the understanding of the effect of structured analysis, it does highlight that predictability is an achievable goal, as humans can predict examples at a higher level than that of a random selector. This finding may warrant future research into whether this score can either increase or decrease.

3.4.2 Readability and Cosine Similarity

In looking at the per-question performance of cosine similarity (i.e., step two of the survey), I decided to look at readability as an indicator to explain the lack of or presence of significance in the difference in medians between the control and experimental groups. I selected readability as a metric. Readability, as defined by Flesch and Gould [135] is the ease with which a person can read a sentence and is based upon the total number of words per sentence and syllables per word. The Flesch-Kincaid Grade Level is the most common way in which this metric is quantified in the United States. This metric indicates the approximate grade level (in the American school system) that is required to understand the text.

I quantified the readability of both the model input and the model response. I then looked at the difference in readability between the input and response. Note that a positive difference indicates that a response is *more* readable than the input. This is because the higher the score, the less readable something is, as a higher grade level is required to understand it.

Question #	Difference in readability between in input and response
1	-0.7371
2	2.6276
3	1.4643
4	-2.3848
5	4.4302
6	4.7321
7	-0.2358
8	0.7011
9	5.8904
10	2.1500
11	7.1032
12	0.8717
13	0.0000
14	0.2064
15	2.2136
16	1.6617
17	2.9227
18	1.1767
19	2.9281
20	-3.3891

Table 3.5: Difference in readability for each input/response pair using the Flesch-Kincaid Grade Level

There are two points worth noting concerning readability scores. Firstly, the readability scores for the questions that the users are asked to answer are a mix of positive and negative. This demonstrates the difference in readability and simplicity. Specifically, readability is more related to sentence length and the number of syllables whereas simplicity (as quantified by the SARI score) is a relative measure [132], [135]. Thus, it is useful to analyze the model responses via these metrics. Secondly, the question that the experimental group and control group were found to have a significant difference in median (i.e., the experimental group had a better performance in terms of cosine similarity) was number 11. This is also the question where the model input and response had the **largest** difference in readability. The model input and response for Question #11 are displayed below.

Input: Weelkes was later to find himself in trouble with the Chichester Cathedral authorities for his heavy drinking and immoderate behaviour.
Model Response: Weelkes got into trouble with the Chichester Cathedral authorities for drinking too much and behaving badly.

Thus, I establish, that in the case of the highest difference in readability, the structured interaction helps the human generate significantly better predictions. This suggests that structured interaction may help humans to better predict model outputs when the differences in readability are greater.

This result is further supported by the trend observed through the linear correlation test.

Correlation of Increased Readability and Cosine Similarity

To see if there was a link between the participant's ability to predict the model response (quantified through cosine similarity) and readability difference, I conducted a linear correlation test. Note that I only looked at questions where there was a *positive* difference in readability (i.e., questions 1, 2, 7, and 20 were excluded). I conducted the Pearson linear correlation test on the average cosine similarity with the model difference in readability.

Group	Correlation Coefficient	P-value
Control	-0.5193	0.0393
Experimental	-0.5116	0.0428

Table 3.6: Correlation and significance calculated via the Pearson statistical test between questions with a positive readability difference and the mean cosine similarity

I observed that there was a clear and significant correlation in both the control and experimental groups. This negative correlation means that as the average cosine similarity grows, the difference in readability decreases. Thus, in both the control and experimental groups, participants are more likely to accurately predict the model response when the change in the readability level of the sentence is smaller.

Meaning of This Finding

This is both a trend towards less predictability in cases with higher differences and a more noticeable effect of the structured interaction in these cases with high differences. **I theorize that the more complex a task, the less likely humans are to predict the outcome, and the more helpful a structured interaction with model outputs will be**. While I see initial trends in this direction, future work must explore cases with even greater differences in the input and response of AI models that make even greater modifications in task complexity. I theorize that a structured interaction could help address this predictability gap.

3.4.3 Trust, Predictability, and Process Evaluation

I also briefly want to highlight the relative significance of the greater score in the experimental group on the trust scale. I do not comment on the trust calibration of this study, nor does this study aim to make significant claims in the field of AI trust. However, the nature of the second trust question is particularly notable.

In particular, the second question conducted in the trust evaluation concerns predictability. I find evidence that participants who complete the structured interaction find the AI model more predictable than those who do not. This indicates a potential trend, and I expect to see an increase in this significance with more complex tasks.

3.4.4 Overall Effect of a Structured Analysis on Model Predictability

These results and the subsequent analysis highlight three trends that require further investigation. In particular, I find the following effect of a *structured interaction*:

1. Participants can better identify the similarity of their prediction to the model prediction.

2. Initial results outlined here suggest that in cases of more complex tasks, a structured interaction can help humans better predict the model behaviour. This kind of interaction will have less effect with less complicated tasks. Future work can explore these relationships further.

3. Participants perceive an AI model to be more predictable after they have completed a structured interaction. I also expect this difference in perception to rise with more complex tasks.

Overall, this study points to a new way of thinking about AI system behaviour. It investigates and argues that it is feasible to predict how a model will behave, and while AI is "unpredictable" it is not impossible to develop an understanding of this behaviour. By participating in thoughtful engagement with the model inputs and responses over a series of examples, I observed that participants were better able to predict model behaviour for a more complex task. Thus, while AI may be "unpredictable" in principle, there are steps that humans can take to understand its behaviour in deployment contexts. Furthermore, while our work suggests that there are steps that humans can take to better understand model behaviour, future work can and should explore these possibilities further.

3.5 Summary of Chapter Findings

Overall, I find evidence that it is possible to move the human baseline for predicting an AI system's behaviour. I find that a structured interaction is one way to do this. These significant results indicate a trend for future work. In general, the evidence I collected suggests that structured analysis and careful consideration of model inputs and responses can influence a person's ability to predict system behaviour. I hope that future work will serve to investigate this further. For example, I would either investigate a wider range of readability differences or select a different task that has fewer constraints in the kind of output we expect. However, the evidence I found here is an important first indicator of how this methodology can be utilized to influence and modify a person's ability to predict AI behaviour.

Chapter 4

Concluding Thoughts

Overall this thesis has demonstrated that an understanding of system behaviour of an AI model at the time of deployment is attainable and a proficient strategy for policy. This thesis has demonstrated that a liability regime based on this understanding of system behaviour serves to fill in an existing regulatory deficiency or "hole." Chapter 2 established this argument by examining the possibility of ascertaining this understanding of system behaviour through documentation at the deployment stage. Chapter 2 also proposed an approach to liability based on this documentation. The other part of this argument was outlined in Chapter 3. In this chapter, it was shown that there are ways in which a person can better understand a generative task. Importantly, an overarching understanding of system behaviour is not impossible to achieve, and cannot only be established but can be improved. Together this thesis has provided a methodical guide to understanding AI system behaviour and establishing a new pathway for effective regulation.

The AI regulatory landscape is fast evolving. While there is a developing trend towards risk assessment, this is not going to be a sufficient policy basis moving forward. AI is often thought of as a 'new age,' exceptional technology regarding regulation. While we can not say that AI is identical to past technologies, and it does require special consideration, we can not forget that already existing regulatory structures, tactics, and opportunities exist. The frameworks we choose to adopt will have the same strengths and weaknesses that they did in the past, manifested in new ways. This is exactly the trend with risk management frameworks. The need for a civil liability framework to accompany the risk management approach. This thesis provides an answer to this question and illustrates that civil liability based on *existing* and *effective* regulatory strategies can be applied to AI system behaviour.

Appendix A

The Deployment Task and Survey

Below is the task that was presented to participants:

Welcome!

In this survey, you will be evaluating an AI Model.

The goal of this research is to determinehow interacting with a series of AI model outputs influences your understanding of future outputs.

