Causal Framework of Artificial Autonomous Agent Responsibility

Matija Franklin
matija.franklin@ucl.ac.uk
University College London
London, United Kingdom

Hal Ashton
henry.ashton.17@ucl.ac.uk
University College London
London, United Kingdom

Edmond Awad
e.awad@exeter.ac.uk
University of Exeter
Exeter, United Kingdom

David Lagnado
d.lagnado@ucl.ac.uk
University College London
London, United Kingdom

ABSTRACT

Recent empirical work on people's attributions of responsibility toward artificial autonomous agents (such as Artificial Intelligence agents or robots) has delivered mixed findings. The conflicting results reflect differences in context, the roles of AI and human agents, and the domain of application. In this article, we outline a causal framework of responsibility attribution which integrates these findings. It outlines nine factors that influence responsibility attribution - *causality, role, knowledge, objective foreseeability, capability, intent, desire, autonomy, and character*. We propose a framework of responsibility that outlines the causal relationships between the nine factors and responsibility. To empirically test the framework we discuss some initial findings and outline an approach to using serious games for causal cognitive research on responsibility attribution. Specifically, we propose a game that uses a generative approach to creating different scenarios, in which participants can freely inspect different sources of information to make judgments about human and artificial autonomous agents.

CCS CONCEPTS

• **Computing methodologies** → **Cognitive science**; • **Security and privacy** → *Human and societal aspects of security and privacy*; • **Social and professional topics** → *Government technology policy*.

KEYWORDS

Blame, Responsibility, Causal Cognition, Attribution

ACM Reference Format:
Matija Franklin, Hal Ashton, Edmond Awad, and David Lagnado. 2022. Causal Framework of Artificial Autonomous Agent Responsibility. In *Proceedings of 5th AAAI/ACM Conference on AI, Ethics, and Society (AIES '22).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3514094.3534140

AIES '22, August 01–03, 2022, Oxford, UK

ACM ISBN 978-1-4503-9247-1/22/08…$15.00
https://doi.org/10.1145/3514094.3534140

1 INTRODUCTION

Recent technological advances are leading to a future in which machines have new societal roles. In these new roles, artificial autonomous agents (henceforth A-bots, used interchangeably with AI or machines) will mediate and govern human relationships, making decisions on the distribution of resources and risks among humans, potentially having a substantial impact on our lives. Some of today's examples include Artificial Intelligence (AI) contributing to decisions on parole [77], hiring [81], and health care [12].

These new roles assume new responsibilities and may change how responsibility is attributed when something goes wrong. It may also complicate how responsibility is distributed among the different actors involved in decision-making, to a degree that it becomes unclear who should be blamed. Some have called this the "AI Responsibility Gap" [71], or the "moral crumple zone" [24]. Furthermore, a machine's performance may be evaluated against novel standards that differ from those used to evaluate humans in the same role. For example, an autonomous vehicle (AV) that hits a pedestrian will be blamed more than a human driver in a similar situation [37]. While it can be understandable in some situations that humans and machines are making these decisions differently, it remains unclear whether a machine is seen as a human adult, a baby, an object, an animal, or something else [38]. Understanding how humans perceive machines and think about their mental states will be critical to understand [39].

An important part of this problem is the cognitive factor. The public's views and trust toward machines majorly predicts their adoption [17, 30, 61, 62]. Understanding people's attitudes is thus critical for addressing any potential concerns [16, 20, 84, 85, 92]. For example, people require AVs to be much safer than human drivers [66, 86]. Moreover, negative public reactions may result in inflated prices [31] and may shape how a tort-based regulatory scheme would turn out, both of which can influence the rate of adoption [9].

Recently, empirical studies have been conducted on how people assign responsibility for A-bots [8, 41, 94]. However, these experiments deliver mixed findings, in some cases showing that people prefer to blame humans rather than A-bots, in other cases that they assign greater blame to the A-bots. These conflicting results reflect differences in context, the roles of A-bots and human agents, the perceived capacity and level of autonomy of A-bots, and possibly the domain itself. However, there has been no attempt to integrate these findings into a systematic framework of attribution.