Important notes on this survey:

- Participation is voluntary
- You may decline to answer any or all questions
- You may decline further participation, at any time, without adverse consequences, by exiting the survey
- Your anonymity is assured

Below is a description of your role and the task:

You work for Company XYZ. Company XYZ provides language education and training services to its clients. They specifically target non-native English speakers.

Company XYZ is considering deploying an AI model to allow its clients to take any complicated sentence, and output a sentence that is simpler, and can be understood by a non-native English speaker while preserving the meaning present in the sentence.

Company XYZ has asked you to evaluate this model for deployment. Specifically, you are evaluating that this model can meet its specified desiderata.

You will be shown a series of sample input/response pairs that have been generated by the model.

Then, you will be presented with a series of inputs to the model and asked to predict the model's response.

Click continue to confirm you understand the information presented here, and that you consent to participate.

For any questions, please reach out to tlcurtis@mit.edu.

You will then learn more about the prediction process and remuneration for completing this survey.
If you wish to view the whole survey, please view the survey at:
https://mit.co1.qualtrics.com/jfe/preview/previewId/83866b12-2f26-4479-b4c0-7c7871a3e6b2/SV_a9oVQVpxsLHGnFc?Q_CHL=preview&Q_SurveyVersionID=current.

This URL takes you to a link to preview the entire survey. If you are interested, you may complete the survey, and no response will be logged.

Appendix B

Open Coding for Machine Learning Workflow

The Open Coding for Machine Learning Workflow has two key stages. The first stage is the annotation stage, where the user is shown a series of Input/Response pairs and asked to annotate them for the desiderata in any way they see fit.

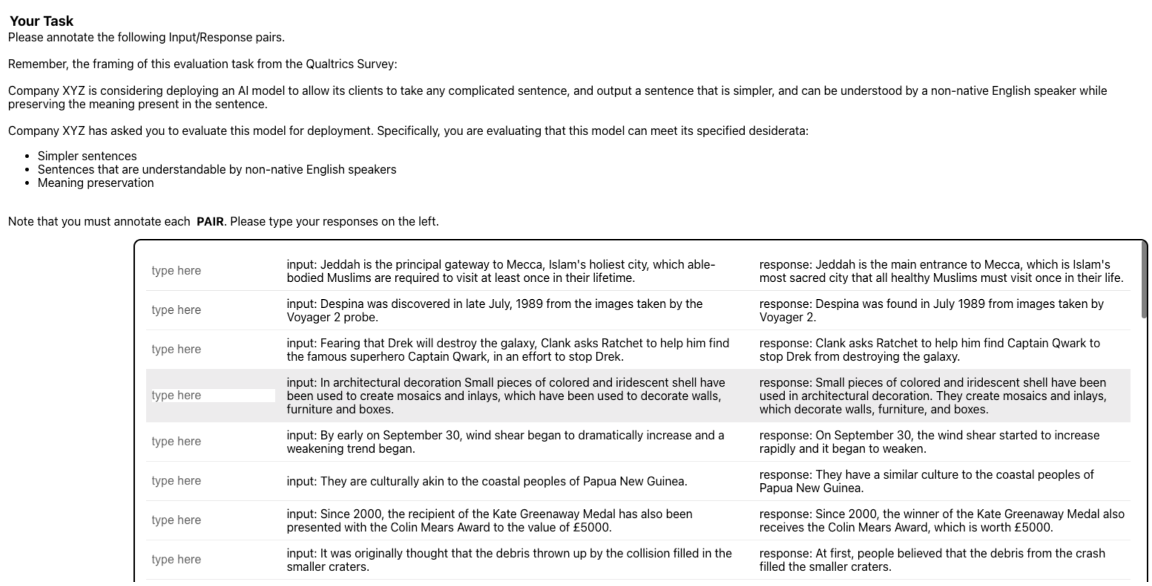

Figure B.1: The annotation page

After completing these annotations, the user is then presented with a second interface, where they can group together similar annotations to generate a "code". Note that a person can choose to put together certain annotations, or create a direct one-to-one mapping of their annotation to their groups.

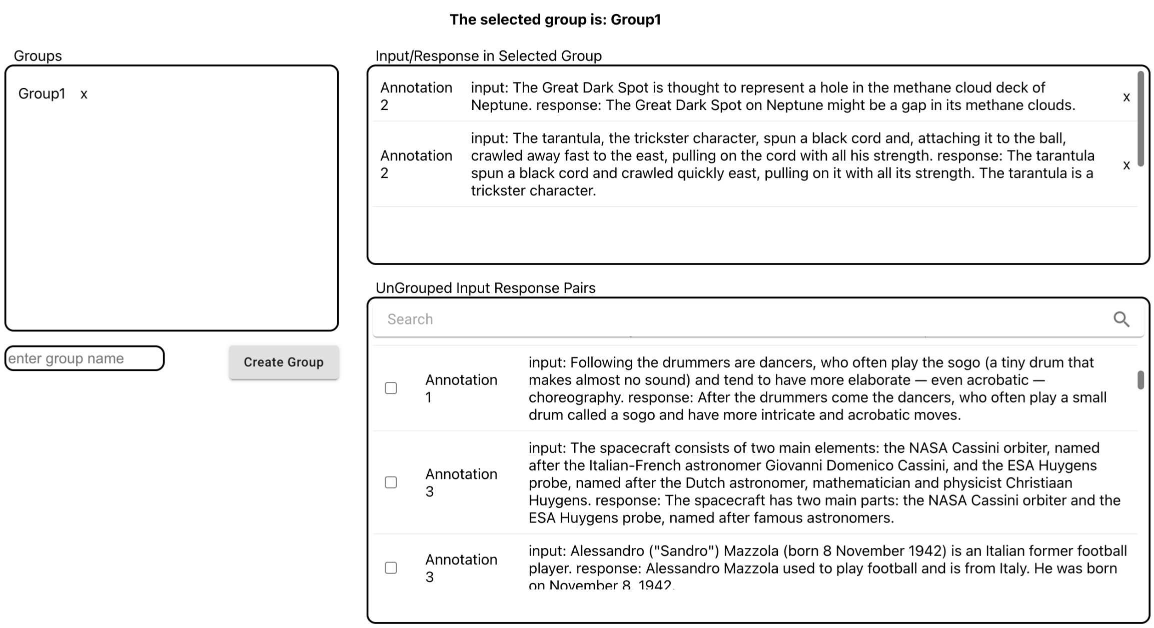

Figure B.2: The grouping page

Appendix C

Input/Response Pairs Used for The User Study

The test data is the Input/Response pairs that users are asked to predict.

Input	Response
Like previous assessment reports, it consists of four reports, three of them from its working groups.	The assessment report, similar to previous ones, has four reports, three of which come from working groups.
Their granddaughter Hélène Langevin-Joliot is a professor of nuclear physics at the University of Paris, and their grandson Pierre Joliot, who was named after Pierre Curie, is a noted biochemist.	Their granddaughter is a nuclear physics professor at the University of Paris, and their grandson is a biochemist named after Pierre Curie.
This stamp remained the standard letter stamp for the remainder of Victoria's reign, and vast quantities were printed.	This stamp was used as the standard letter stamp for the rest of Victoria's reign, and a lot of them were printed.
The International Fight League was an American mixed martial arts (MMA) promotion billed as the world's first MMA league.	The International Fight League was an American organization that promoted mixed martial arts (MMA) and was considered the first MMA league.
Giardia lamblia (synonymous with Lamblia intestinalis and Giardia duodenalis) is a flagellated protozoan parasite that colonises and reproduces in the small intestine, causing giardiasis.	Giardia lamblia (also known as Lamblia intestinalis and Giardia duodenalis) is a parasite that lives and multiplies in the small intestine, causing giardiasis.
Aside from this, Cameron has often worked in Christian-themed productions, among them the post-Rapture films Left Behind: The Movie, Left Behind II: Tribulation Force, and Left Behind: World at War, in which he plays Cameron "Buck" Williams.	Cameron has acted in many Christian-themed movies, including the Left Behind series where he plays the character of Cameron "Buck" Williams.
This was the area east of the mouth of the Vistula River, later sometimes called "Prussia proper".	This was the area east of the Vistula River's mouth, sometimes known as "Prussia proper" later on.
After graduation he returned to Yerevan to teach at the local Conservatory and later he was appointed artistic director of the Armenian Philarmonic Orchestra.	After graduating, he taught at the Yerevan Conservatory and became the artistic director of the Armenian Philarmonic Orchestra.
The story of Christmas is based on the biblical accounts given in the Gospel of Matthew, namely - and the Gospel of Luke, specifically -.	The story of Christmas comes from the Bible in the Gospel of Matthew and Luke.