Another shortcoming of prior empirical research is an over-focus on simple comparisons between humans and A-bots (e.g., [41, 94]), rather than looking at the broader system that includes both humans and A-bots interacting, possibly including several different types of human agents (designer, deployer, user, annotator, auditor) and different types of A-bots (algorithms, AIs, systems, robots). It neglects the complexity of real-world situations, where several parties are involved with the A-bots, each with varying degrees of responsibility.

To address this complexity we outline the methods and requirements of a retrospective causal model of responsibility. Rooted in models from philosophy, law, and psychological research, we propose a framework that captures the concepts that underpin attributions of responsibility [59, 69, 97]. Here, responsibility relates to the concept of outcome responsibility which looks at actions and outcomes that occurred in the past for which an agent is either blameworthy or praiseworthy [76]. The framework aims to explain and predict people's attributions. It is constructed on the following core concepts: *causality, role, knowledge, objective foreseeability, capability, intent, desire, autonomy and character* (see Figure 1). People's application of these concepts to attributing responsibility towards A-bots is rooted in their prior knowledge of human agents interacting in social systems [44]. We also present some initial findings that providence evidence for our proposed causal framework. This case study focuses on the role objective foreseeability has on responsibility attributions. To further empirically test the framework we also proposes novel methodologies in serious games and apply them to the study of attribution.

2 BACKGROUND

2.1 Responsibility attribution

Over millennia people have developed a rich system of responsibility attribution, allowing us to hold individuals responsible for their actions, and assign praise and blame in complex social situations [72, 79]. A crucial precursor to assessing responsibility is our ability to build causal models to explain how and why people do things, to infer the hidden beliefs and intentions behind their actions, and to tell a story that makes sense of social behavior [4]. We typically blame someone for intentionally (or carelessly) causing harm but excuse them otherwise; however, there is significant cross-cultural variation in intention's effect on blame [11]. Similarly, legal systems have developed a complex set of concepts, such as intent, foreseeability, negligence, and reasonable care, to assign culpability or liability to human agents and corporations. These too usually depend on establishing a causal link between an agent's actions and consequent harm and showing that the agent intended or foresaw the outcome (or that they ought reasonably to have done so).

The advent of autonomous systems presents several challenges to these frameworks. AIs do not fit naturally into our standard classifications of what it is to be an acting and knowing agent. While AIs are capable of learning and making autonomous decisions, they lack typical human qualities such as intent, integrity, and moral sense. To tackle this problem, various proposals have been made as to how to fit AIs into our current schemes – treating them as fully-fledged agents, animal-like, or mechanical products, with differing implications for their legal and moral status [32, 64, 93].

Another classic issue in responsibility attribution is the problem of many hands, where multiple different parties combine to bring about an outcome [90]. How should causality and responsibility be distributed across this complex interacting system, to ensure fair and proportionate blame? This problem is exacerbated when the parties have different roles, levels of skill, and perhaps differing knowledge about each other.

A-bots presents novel challenges because there are various different types of principal-agent relationships and opportunities for various types of feedback loops [42, 47]. For example, the principal can have varying levels of oversight over the agent, in terms of the explicitness of their orders and instructions, the monitoring of the agent's progress, and how they might intervene in the agent's task. Likewise, the agent can have various ways it carries out its task, and how it responds to the principal's guidance or interference.

2.2 Artificial Autonomous Agent responsibility

Although A-bots may have causal efficacy that can produce states of affairs, it is not clear who should be held responsible [45]. In recent years, these issues have been investigated empirically. Current research has identified certain patterns in people's judgments of AI. Hidalgo et al. [41] report that 1) people tend to judge humans more for their intentions and machines more for the outcomes of their actions; 2) people infer more extreme intentions to humans and narrow intentions to machines. People are more willing to excuse humans for accidents than machines. Further, machines are judged more harshly for scenarios involving physical harm, while humans are for scenarios involving unfairness. Finally, they found that people are more likely to centralize responsibility up the chain of command for machine mistakes.

Research suggests that role and causality impact blame attributions. When a car is being collaboratively driven under the shared control of a human and artificial autonomous driver, less blame and causality was attributed to the machine when both drivers made an error [8]. There was no difference in blame in a scenario where one driver had the role of driving, and the other was allowed to intervene. A similar finding emerges when an A-bots take the role of an advisor - doctors advised by AIs were judged as more blameworthy than those advised by human doctors [99]. This finding held when the AI adviser had the ability to intervene in the doctor's decisions. Finally, AI agents, compared to human agents, were expected to make utilitarian moral choices and got blamed more for not doing so [70].

Increased A-bot autonomy also increases blame. First, higher machine autonomy is associated with intent inferences towards A-bot being more similar to that of humans [10]. This is supported by research showing that when robots are described as autonomous, participants attribute nearly as much blame to them as they do to humans [28]. Further, more autonomous technologies decrease the perceived control a user has over it, which in turn decreases the praise the user receives for positive outcomes [46]. Finally, drivers of manually controlled vehicles are deemed more responsible than drivers of automated vehicles (AVs) [73].

Finally, research shows that people's expectations of A-bots' capabilities form consistent patterns. First, people generally expect automation to be perfect and people to be imperfect [67]. People

rely more on algorithmic advice as task difficulty increases [15]. Second, people refuse to use A-bots for making moral decisions [21]. This aversion is mediated by perceptions that machines cannot fully think or feel [14]. It may also be due to people's perceptions of A-bots as selfish and uncooperative [43]. Finally, people do not trust AIs to deal with emotions [98], and rely less on A-bots for tasks that seem subjective rather than objective [18].

2.3 Legal perspectives

It is the current position that Autonomous AI agents are not legal persons and therefore cannot commit crimes and bear any responsibility for any damages that they might cause. That falls instead on the owner of the AI system but there is a growing debate about what sort of liability should apply [95]. At one extreme, there are those who argue for a strict-negligence regime where proof of harm is the only thing that should be considered [63]. These arguments often rely on the concept of respondeat superior, which historically made harms caused by animals and slaves the responsibility of their owner and more recently applies to the harms of employees caused in the course of their job. In some cases, this approach seems inequitable as in the case of the Swiss Random Darknet shopper [56] which was programmed to buy items at random from the darknet as part of an art installation. When the algorithm bought some illegal drugs, the artist creator was initially charged with a drugs possession offense. After the design and purpose algorithm was explained to the police, charges were dropped. Abbott [1] argues that negligence should be the standard since strict liability is too draconian. Civil negligence requires it to be shown that on the balance of probabilities, a reasonable actor would not have caused the harm. This poses further questions regarding what standard of reasonability should be considered since what is reasonable for A-bot might not agree with the standard for a human.

The types of factors which civil and criminal consider when deciding responsibility, are instructive. Firstly there are questions of causality to be answered, both in the physical or 'factual' sense which typically corresponds to a but-for analysis of causation and a more stylized legal sense. This concept of legal causation is related to the folk-judgment of causality and is especially concerned with the 'proximity' of the cause to the outcome [59]. In particular, the presence of a novus actus interveniens, or intervening action committed by a third party absolves the actor of wrongdoing. As well as causation of harm through action (and occasionally omission of action) also known as actus reus, Criminal law requires an assessment of the mental state of the accused at the point of commission. It has defined different standards of mens rea (guilty mind) which correspond to different levels of culpability for any particular harm. The US Model Penal Code mentions four levels of culpability, which are in descending gravity: Purpose, Knowledge, Recklessness, and Negligence. Crimes committed with Purpose and Knowledge are done so with knowledge concerning the likely consequence of the agent's actions with Purpose crimes requiring a desire or aim on the part of the perpetrator for the outcome. Recklessness and Negligence require external standards of good conduct to be applied against the behavior of the accused. Recklessness requires the actor to unreasonably disregard the danger of acting in the way they did. In sum, the law ascribes responsibility for harm by considering causation, intent to commit harm, foreseeability of the harmful outcome occurring, and by applying an objective standard as to whether the actions of the accused were reasonable. Experimental work has been conducted to marry these four levels of culpability with lay judgments of blame [87]. Recently the study of mens rea judgments has begun to be extended towards robots [91].

3 TOWARDS A FRAMEWORK OF RESPONSIBILITY ATTRIBUTION

Drawing inspiration from philosophy [97], law [23], and psychology [4, 69] we propose an overarching explanatory framework that captures the core concepts that underpin judgments of responsibility. Rather than simply describe patterns of attribution we seek to build a framework with the power to predict and explain people's judgments. The framework will be constructed around several core concepts that are described below.