Input	Response
Weelkes was later to find himself in trouble with the Chichester Cathedral authorities for his heavy drinking and immoderate behaviour.	Weelkes got into trouble with the Chichester Cathedral authorities for drinking too much and behaving badly.
So far the 'celebrity' episodes have included Vic Reeves, Nancy Sorrell, Gaby Roslin, Scott Mills, Mark Chapman, Simon Gregson, Sue Cleaver, Carol Thatcher, Paul O'Grady and Lee Ryan.	The 'celebrity' episodes have had guests such as Vic Reeves, Nancy Sorrell, Gaby Roslin, and others.
It was discovered by Stephen P. Synnott in images from the Voyager 1 space probe taken on March 5, 1979 while orbiting around Jupiter.	Stephen P. Synnott discovered it in images from the Voyager 1 space probe on March 5, 1979 while orbiting Jupiter.
Gomaespuma was a Spanish radio show, hosted by Juan Luis Cano and Guillermo Fesser.	Gomaespuma was a Spanish radio show hosted by Juan Luis Cano and Guillermo Fesser.
On 16 June 2009, the official release date of The Resistance was announced on the band's website.	On June 16, 2009, the band's website announced the official release date of The Resistance.
He is also a member of another Jungiery boyband 183 Club.	He is also part of another boyband called 183 Club.
The Apostolic Tradition, attributed to the theologian Hippolytus, attests the singing of Hallel psalms with Alleluia as the refrain in early Christian agape feasts.	In early Christian feasts, people sang Hallel psalms with the refrain Alleluia, as described in the Apostolic Tradition written by Hippolytus.
In return, Rollo swore fealty to Charles, converted to Christianity, and undertook to defend the northern region of France against the incursions of other Viking groups.	Rollo promised loyalty to Charles, became Christian, and agreed to protect northern France from other Viking attacks.
It is derived from Voice of America (VoA) Special English.	It comes from Voice of America (VoA) Special English.
Disney received a full-size Oscar statuette and seven miniature ones, presented to him by 10-year-old child actress Shirley Temple.	Disney received a big Oscar and seven small ones from Shirley Temple, who was 10 years old.
It was the first asteroid to be discovered by a spacecraft.	A spacecraft discovered this asteroid before any others.

Table C.1: Step two: The test input/response pairs

Input	Response
It will then dislodge itself and sink back to the river bed in order to digest its food and wait for its next meal.	After eating, the animal will sink back to the river bed and wait for its next meal.
Furthermore, research has shown children are less likely to report a crime if it involves someone that he or she knows, trusts, and / or cares about.	Research has shown that children are less likely to report a crime if it involves someone they know, trust, or care about.
Today, Landis' father has become a hearty supporter of his son and regards himself as one of Floyd's biggest fans.	Landis' father is now a big fan and supporter of his son, Floyd.
Shortly after attaining Category 4 status, the outer convection of the hurricane became ragged.	After becoming a Category 4 hurricane, its outer convection became ragged.
The equilibrium price for a certain type of labor is the wage.	The wage is the equilibrium price for a certain type of labor.
Convinced that the grounds were haunted, they decided to publish their findings in a book An Adventure (1911), under the pseudonyms of Elizabeth Morison and Frances Lamont.	They believed the grounds were haunted, and wrote a book about it called "An Adventure" (1911) under the names Elizabeth Morison and Frances Lamont.
He settled in London, devoting himself chiefly to practical teaching.	He moved to London and focused mainly on teaching.
Brunstad has several fast food restaurants, a cafeteria-style restaurant, coffee bar, and its own grocery store.	Brunstad has many places to eat, including fast food restaurants, a cafeteria-style restaurant, a coffee bar, and its own grocery store.
He left a detachment of 11,000 troops to garrison the newly conquered region.	He left 11,000 troops to guard the region he just conquered.
In 1438 Trevi passed under the temporal rule of the Church as part of the legation of Perugia, and thenceforth its history merges first with that of the States of the Church, then (1860) with the united Kingdom of Italy.	Trevi came under the rule of the Church in 1438 as part of Perugia's legation. From then on, its history was intertwined with the States of the Church and later with the Kingdom of Italy.
The depression moved inland on the 20th as a circulation devoid of convection, and dissipated the next day over Brazil, where it caused heavy rains and flooding.	On the 20th, the depression moved inland without any thunderstorms and disappeared the following day in Brazil, which resulted in floods and heavy rainfall.
The New York City Housing Authority Police Department was a law enforcement agency in New York City that existed from 1952 to 1995.	The New York City Housing Authority Police Department was a police agency in New York City that was around from 1952 to 1995.

Input	Response
The current lineup of the band comprises Flynn (vocals, guitar), Duce (bass), Phil Demmel (guitar), and Dave McClain (drums).	The band's current lineup includes Flynn (singer, guitarist), Duce (bassist), Phil Demmel (guitarist), and Dave McClain (drummer).
Advocacy Countries with a minority Muslim population are more likely than Muslim-majority countries of the Greater Middle East to use mosques as a way to promote civic participation.	Countries with fewer Muslims are more likely than Muslim-majority countries in the Greater Middle East to use mosques to encourage civic involvement.
The characters are foul-mouthed extensions of their earlier characters Pete and Dud.	The characters are vulgar versions of their previous personas, Pete and Dud.
Johan was also the original bassist of the Swedish power metal band HammerFall, but quit before the band ever released a studio album.	Johan was part of the band HammerFall as the bassist, but he left before they made an album.
In 1998, Culver ran for Iowa Secretary of State and was victorious.	Culver won the election for Iowa Secretary of State in 1998.
In 1990, Mark Messier took the Hart over Ray Bourque by a margin of two votes, the difference being a single first-place vote.	Mark Messier won the Hart award in 1990 by two votes, with only one more first-place vote than Ray Bourque.
Shade sets the main plot of the novel in motion when he impetuously defies that law, and inadvertently initiates a chain of events that leads to the destruction of his colony's home, forcing their premature migration, and his separation from them.	Shade starts the main story of the book by breaking a law, which causes a chain of events leading to his colony's home being destroyed. They have to move early and Shade is separated from them.
The female equivalent is a daughter.	The female counterpart is a daughter.