3.0.1 ***Causality.*** Causality is a crucial concept in both everyday and legal reasoning, and a precursor to attributing responsibility. However, an agent can cause an outcome but not be blamed for it. To give an example, although A-bot can cause an event, the A-bot user may receive the blame. One framework proposes that when people assign blame they prospectively determine an agent's criticality - the extent to which an outcome is dependent on an agent's action [57]. Further, people retrospectively determine an agent's pivotality - its causal contribution to an outcome. When allocating blame amongst multiple agents, attributions are sensitive to an agent's pivotality [34]. Finally, an agent's location in a causal chain influences causal inference - agents causing later events are seen as more causal and blameworthy [58]. This will impact how an A-bot is judged when executing an action directly, versus when giving advice to a human who executes the final action. This will be further discussed in relation to role.

Advances in computer science have led to a comprehensive formal framework for causal inference [40, 75], and we will use this framework as a basis for exploring people's causal judgments [34, 60, 89]. The framework allows us to model complex interactions between human agents and A-bots, and also to capture the mental states of the agents, as well as prior histories, capacities, and competencies [35]. Thus far, the causal modeling approach has mostly been applied to physical or social systems, but we will extend it to include AIs. The framework allows us to model counterfactual inferences, which are crucial in establishing causation and responsibility.

3.0.2 ***Role.*** Role attributions relate to criticality - agents are responsible for carrying out actions according to their role [57]. In group and organizational contexts, people can take a variety of different roles, performing tasks of differing importance and requiring different skills and competence [36]. Often there is a hierarchy of control, such that higher-level agents take on more responsibility than lower-level agents in the chain (e.g., employer versus employee relations). How does role responsibility affect people's judgments, and how does this translate to systems that include both humans and AIs, possibly with different levels of competence and skill?

The effect role has on responsibility attribution is sensitive to the specific type of human-AI interaction [82]. Current frameworks propose that an A-bots can take several roles in interacting with humans; namely: advisor (i.e., human receiving advice from an A-bot), partner (i.e., human collaborating with A-bot), delegatee (i.e., human delegating to, and overseeing an A-bot), role model (i.e., humans imitating A-bots) [54]. Ashton and Franklin [5] further propose two further roles - A-bot as a boss and A-bot as an adversary.

3.0.3 ***Knowledge.*** Knowledge relates to the degree to which an agent has the knowledge to subjectivley foresees the outcomes of its actions [3]. Agents are blamed more for highly foreseeable mistakes, and less for extremely unlikely events [50, 58]. People receive more blame when they bring about the harm that was predictable and is not held to account for extremely unlikely eventualities. In legal contexts, predictability is assessed relative to a reasonable person standard, with people being held liable or responsible for harm if a reasonable person could have predicted the outcome. AI complicates the relationship between knowledge and blame in at least two ways. First, what reasonable knowledge and thus forseeability should people ascribe to an AI, when its information processing is in a "black box" (this is not necessarily always the case for A-bots). Second, given AI's autonomy, what reasonable foreseeability will people ascribe to a person deploying or developing an AI?

3.0.4 ***Objective foreseeability.*** Objective Foreseeability represents how likely an outcome really is, irrespective of an agent's expectations about what is going to happen [58]. It thus differs from *subjective foreseeability* - the likeliness of an event from an agent's point of view - and *reasonable foreseeability* - what would be reasonable for an agent to expect given the available information. Both objective and subjective foreseeability have been shown to have an effect of people's blame attributions [58]. In relation to AI, one could argue that subjective and reasonable foreseeability is evidenced by how well the AI is evaluated on benchmarks during development.

3.0.5 ***Capability.*** Capability refers to an agents' competencies and skills. Expectations of an agent's skill influence blame attributions for the outcomes it produces, with high prior expectations resulting in more blame for negative outcomes when an agent underperforms [35]. In other words, people tend to blame skilled and capable human agents more than unskilled ones. In turn, low prior expectations result in more praise for a positive outcome when it exceeds its performance [33].

Recent research suggests that people's judgments of blame are sensitive to the different levels of complexity of AI [53], but these questions have not been tested in broader human-AI systems. Given that AI often exceeds human ability in specific tasks, it may be blamed more for its mistakes. However, one study found that although some situations are more difficult for AVs (such as those requiring commons sense) and others for human drivers (such as those requiring fast reaction times), people's blame judgments were higher for AVs, irrespective of the context of the mistake [27].

3.0.6 ***Intent.*** Intent influences blame attributions because they allow one to distinguish between the effects an agent did or did not intend [51]. People rate intentional actions as more blameworthy than unintentional actions [58]. Although intent has traditionally been ascribed to human agents, as well as certain group agents such as corporations, the autonomous behavior of A-bot agents may encourage people to infer intent to them [65]. Alternatively, people may infer intent towards the AI's user [45].

Intent has traditionally been ascribed only to human agents, although also extended to certain groups such as corporations. However, the increasing sophistication of new AI approaches that make autonomous decisions, might encourage people to assign some degree of intent to A-bots. Research in experimental jurisprudence suggests that this is the case, with people ascribing equal amounts of intent to humans and A-bots [6]. But perhaps more importantly, the use of highly sophisticated AI might change our attributions to human agents in the broader human-AI system. For example, if an AI system undertakes a task in a novel way (not predicted by the human user or even designer) we might expect this to reduce the responsibility assigned to the human (and thus make them less liable). More broadly, we also need to consider ascriptions of intention across multiple human and AI systems: for example, when a contractor, designer, and AI system combine to produce an outcome (Johnson and Verdicchio [45] introduce the term 'triadic agency' for this context).

3.0.7 ***Desire or Aim.*** Desires have been treated as conceptually separate from intentions in that 1) intentions involve committing to performing the intended actions, while desires do not, 2) intentions are based on reasoning, while desires are an input to this reasoning, and 3) desires can be directed, while intentions are directed at the intender's own actions [68]. A further distinction comes from legal ideas surrounding the two concepts, establishing that outcomes which are not desired can be intended [100]. Like intention, desire influences blame attributions [19]. In relation to autonomous artificial agents, desires may be better thought of as aims, as it avoids anthropomorphism.

3.0.8 ***Autonomy.*** Autonomy is when an agent is able to make its own decision, without the control of another agent. People attribute blame in terms of an agent's control over the outcome of its actions [3]. It thus relates to the extent to which an agent's behavior is purposeful and the extent to which an agent knows what it's doing. Perceived autonomy is also associated with perceived intent, in that more autonomous agents are seen as more intentional [10]. Attributions of autonomy to A-bots agents will depend on whether or not people perceive it as controlled by its developers and users. An A-bot getting an initial order, and then executing a task independently will be viewed as more autonomous than an A-bot that needs to seek permision in order to execute tasks.

3.0.9 ***Character.*** Character refers to an agent's moral character [78]. In relation to responsibility, is the extent to which the agent is "responsible" or the opposite - "irresponsible" [96]. Character thus capture an aspect of agents that is outside of their capacity to be responsible. Inferences about moral character affect both blame and praise judgments [88]. Further inferring bad character amplified the effect the severity of outcome (i.e., consequences) would have on blame and praise judgments. In an experiment where either an A-bot or human engaged in either virtuous or vicious behavior, the participants showed weakened moral attributions for A-bots

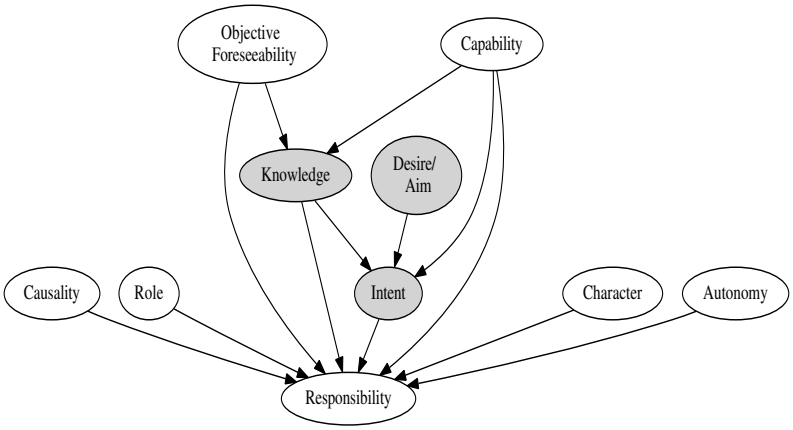

Figure 1: Causal Framework of Responsibility: Causal diagram showing the relationship between responsibility, causality, role, objective foreseeability, capability, knowledge, desire/aim, intent, character, autonomy. Factors in grey represent *mental states*

compared to humans [29]. Thus it may be the case that an A-bot's character influenced responsibility judgments less than human character. It may also be the case that an A-bot's character is more related to the perceived character of its developers.