Table C.2: Step three: The model under analysis input/response pairs

Input	Response
It will then dislodge itself and sink back to the river bed in order to digest its food and wait for its next meal.	It will then dislodge itself and sink back to the river bed to digest its food.
Furthermore, research has shown children are less likely to report a crime if it involves someone that he or she knows, trusts, and / or cares about.	Furthermore, research shows children are less likely to report a crime if it involves someone he knows, trusts, and / or cares about.
Today, Landis' father has become a hearty supporter of his son and regards himself as one of Floyd's biggest fans.	Landis' dad was a supporter of his son. He gave him lots of support.
Shortly after attaining Category 4 status, the outer convection of the hurricane became ragged.	Shortly after reaching Category 4 status, the outer convection of the hurricane became ragged.
The equilibrium price for a certain type of labor is the wage.	The price for certain labor is the wage.
Convinced that the grounds were haunted, they decided to publish their findings in a book An Adventure (1911), under the pseudonyms of Elizabeth Morison and Frances Lamont.	They found the grounds were haunted and decided to publish their findings in a book, An Adventure.
He settled in London, devoting himself chiefly to practical teaching.	He settled in London, Devoting himself to practical teaching.
Brunstad has several fast food restaurants, a cafeteria-style restaurant, coffee bar, and its own grocery store.	Brunstad has several fast food restaurants, coffee bar, cafeteria, and grocery store.
He left a detachment of 11,000 troops to garrison the newly conquered region.	He left a detachment of 11,000 troops.
In 1438 Trevi passed under the temporal rule of the Church as part of the legation of Perugia, and thenceforth its history merges first with that of the States of the Church, then (1860) with the united Kingdom of Italy.	In 1438 Trevi was under the temporal rule of the Church as part of the legation of Perugia. Thenceforth its history merges first with that of the States of the Church, then (1860) with the United Kingdom of Italy.
The depression moved inland on the 20th as a circulation devoid of convection, and dissipated the next day over Brazil, where it caused heavy rains and flooding.	The depression moved inland as a movement without convection. It dissipated the next day over Brazil. It caused heavy rains and flooding.
The New York City Housing Authority Police Department was a law enforcement agency in New York City that existed from 1952 to 1995.	The New York City Housing Authority Police Department was an agency in New York City from 1952 to 1995.

Input	Response
The current lineup of the band comprises Flynn (vocals, guitar), Duce (bass), Phil Demmel (guitar), and Dave McClain (drums).	The current band has Flynn, Duce, Phil Demmel, and Dave McClain.
Advocacy Countries with a minority Muslim population are more likely than Muslim-majority countries of the Greater Middle East to use mosques as a way to promote civic participation.	Country's with a Muslim minority population are more likely than Muslim-majority countries of the Greater Middle East to use mosques as a way to promote civic participation.
The characters are foul-mouthed extensions of their earlier characters Pete and Dud.	The characters are extensions of Pete and Dud.
Johan was also the original bassist of the Swedish power metal band HammerFall, but quit before the band ever released a studio album.	Johan was the bassist of the Swedish power metal band HammerFall. He had to quit before the band released an album.
In 1998, Culver ran for Iowa Secretary of State and was victorious.	Culver was elected Iowa Secretary of State in 1998.
In 1990, Mark Messier took the Hart over Ray Bourque by a margin of two votes, the difference being a single first-place vote.	Mark Messier won Hart over Ray Bourque by a margin of two votes. Bourque lost by just one vote.
Shade sets the main plot of the novel in motion when he impetuously defies that law, and inadvertently initiates a chain of events that leads to the destruction of his colony's home, forcing their premature migration, and his separation from them.	Shade breaks a law to set the plot in motion of the novel. It leads to the destruction of the colony's home and he's separation from them.
The female equivalent is a daughter.	Female equivalent is a daughter.

Table C.3: Step three: The comparison model input/response pairs

Appendix D

Input Response Pairs that The User is Asked to Predict

Input	Response
Like previous assessment reports, it consists of four reports, three of them from its working groups.	The assessment report, similar to previous ones, has four reports, three of which come from working groups.
Their granddaughter Hélène Langevin-Joliot is a professor of nuclear physics at the University of Paris, and their grandson Pierre Joliot, who was named after Pierre Curie, is a noted biochemist.	Their granddaughter is a nuclear physics professor at the University of Paris, and their grandson is a biochemist named after Pierre Curie.
This stamp remained the standard letter stamp for the remainder of Victoria's reign, and vast quantities were printed.	This stamp was used as the standard letter stamp for the rest of Victoria's reign, and a lot of them were printed.
The International Fight League was an American mixed martial arts (MMA) promotion billed as the world's first MMA league.	The International Fight League was an American organization that promoted mixed martial arts (MMA) and was considered the first MMA league.
Giardia lamblia (synonymous with Lamblia intestinalis and Giardia duodenalis) is a flagellated protozoan parasite that colonises and reproduces in the small intestine, causing giardiasis.	Giardia lamblia (also known as Lamblia intestinalis and Giardia duodenalis) is a parasite that lives and multiplies in the small intestine, causing giardiasis.
Aside from this, Cameron has often worked in Christian-themed productions, among them the post-Rapture films Left Behind: The Movie, Left Behind II: Tribulation Force, and Left Behind: World at War, in which he plays Cameron "Buck" Williams.	Cameron has acted in many Christian-themed movies, including the Left Behind series where he plays the character of Cameron "Buck" Williams.
This was the area east of the mouth of the Vistula River, later sometimes called "Prussia proper".	This was the area east of the Vistula River's mouth, sometimes known as "Prussia proper" later on.
After graduation he returned to Yerevan to teach at the local Conservatory and later he was appointed artistic director of the Armenian Philarmonic Orchestra.	After graduating, he taught at the Yerevan Conservatory and became the artistic director of the Armenian Philarmonic Orchestra.
The story of Christmas is based on the biblical accounts given in the Gospel of Matthew, namely - and the Gospel of Luke, specifically -.	The story of Christmas comes from the Bible in the Gospel of Matthew and Luke.
Weelkes was later to find himself in trouble with the Chichester Cathedral authorities for his heavy drinking and immoderate behaviour.	Weelkes got into trouble with the Chichester Cathedral authorities for drinking too much and behaving badly.

Input	Response
So far the 'celebrity' episodes have included Vic Reeves, Nancy Sorrell, Gaby Roslin, Scott Mills, Mark Chapman, Simon Gregson, Sue Cleaver, Carol Thatcher, Paul O'Grady and Lee Ryan.	The 'celebrity' episodes have had guests such as Vic Reeves, Nancy Sorrell, Gaby Roslin, and others.
It was discovered by Stephen P. Synnott in images from the Voyager 1 space probe taken on March 5, 1979 while orbiting around Jupiter.	Stephen P. Synnott discovered it in images from the Voyager 1 space probe on March 5, 1979 while orbiting Jupiter.
Gomaespuma was a Spanish radio show, hosted by Juan Luis Cano and Guillermo Fesser.	Gomaespuma was a Spanish radio show hosted by Juan Luis Cano and Guillermo Fesser.
On 16 June 2009, the official release date of The Resistance was announced on the band's website.	On June 16, 2009, the band's website announced the official release date of The Resistance.
He is also a member of another Jungiery boyband 183 Club.	He is also part of another boyband called 183 Club.
The Apostolic Tradition, attributed to the theologian Hippolytus, attests the singing of Hallel psalms with Alleluia as the refrain in early Christian agape feasts.	In early Christian feasts, people sang Hallel psalms with the refrain Alleluia, as described in the Apostolic Tradition written by Hippolytus.
In return, Rollo swore fealty to Charles, converted to Christianity, and undertook to defend the northern region of France against the incursions of other Viking groups.	Rollo promised loyalty to Charles, became Christian, and agreed to protect northern France from other Viking attacks.
It is derived from Voice of America (VoA) Special English.	It comes from Voice of America (VoA) Special English.
Disney received a full-size Oscar statuette and seven miniature ones, presented to him by 10-year-old child actress Shirley Temple.	Disney received a big Oscar and seven small ones from Shirley Temple, who was 10 years old.
It was the first asteroid to be discovered by a spacecraft.	A spacecraft discovered this asteroid before any others.

Table D.1: Input/Response pairs that the user is asked to predict, listed in order

References

[1] D. A. Stone, *Policy paradox: the art of political decision making*, eng, Third edition. New York: W.W. Norton & Company, 2022, viii, 408 pages, ISBN: 9780393912722.

[2] L. Winner, "Do artifacts have politics?" *Computer Ethics*, pp. 177–192, May 2017. DOI: 10.4324/9781315259697-21/ARTIFACTS-POLITICS-LANGDON-WINNER. [Online]. Available: https://www.taylorfrancis.com/chapters/edit/10.4324/9781315259697-21/artifacts-politics-langdon-winner.

[3] M. E. Kaminski, "Regulating the risks of ai," *Forthcoming, Boston University Law Review*, vol. 103, 2023.