3.1 Causal relationships

We propose a Causal Framework of Responsibility (see Figure 1) - a Directed Acyclic Graph which outlines the causal relationship between the nine aforementioned factors and responsibility. As well as investigating how each of these components affect attributions of responsibility, our framework also outlines the interrelations between these different components. These relations will drawn from the previously discussed empirical evidence. Causality, role, character, and autonomy independently have causal influence over responsibility. Objective foreseeability influences both responsibility, and knowledge, which in turn also influences responsibility and intent. As previously discussed, desire influences intent, which in turn influences responsibility. Finally, capability influences knowledge, intent and responsibility.

We view the proposed interrelations in our Causal Framework of Responsibility as a starting point. The framework will be refined and revised in light of new conceptual and empirical work. For example, the relation between intent and knowledge is complex: intentional actions are often accompanied by high knowledge but these can sometimes come apart (for example, when someone undertakes a highly speculative course of action). The methods used for the framework validation will be further discussed.

By articulating the relationships between these components in a causal framework we are better placed to investigate how they might combine and interact, drawing on formal tools from causal modeling. Rather than simply generate lists of factors we can explore the causal mechanisms and pathways that lead to responsibility attributions. Crucially, developing such a causal framework serves as both a predictive and explanatory tool. For example, if an agent intentionally brings about an adverse outcome we can predict high levels of blame, but this will be modulated by the agent's degree of knowledge. Moreover, if an agent brings about an adverse outcome with high foresight and capability, we will infer a high level of intent (in the absence of other information) [49].

4 FRAMEWORK VALIDATION

The causal modeling framework gives us a starting point for empirical work on responsibility attribution towards AI. Rather than just focus on how we ought to categorize AIs for legal or moral purposes, we will focus on how people (both the public and expert users) actually incorporate AIs into their conceptual framework, and how this shapes judgments of responsibility. We argue that this work provides a critical stepping stone for developing new methods to enhance people's trust and usage of AIs.

In this section we will outline methods for developing and testing the framework by researching individual factors, as well as multiple factors with the use of serious games [22]. We will focus on key areas where questions of responsibility arise in AI, both for expert users and laypeople, across several domains, including financial and medical decision making, and the use of smart systems in the household and workplace. We have selected these domains because they are rapidly expanding in their deployment of AI, and the general public is increasingly faced with questions of usage and trust.

4.1 Exploring individual factors

We propose for the use of simulation experiments to investigate how people's responsibility judgments are influenced by key factors outlined above. Particular focus should be directed towards multiple-agent situations that include humans and AIs, to see how these components trade-off and shape how people distribute responsibility across the different agents and components in the system.

These studies can be conducted online by recruiting members of the general population. We propose for a standard procedure in Causal Cognition research, where participants are first given a background scenario, such as the description of a medical decision-making context with both an AI diagnostic system and a doctor. Participants are then probed for their understanding and mental models of the set-up (using both verbal questions and a novel drawing approach for eliciting people's causal models [49]), and asked for prospective judgments of knowledge, capability, and role. In the next stage participants are given outcome information (e.g., a positive or negative health outcome) and then asked to rate each agent in the system on various dimensions including causal contribution, responsibility, and intent. Participants should also be asked to give narrative explanations of the scenarios and their reason for their responsibility judgments.

Using this process across a series of experiments allows one to manipulate the key factors from the explanatory framework, such as the level of autonomous decision-making of the AIs, the degree of foreseeability of the outcomes, and the role relations between different agents. One can also explore how people's prior models of the scenarios affect their subsequent responsibility judgments.

Based on the findings from the simulation studies we will further develop the experimental materials so that we can conduct a massive online experiment in the form of a serious game.

Case study: Objective foreseeability. A-bots, once developed and deployed, behave in ways that cannot be predicted by their developers and users. Our pilot study used four experimental groups to explore people's judgments of 1) AI's that were predictable; 2) AI's that were unpredictable; 3) Humans that were predictable; 4) Humans that were unpredictable. Participants (N= 273) observed as the agents performed actions that were either expected or unexpected for a given situation and produced outcomes that were either intended by the user or not intended by the user. The user was someone who employed to human or machine agent to work on a task.

Participants were given scenarios where a "senior trader" employs a human or AI to invest in the stock market. In each scenario, the agents choose between two stocks, with one being the more expected choice, and the other being less expected. The study was comprised of three phases. During the training phase, participants were introduced to the agents and given validation questions which served to confirm whether the participants understood the agent they would be judging - predictable or unpredictable, human or AI. During the test phase. Participants gave prospective judgments about an agent's foresight and capability for a given situation. Once they saw the outcomes of the agent's actions, they gave retrospective judgments about an agent's causality, responsibility, and intent. Finally, in the evaluation phase, participants made judgments about the agent's autonomy and role. All judgments were made on on 100-point analogue scales.

Within-subject ANOVAs revealed that, compared to the users, machine and predictable agents were attributed more responsibility [$F(4,269)=4.86$, $p=.028$]. Further, the results of a multiple linear regression that predicted judgments of an agent's responsibility found that intent and foresight were significant predictors for blaming human agents [$F(267,6)=108.93$, $p<.001$, $R2=.841$], while role and capacity were significant for machine agents [$F(267,6)=164.67$, $p<.001$, $R2=.88$]. Capability and intent were significant predictors of blame towards predictable agents [$F(267,6)=154.76$, $p<.001$, $R2=.881$], while foresight was for unpredictable agents [$F(267,6)=121.44$, $p<.001$, $R2=.838$]. The results thus build on previous research finding that intent was crucial for judging humans and outcome was crucial for judging machines [41].

Altogether, the pilot study indicates that participants are willing to make these inferences and attributions towards AI agents. We have identified initial patterns that are consistent with previous research. Finally, the initial findings were consistent with our proposed framework.

4.2 The Blame Game: A serious games approach to studying responsibility attributions

Serious games are games which have an aim other than entertainment, such as delivering an intervention, educating, or collecting data for scientific exploration [22]. Although the concept of serious games was first coined in the 1970s [2], the use of serious games in research has been growing rapidly. Laamarti et al. [55] outline a taxonomy of serious games, including: 1) type of activity performed by the player; 2) the sensory modalities experienced by the player in the game; 3) the interaction of the player with the game; 4) the environment of the digital game; 5) the application domain of the game, which relates to the games goal (e.g., education). Recent research in cognitive science has embraced the serious games approach towards delivering interventions and collecting data in *massive online experiments* [83].

An example for the use of serious games for the purpose of social science research, especially with respect to people's judgment of AI comes from the Moral Machine Experiment [7]. The moral machine platform used serious game elements to gather data on people's decision relating to moral decisions made by self-driving cars. People were shown moral dilemmas that a self-driving car would face. People had two judge which outcome they thought was more acceptable. An important aspect of the Moral Machine Experiment is that it used a generative approach to developing scenarios. Each scenario featured components from a larger set of components. For example, on the left side of the road a scenario could feature a man, a pregnant woman or a cat. The scenarios were generated using randomisation of these components so that the scenarios could explore how species, social value, gender, age, fitness, and utilitarianism influenced judgment. The components were presented visually, but participants could also access text-based descriptions.

Inspired by the Moral Machine example, we designed a serious game, *The Blame Game*, to collect data for this project. In *The Blame Game* players take the role of a detective trying to solve a case where an A-bot has committed a crime. The goal of the game is to collect information about the case and to make a judgment about who is to blame. In this section we outline our approach to using serious games and a generative game design to study people's judgments of responsibility. Specifically, we outline four separate criteria for using a causal cognitive approach for serious games towards gathering data to develop a framework of responsibility attribution towards AI.

First, the game will use a *visual novels* format. Visual novels are interactive games which feature a text and an image-based story. They involve minimal gameplay - players can respond to situations by clicking to keep the text and images moving while making some choices along the way. The benefit of this approach is that visual novels are easy to develop, given that they require only image and text, whilst also being more engaging than purely text-based games. The choices players make along the way can be used to collect data on choices and judgments.