[4] L. Illia, E. Colleoni, and S. Zyglidopoulos, "Ethical implications of text generation in the age of artificial intelligence," *Business Ethics, the Environment & Responsibility*, vol. 32, no. 1, pp. 201–210, 2023.

[5] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "The off-switch game," *arXiv preprint arXiv:1611.08219*, 2016.

[6] A. M. Turner, L. Smith, R. Shah, A. Critch, and P. Tadepalli, "Optimal policies tend to seek power," *arXiv preprint arXiv:1912.01683*, 2019.

[7] R. Ngo, "The alignment problem from a deep learning perspective," *arXiv preprint arXiv:2209.00626*, 2022.

[8] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, *et al.*, "Holistic evaluation of language models," *arXiv preprint arXiv:2211.09110*, 2022.

[9] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, "The woman worked as a babysitter: On biases in language generation," *arXiv preprint arXiv:1909.01326*, 2019.

[10] I. El Atillah. "Man ends his life after an ai chatbot 'encouraged' him to sacrifice himself to stop climate change." (2023), [Online]. Available: https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-.

[11] L. Lucy and D. Bamman, "Gender and representation bias in gpt-3 generated stories," in *Proceedings of the Third Workshop on Narrative Understanding*, 2021, pp. 48–55.

[12] G. Spitale, N. Biller-Andorno, and F. Germani, "Ai model gpt-3 (dis) informs us better than humans," *arXiv preprint arXiv:2301.11924*, 2023.

[13] J. Chamberlain, "The risk-based approach of the european union's proposed artificial intelligence regulation: Some comments from a tort law perspective," *European Journal of Risk Regulation*, vol. 14, pp. 1–13, 1 2023, ISSN: 1867-299X. DOI: DOI:10.1017/err.2022.38. [Online]. Available: https://www.cambridge.org/core/article/riskbased-approach-of-the-european-unions-proposed-artificial-intelligence-regulation-some-comments-from-a-tort-law-perspective/A996034CC512B6B8A77B73FE39E77DAE.

[14] M. Bovens, "Analysing and assessing accountability: A conceptual framework1," *European Law Journal*, vol. 13, pp. 447–468, 4 Jul. 2007, ISSN: 1468-0386. DOI: 10.1111/J.1468-0386.2007.00378.X. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1468-0386.2007.00378.x%20https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0386.2007.00378.x%20https://onlinelibrary.wiley.com/doi/10.1111/j.1468-0386.2007.00378.x.

[15] S. I. Lindberg, "Mapping accountability: Core concept and subtypes," *International Review of Administrative Sciences*, vol. 79, pp. 202–226, 2 Jun. 2013, ISSN: 00208523. DOI: 10.1177/0020852313477761/ASSET/IMAGES/LARGE/10.1177_0020852313477761-FIG1.JPEG. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0020852313477761.

[16] R. Mulgan, "'accountability': An ever-expanding concept?" *Public Administration*, vol. 78, pp. 555–573, 3 Jan. 2000, ISSN: 1467-9299. DOI: 10.1111/1467-9299.00218. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/1467-9299.00218%20https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9299.00218%20https://onlinelibrary.wiley.com/doi/10.1111/1467-9299.00218.

[17] I. Thynne and J. Goldring, "Accountability and control: Government officials and the exercise of power," *(No Title)*, 1987.

[18] C. Novelli, M. Taddeo, and L. Floridi, "Accountability in artificial intelligence: What it is and how it works," *AI and Society*, vol. 1, pp. 1–12, Feb. 2023, ISSN: 14355655. DOI: 10.1007/S00146-023-01635-Y/FIGURES/1. [Online]. Available: https://link.springer.com/article/10.1007/s00146-023-01635-y.

[19] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai," *Berkman Klein Center Research Publication*, no. 2020-1, 2020.

[20] J. A. Kroll, "Outlining traceability: A principle for operationalizing accountability in computing systems," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21, Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 758–771, ISBN: 9781450383097. DOI: 10.1145/3442188.3445937. [Online]. Available: https://doi.org/10.1145/3442188.3445937.

[21] N. Telecommunications and I. A. (NTIA), *Ai accountability policy report.* [Online]. Available: https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report.

[22] C. Percy, S. Dragicevic, S. Sarkar, and A. d'Avila Garcez, "Accountability in ai: From principles to industry-specific accreditation," *AI Communications*, vol. 34, pp. 181–196, 2021, ISSN: 1875-8452. DOI: 10.3233/AIC-210080. [Online]. Available: https://content.iospress.com:443/articles/ai-communications/aic210080.

[23] C. Curtis, N. Gillespie, and S. Lockey, "Ai-deploying organizations are key to addressing 'perfect storm' of ai risks," *AI and Ethics*, vol. 3, pp. 145–153, 1 2023, ISSN: 2730-5961. DOI: 10.1007/s43681-022-00163-7. [Online]. Available: https://doi.org/10.1007/s43681-022-00163-7.

[24] K. Crawford, *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021, ISBN: 0300252390.

[25] C. O'neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017, ISBN: 0553418831.

[26] K. Crawford and R. Calo, "There is a blind spot in ai research," *Nature*, vol. 538, pp. 311–313, 7625 2016, ISSN: 1476-4687. DOI: 10.1038/538311a. [Online]. Available: https://doi.org/10.1038/538311a.

[27] M. Altman, A. Wood, and E. Vayena, "A harm-reduction framework for algorithmic fairness," *IEEE Security & Privacy*, vol. 16, pp. 34–45, 3 2018, ISSN: 1558-4046. DOI: 10.1109/MSP.2018.2701149.

[28] P. G. R. de Almeida, C. D. dos Santos, and J. S. Farias, "Artificial intelligence regulation: A framework for governance," *Ethics and Information Technology*, vol. 23, no. 3, pp. 505–525, 2021.

[29] N. AI, "Artificial intelligence risk management framework (ai rmf 1.0)," 2023.

[30] M. Almada and N. Petit, "The eu ai act: A medley of product safety and fundamental rights?" *SSRN Electronic Journal*, Oct. 2023. DOI: 10.2139/SSRN.4308072. [Online]. Available: https://papers.ssrn.com/abstract=4308072.

[31] J. Schuett, "Risk management in the artificial intelligence act," *European Journal of Risk Regulation*, pp. 1–19, 2023, ISSN: 1867-299X. DOI: DOI:10.1017/err.2023.1. [Online]. Available: https://www.cambridge.org/core/product/2E4D5707E65EFB3251A76E288BA74068.

[32] R. Dotan, B. Blili-Hamelin, R. Madhavan, J. Matthews, and J. Scarpino, "Evolving ai risk management: A maturity model based on the nist ai risk management framework," *arXiv preprint arXiv:2401.15229*, 2024.

[33] A. Dimou, "Airo: An ontology for representing ai risks based on the proposed eu ai act and iso risk management standards," *Towards a Knowledge-Aware AI: SEMANTiCS*, p. 51, 2022.

[34] E. Kazim, C. Kerrigan, and A. Koshiyama, "Eu proposed ai legal framework," *Available at SSRN 3846898*, 2021.

[35] A. Kak and S. M. West, "General purpose ai poses serious risks, should not be excluded from the eu's ai act: Policy brief," AI Now Institute, Apr. 2023.

[36] J. De Cooman, "Humpty dumpty and high-risk ai systems: The ratione materiae dimension of the proposal for an eu artificial intelligence act," *Market and Competition Law Review*, vol. 6, pp. 49–88, 1 2022. [Online]. Available: https://heinonline.org/HOL/P?h=hein.journals/mclr6&i=49%20https://heinonline.org/HOL/PrintRequest?handle=hein.journals/mclr6&collection=journals&div=6&id=49&print=section&sction=6.

[37] M. Veale and F. Z. Borgesius, "Demystifying the draft eu artificial intelligence act — analysing the good, the bad, and the unclear elements of the proposed approach," vol. 22, pp. 97–112, 4 2021. DOI: doi:10.9785/cri-2021-220402. [Online]. Available: https://doi.org/10.9785/cri-2021-220402.

[38] R. Baldwin, M. Cave, and M. Lodge, "281Risk-based Regulation," in *Understanding Regulation: Theory, Strategy, and Practice*, Oxford University Press, Oct. 2011, ISBN: 9780199576081. DOI: 10.1093/acprof:osobl/9780199576081.003.0013. eprint: https://academic.oup.com/book/0/chapter/151921999/chapter-ag-pdf/44974181/book_7235_section_151921999.ag.pdf. [Online]. Available: https://doi.org/10.1093/acprof:osobl/9780199576081.003.0013.

[39] K. R. Hammond, *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. Oxford University Press, 1996, ISBN: 0190282665.

[40] R. B. Stewart, "The role of the courts in risk management," *Environmental Law Reporter News & Analysis*, vol. 16, 1986. [Online]. Available: https://heinonline.org/HOL/Page?handle=hein.journals/elrna16&id=206&div=59&collection=journals.

[41] A. Lior, "Ai entities as ai agents: Artificial intelligence liability and the ai respondeat superior analogy," *Mitchell Hamline Law Review*, vol. 46, p. 2, 5 2020. [Online]. Available: https://open.mitchellhamline.edu/cgi/viewcontent.cgi?article=1223&context=mhlr.

[42] I. Giuffrida, "Liability for ai decision-making: Some legal and ethical considerations symposium: Rise of the machines: Artificial intelligence, robotics, and the reprogramming of law," *Fordham Law Review*, vol. 88, p. 439, 2020.

[43] H. Zech, "Liability for ai: Public policy considerations," *ERA Forum*, vol. 22, pp. 147–158, 1 2021, ISSN: 1863-9038. DOI: 10.1007/s12027-020-00648-0. [Online]. Available: https://doi.org/10.1007/s12027-020-00648-0.

[44] C. Wendehorst, "Strict liability for ai and other emerging technologies," vol. 11, pp. 150–180, 2 2020. DOI: doi:10.1515/jetl-2020-0140. [Online]. Available: https://doi.org/10.1515/jetl-2020-0140.

[45] R. Crootof, M. E. Kaminski, W. Price, and I. I. Nicholson, "Humans in the loop," *Vand. L. Rev.*, vol. 76, p. 429, 2023. [Online]. Available: https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/vanlr76&section=14&casa_token=83VtGvFMW8IAAAAA:p8GHmIZGq7lNSpC95MZVap8qwtPFhftZ531RM7qhRrD14dgLDwC93uJBA3PQEqJ_r9XIY8_l8w.

[46] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 33–44.

[47] I. Naja, M. Markovic, P. Edwards, and C. Cottrill, "A semantic framework to support ai system accountability and audit," in *The Semantic Web*, R. Verborgh, K. Hose, H. Paulheim, P.-A. Champin, M. Maleshkova, O. Corcho, P. Ristoski, and M. Alam, Eds., Cham: Springer International Publishing, 2021, pp. 160–176, ISBN: 978-3-030-77385-4.

[48] F. Königstorfer and S. Thalmann, "Ai documentation: A path to accountability," *Journal of Responsible Technology*, vol. 11, p. 100 043, 2022.

[49] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, *et al.*, "Toward trustworthy ai development: Mechanisms for supporting verifiable claims," *arXiv preprint arXiv:2004.07213*, 2020.

[50] F. Königstorfer and S. Thalmann, "Software documentation is not enough! requirements for the documentation of ai," *Digital Policy, Regulation and Governance*, vol. 23, no. 5, pp. 475–488, 2021.

[51] M. C. Buiten, "Towards intelligent regulation of artificial intelligence," *European Journal of Risk Regulation*, vol. 10, pp. 41–59, 1 2019, ISSN: 1867-299X. DOI: DOI:10.1017/err.2019.8. [Online]. Available: https://www.cambridge.org/core/article/towards-intelligent-regulation-of-artificial-intelligence/AF1AD1940B70DB88D2B24202EE933F1B.

[52] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[53] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, A. Jacquet, A. 2. Jacquet, and A. I. Responsible, "73 responsible ai pattern catalogue: A collection of best practices for ai governance and engineering," *ACM Comput. Surv*, vol. 56, 2024. DOI: 10.1145/3626234. [Online]. Available: https://doi.org/10.1145/3626234.

[54] R. M. Unger, *The left alternative.* Verso Books, 2020.

[55] B. Rakova, J. Yang, H. Cramer, and R. Chowdhury, "Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–23, 2021.

[56] B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 10, no. 4, pp. 1–31, 2020.

[57] Cohere, OpenAI, and A. Labs, *Best practices for deploying language models*, Blog, 2022. [Online]. Available: https://openai.com/blog/best-practices-for-deploying-language-models/.

[58] A. Kloker, J. Fleiß, C. Koeth, T. Kloiber, P. Ratheiser, and S. Thalmann, "Caution or trust in ai? how to design xai in sensitive use cases?," 2022.

[59] L. Zhu, X. Xu, Q. Lu, G. Governatori, and J. Whittle, "Ai and ethics—operationalizing responsible ai," *Humanity Driven AI: Productivity, Well-being, Sustainability and Partnership*, pp. 15–33, 2022.

[60] P. Boza and T. Evgeniou, "Implementing ai principles: Frameworks, processes, and tools," 2021.

[61] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell, "Towards accountability for machine learning datasets: Practices from software engineering and infrastructure," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21, Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 560–575, ISBN: 9781450383097. DOI: 10.1145/3442188.3445918. [Online]. Available: https://doi.org/10.1145/3442188.3445918.

[62] H. Khlaaf, P. Mishkin, J. Achiam, G. Krueger, and M. Brundage, *A hazard analysis framework for code synthesis large language models*, 2022. arXiv: 2207.14157 [cs.SE].

[63] D. S. Brown, J. Schneider, A. Dragan, and S. Niekum, "Value alignment verification," in *International Conference on Machine Learning*, PMLR, 2021, pp. 1105–1115.

[64] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, "Red teaming language models with language models," *arXiv preprint arXiv:2202.03286*, 2022.

[65] F. Wotawa, "On the use of available testing methods for verification & validation of ai-based software and systems.," in *SafeAI@ AAAI*, 2021.

[66] M. Borg, C. Englund, K. Wnuk, B. Duran, C. Levandowski, S. Gao, Y. Tan, H. Kaijser, H. Lönn, and J. Törnqvist, "Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry," *arXiv preprint arXiv:1812.05389*, 2018.

[67] R.-J. Yew and D. Hadfield-Menell, "A penalty default approach to preemptive harm disclosure and mitigation for ai systems," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 823–830.

[68] J. Roski, E. J. Maier, K. Vigilante, E. A. Kane, and M. E. Matheny, "Enhancing trust in ai through industry self-governance," *Journal of the American Medical Informatics Association*, vol. 28, no. 7, pp. 1582–1590, 2021.

[69] G. Chaudhary, "Artificial intelligence: The liability paradox," *ILI Law Review*, 2020.

[70] J. K. Kingston, "Artificial intelligence and legal liability," in *Research and Development in Intelligent Systems XXXIII: Incorporating Applications and Innovations in Intelligent Systems XXIV 33*, Springer, 2016, pp. 269–279.

[71] M. E. Gerstner, "Liability issues with artificial intelligence software," *Santa Clara L. Rev.*, vol. 33, p. 239, 1993.

[72] P. H. Padovan, C. M. Martins, and C. Reed, "Black is the new orange: How to determine ai liability," *Artificial Intelligence and Law*, vol. 31, no. 1, pp. 133–167, Mar. 2023, ISSN: 1572-8382. DOI: 10.1007/s10506-022-09308-9. [Online]. Available: https://doi.org/10.1007/s10506-022-09308-9.

[73] M. E. Kaminski, "Binary governance: Lessons from the gdpr's approach to algorithmic accountability," *S. Cal. L. Rev.*, vol. 92, p. 1529, 2018.

[74] E. Aghajani, C. Nagy, O. L. Vega-Marquez, M. Linares-Vasquez, L. Moreno, G. Bavota, and M. Lanza, "Software documentation issues unveiled," *Proceedings - International Conference on Software Engineering*, vol. 2019-May, pp. 1199–1210, May 2019, ISSN: 02705257. DOI: 10.1109/ICSE.2019.00122.

[75] *European parliament resolution of 20 october 2020 with recommendations to the commission on a civil liability regime for artificial intelligence*, Oct. 2020.

[76] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, pp. 86–92, 12 2021, ISSN: 0001-0782.

[77] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," 2019, pp. 220–229. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3287560.3287596.

[78] R. Dobbe, T. K. Gilbert, and Y. Mintz, "Hard choices in artificial intelligence," *Artificial Intelligence*, vol. 300, p. 103 555, 2021.

[79] I. Solaiman, "The gradient of generative ai release: Methods and considerations," *ACM International Conference Proceeding Series*, pp. 111–122, Jun. 2023. DOI: 10.1145/3593013.3593981. [Online]. Available: https://dl.acm.org/doi/10.1145/3593013.3593981.

[80] S. Casper, C. Ezell, C. Siegmann, *et al.*, "Black-box access is insufficient for rigorous ai audits," vol. 1, 2024.

[81] R. Bommasani, D. A. Hudson, E. Adeli, *et al.*, "On the opportunities and risks of foundation models," Aug. 2021. [Online]. Available: https://arxiv.org/abs/2108.07258v3.

[82] R. Dwivedi, D. Dave, H. Naik, *et al.*, "Explainable ai (xai): Core ideas, techniques, and solutions," *ACM Comput. Surv.*, vol. 55, no. 9, Jan. 2023, ISSN: 0360-0300. DOI: 10.1145/3561048. [Online]. Available: https://doi.org/10.1145/3561048.

[83] T. Miller, *Explainable ai is dead, long live explainable ai! hypothesis-driven decision support*, 2023. arXiv: 2302.12389 [cs.AI].

[84] F. F.-H. Nah, R. Zheng, J. Cai, K. Siau, and L. Chen, "Generative ai and chatgpt: Applications, challenges, and ai-human collaboration," *Journal of Information Technology Case and Application Research*, vol. 25, pp. 277–304, 3 Jul. 2023, ISSN: 23336897. DOI: 10.1080/15228053.2023.2233814. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/15228053.2023.2233814.

[85] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, Mar. 2021. DOI: 10.1145/3442188.3445922. [Online]. Available: https://dl.acm.org/doi/10.1145/3442188.3445922.

[86] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, and A. Kumar, "Holistic evaluation of language models," *arXiv preprint arXiv:2211.09110*, 2022.

[87] I. Rahwan, M. Cebrian, N. Obradovich, *et al.*, "Machine behaviour," *Nature 2019 568:7753*, vol. 568, pp. 477–486, 7753 Apr. 2019, ISSN: 1476-4687. DOI: 10.1038/s41586-019-1138-y. [Online]. Available: https://www.nature.com/articles/s41586-019-1138-y.

[88] C. A. Zhang, S. Cho, and M. Vasarhelyi, "Explainable artificial intelligence (xai) in auditing," *International Journal of Accounting Information Systems*, vol. 46, p. 100 572, 2022.

[89] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, and H. Lakkaraju, "Openxai: Towards a transparent evaluation of model explanations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 15 784–15 799, 2022.

[90] N. V. Berkel, J. Opie, O. F. Ahmad, L. Lovat, D. Stoyanov, and A. Blandford, "Initial responses to false positives in ai-supported continuous interactions: A colonoscopy case study," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 12, 1 Mar. 2022, ISSN: 21606463. DOI: 10.1145/3480247. [Online]. Available: https://dl.acm.org/doi/10.1145/3480247.

[91] E. Bondi, R. Koster, H. Sheahan, M. Chadwick, Y. Bachrach, T. Cemgil, U. Paquet, and K. Dvijotham, "Role of human-ai interaction in selective prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 5286–5294, 5 Jun. 2022, ISSN: 2374-3468. DOI: 10.1609/AAAI.V36I5.20465. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/20465.

[92] N. V. Berkel, M. B. Skov, and J. Kjeldskov, "Human-ai interaction," *Interactions*, vol. 28, pp. 67–71, 6 Nov. 2021, ISSN: 15583449. DOI: 10.1145/3486941. [Online]. Available: https://dl.acm.org/doi/10.1145/3486941.

[93] M. Tahaei, M. Constantinides, D. Quercia, S. Kennedy, M. Muller, S. Stumpf, Q. V. Liao, R. Baeza-Yates, L. Aroyo, and J. Holbrook, "Human-centered responsible artificial intelligence: Current & future trends," 2023, pp. 1–4. [Online]. Available: https://arxiv.org/abs/2302.08157.

[94] W. Xu, M. J. Dainoff, L. Ge, and Z. Gao, "Transitioning to human interaction with ai systems: New challenges and opportunities for hci professionals to enable human-centered ai," 2022.

[95] M. Riveiro and S. Thill, ""that's (not) the output i expected!" on the role of end user expectations in creating explanations of ai systems," *Artificial Intelligence*, vol. 298, p. 103 507, Sep. 2021, ISSN: 0004-3702. DOI: 10.1016/J.ARTINT.2021.103507.

[96] C. Pelau, D. C. Dabija, and I. Ene, “What makes an ai device human-like? the role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry,” *Computers in Human Behavior*, vol. 122, p. 106 855, Sep. 2021, ISSN: 0747-5632. DOI: 10.1016/J.CHB.2021.106855.

[97] M. Price, “Open coding for machine learning,” 2022.

[98] K. Charmaz, “Constructing grounded theory,” pp. 1–416, 2014.

[99] S. S. Sundar, “Rise of machine agency: A framework for studying the psychology of human–ai interaction (haii),” *Journal of Computer-Mediated Communication*, vol. 25, pp. 74–88, 1 Mar. 2020, ISSN: 10836101. DOI: 10.1093/JCMC/ZMZ026. [Online]. Available: https://dx.doi.org/10.1093/jcmc/zmz026.

[100] S. Amershi, D. Weld, M. Vorvoreanu, *et al.*, “Guidelines for human-ai interaction,” *Conference on Human Factors in Computing Systems - Proceedings*, p. 13, May 2019. DOI: 10.1145/3290605.3300233. [Online]. Available: https://dl.acm.org/doi/10.1145/3290605.3300233.

[101] Q. Yang, A. Steinfeld, C. Rosé, and J. Zimmerman, “Re-examining whether, why, and how human-ai interaction is uniquely difficult to design,” *Conference on Human Factors in Computing Systems - Proceedings*, Apr. 2020. DOI: 10.1145/3313831.3376301. [Online]. Available: https://dl.acm.org/doi/10.1145/3313831.3376301.

[102] U. Ehsan and M. O. Riedl, “Human-centered explainable ai: Towards a reflective sociotechnical approach,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12424 LNCS, pp. 449–466, 2020, ISSN: 16113349. DOI: 10.1007/978-3-030-60117-1_33/TABLES/3. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-60117-1_33.

[103] M. Amirizaniani, J. Yao, A. Lavergne, E. S. Okada, T. Roosta, U. C. Berkeley, and C. Shah, “Developing a framework for auditing large language models using human-in-the-loop,” vol. 1, 2024.

[104] S. Shankar, J. D. Zamfirescu-Pereira, B. Hartmann, A. G. Parameswaran, and I. Arawjo, “Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences,” *Proceedings of UIST (Submitted for Review to UIST ’24)*, vol. 1, 2024. DOI: XXXXXXX.XXXXXXX.

[105] A. Bandi, P. V. S. R. Adapa, and Y. E. V. P. K. Kuchi, “The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges,” *Future Internet 2023, Vol. 15, Page 260*, vol. 15, p. 260, 8 Jul. 2023, ISSN: 1999-5903. DOI: 10.3390/FI15080260. [Online]. Available: https://www.mdpi.com/1999-5903/15/8/260/htm%20https://www.mdpi.com/1999-5903/15/8/260.

[106] S. Nikolaidis and J. Shah, "Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy," *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 33–40, 2013, ISSN: 21672148. DOI: 10.1109/HRI.2013.6483499.

[107] S. Nikolaidis, P. Lasota, R. Ramakrishnan, and J. Shah, "Improved human-robot team performance through cross-training, an approach inspired by human team training practices," *International Journal of Robotics Research*, vol. 34, pp. 1711–1730, 14 Dec. 2015, ISSN: 17413176. DOI: 10.1177/0278364915609673/ASSET/IMAGES/LARGE/10.1177_0278364915609673-FIG13.JPEG. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/0278364915609673.

[108] D. Pan, D. Zhao, Y. Pu, L. Wang, and Y. Zhang, "Use of cross-training in human–robot collaborative rescue," *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol. 34, pp. 261–276, 3 May 2024, ISSN: 1520-6564. DOI: 10.1002/HFM.21025. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/hfm.21025%20https://onlinelibrary.wiley.com/doi/abs/10.1002/hfm.21025%20https://onlinelibrary.wiley.com/doi/10.1002/hfm.21025.

[109] L. Zhou, P. A. Moreno-Casares, F. Martínez-Plumed, *et al.*, "Predictable artificial intelligence," Oct. 2023. [Online]. Available: https://arxiv.org/abs/2310.06167v1.

[110] D. Ganguli, D. Hernandez, L. Lovitt, *et al.*, "Predictability and surprise in large generative models," *ACM International Conference Proceeding Series*, pp. 1747–1764, Jun. 2022. DOI: 10.1145/3531146.3533229. [Online]. Available: https://dl.acm.org/doi/10.1145/3531146.3533229.

[111] G. Fink and M. Bishop, "Property-based testing," *ACM SIGSOFT Software Engineering Notes*, vol. 22, pp. 74–80, 4 Jul. 1997, ISSN: 0163-5948. DOI: 10.1145/263244.263267. [Online]. Available: https://dl.acm.org/doi/10.1145/263244.263267.

[112] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of nlp models with checklist," May 2020. [Online]. Available: https://arxiv.org/abs/2005.04118v1.

[113] S. E. Middleton, E. Letouzé, A. Hossaini, and A. Chapman, "Trust, regulation, and human-in-the-loop ai," *Communications of the ACM*, vol. 65, pp. 64–68, 4 Mar. 2022, ISSN: 15577317. DOI: 10.1145/3511597. [Online]. Available: https://dl.acm.org/doi/10.1145/3511597.

[114] S. Kandul, V. Micheli, J. Beck, T. Burri, F. Fleuret, M. Kneer, and M. Christen, "Human control redressed: Comparing ai and human predictability in a real-effort task," *Computers in Human Behavior Reports*, vol. 10, p. 100 290, May 2023, ISSN: 2451-9588. DOI: 10.1016/J.CHBR.2023.100290.

[115] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Beyond accuracy: The role of mental models in human-ai team performance," *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, pp. 2–11, Oct. 2019, ISSN: 2769-1349. DOI: 10.1609/HCOMP.V7I1.5285. [Online]. Available: https://ojs.aaai.org/index.php/HCOMP/article/view/5285.

[116] B. Nushi, E. Kamar, and E. Horvitz, "Towards accountable ai: Hybrid human-machine analyses for characterizing system failure," *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 6, pp. 126–135, Jun. 2018, ISSN: 2769-1349. DOI: 10.1609/HCOMP.V6I1.13337. [Online]. Available: https://ojs.aaai.org/index.php/HCOMP/article/view/13337.

[117] P. Hase and M. Bansal, "Evaluating explainable ai: Which algorithmic explanations help users predict model behavior?" *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 5540–5552, May 2020, ISSN: 0736587X. DOI: 10.18653/v1/2020.acl-main.491. [Online]. Available: https://arxiv.org/abs/2005.01831v1.

[118] A. J. Artino, "Bandura, ross, and ross: Observational learning and the bobo doll," *Online Submission*, Jan. 2007.

[119] T. L. Chartrand and R. van Baaren, "Chapter 5 human mimicry," *Advances in Experimental Social Psychology*, vol. 41, pp. 219–274, Jan. 2009, ISSN: 0065-2601. DOI: 10.1016/S0065-2601(08)00405-X.

[120] N. Gueguen, C. Jacob, and A. Martin, "Mimicry in social interaction: Its effect on human judgment and behavior," *European Journal of Social Sciences*, vol. 8, 2 2009.

[121] K. Darling, "'who's johnny?' anthropomorphic framing in human-robot interaction, integration, and policy," *SSRN Electronic Journal*, Mar. 2015. DOI: 10.2139/SSRN.2588669. [Online]. Available: https://papers.ssrn.com/abstract=2588669.

[122] Y. Feng, J. Qiang, Y. Li, Y. Yuan, and Y. Zhu, "Sentence simplification via large language models," 2023. [Online]. Available: https://github.com/BrettFyt/SS.

[123] M. U. Hadi, qasem al tashi, R. Qureshi, *et al.*, "A survey on large language models: Applications, challenges, limitations, and practical usage," *Authorea Preprints*, Oct. 2023. DOI: 10.36227/TECHRXIV.23589741.V1. [Online]. Available: https://www.authorea.com/doi/full/10.36227/techrxiv.23589741.v1?commit=b1cb46f5b0f749cf5f2f33806f7c124904c14967.

[124] J. Yang, U. H. Jin, X. Han, *et al.*, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, p. 32, 6 Apr. 2024, ISSN: 1556-4681. DOI: 10.1145/3649506. [Online]. Available: https://dl.acm.org/doi/10.1145/3649506.

[125] T. Kew, A. Chi, L. Vásquez-Rodríguez, S. Agrawal, D. Aumiller, F. Alva-Manchego, and M. Shardlow, "Bless: Benchmarking large language models on sentence simplification," 2023. [Online]. Available: https://chat.openai.com/.

[126] F. Alva-Manchego, L. Martin, A. Bordes, C. Scarton, B. Sagot, and L. Specia, "Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations," D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Association for Computational Linguistics, Jul. 2020, pp. 4668–4679. DOI: 10.18653/v1/2020.acl-main.424. [Online]. Available: https://aclanthology.org/2020.acl-main.424.

[127] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects institute for human and machine cognition," 2018.

[128] B. Cahour and J. F. Forzy, "Does projection into use improve trust and exploration? an example with a cruise control system," *Safety Science*, vol. 47, pp. 1260–1270, 9 Nov. 2009, ISSN: 0925-7535. DOI: 10.1016/J.SSCI.2009.03.015.

[129] B. D. Adams, L. E. Bruyn, and S. Houde, "Trust in automated systems literature review," 2003.

[130] OpenAI, *Embeddings - openai api*, 2024. [Online]. Available: https://platform.openai.com/docs/guides/embeddings.

[131] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[132] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "Optimizing statistical machine translation for text simplification," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 401–415, 2016.

[133] F. Alva-Manchego, L. Martin, C. Scarton, and L. Specia, "EASSE: Easier automatic sentence simplification evaluation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 49–54. DOI: 10.18653/v1/D19-3009. [Online]. Available: https://aclanthology.org/D19-3009.

[134] J. Klayman, "Varieties of confirmation bias," *Psychology of Learning and Motivation - Advances in Research and Theory*, vol. 32, pp. 385–418, C Jan. 1995, ISSN: 0079-7421. DOI: 10.1016/S0079-7421(08)60315-1.

[135] R. F. Flesch and A. J. Gould, "The art of readable writing," *(No Title)*, 1949.