Second, we use a generative approach towards developing the visual novel scenarios. Specifically, this involves generating images and text that describe different factors (character, intent, capability, etc) and agents. Players can receive information about factors either in the form of evidence they can analyse or text-based descriptions they can read.

Third, players get to choose what information they would like to learn, and the order in which they do so. Participants are allowed to make their judgments about who is responsible at any point, even if they have not inspected all of the available information. The benefit of this approach is that it gives rich data on what type of information people prefer when making a responsibility judgment

depending on the case they are evaluating. Further, this would allow for exploring how different combinations of inspected factors uniquely influence responsibility judgments. By giving people this freedom, the game will also serve the purpose of education through engaging people with questions pertaining to AI ethics.

Finally, in order to capture context-effect, the game will use a series of different domains (e.g., healthcare, self-driving cars, algorithmic trading). Further, players should interact with both A-bot and human scenarios. This allows for answering question relating to what factors uniquely influence people's judgments towards AIs versus humans.

5 DISCUSSION

This project aims to address the issue of mixed findings in empirical studies on how people assign responsibility to AI. It outlines nine factors that influence responsibility attribution - causality, role, knowledge, objective foreseeability, capability, intent, desire, autonomy and character. We propose a causal framework of responsibility - a Directed Acyclic Graph (DAG) which outlines the causal relationships between the nine factors and responsibility. To empirically test the framework we proposed our approach to using serious games and applying them to causal cognitive research. Specifically, we propose a visual novel game that uses a generative approach to creating different scenarios, in which participants can freely inspect different sources of information to make judgments about humans and A-bots.

Our proposed framework has several implications. By building a cognitive model of responsibility based on psychological evidence, one can provide foundational work for systems that provide automated oversight for A-bots [25]. Specifically, it could inform the development of systems to monitor deployed AIs "in the wild" and inform the creation of policy and regulation for AI governance [80].

The research also has implications for human-AI interaction, in that through understanding the responsibility model that people use when judging AI, one can improve trust in AI. Specifically, psychological research on how people judge AI can be used for improving Explainable AI – models that provide information about why AI acted the way it did [74].

AI responsibility poses many questions and challenges for AI developers. By using the framework to understand people's inferences and attributions, developers will be able to predict people's judgments towards A-bots in novel situations. This has practical applications for identifying the key factors that may lead harsh judgment, and a corresponding decrease in use.

The proposed serious games paradigm has implications for research in causal cognition aiming to systematically understand how different factors interrelate to form judgments, rather than just listing them individually. Our research has the potential to both replicate and contextualise findings in the field, and father new evidence for people's judgments of A-bots.

If the data is collected from a (nationally or internationally) representative sample of the public, the framework can serve as a tool for policymaking. Policymakers usually rely on prescriptive moral judgments, but in some situations they are interested in the opinion of the public. In such cases, our framework can be used as a prediction tool of the public's attribution in several situations. Moreover, it can be integrated into futuristic regulation frameworks that are proposed to cope with the pace of technological advances in AI, such as "adaptive regulation," which are regulatory frameworks designed to be updated automatically [13].

We believe our proposed framework can be extended in several ways to increase its generalisability and practicality. One simplifying assumption we made is that the data will be collected from a homogeneous sample of individuals. This can make sense in some situations. However, a group of individuals drawn form different cultures (societies, countries) or comprised of sub-populations that are characterised by demographic features (e.g., age, gender, race) are better represented by frameworks that deal with heterogeneity. A potential extension can be in employing hierarchical models in which individuals from e.g., one country are drawn from the same distribution [48, 52].

Another possible extension would deal with longitudinal data (i.e., across time). The proposed model assumes that the data is cross sectional (taken at one point in time). This is also a reasonable simplifying assumption. However, public perception is adaptive and can change over time (whether because the role or influence of existing factors have changed or because new factors have become relevant). Developing a framework with adaptive structure would be a big challenge, but an adaptive framework with a fixed structure can be developed as an extension to automatically adapt with newly-acquired data.

6 CONCLUSION

Given the increasingly high degree that AIs are becoming embedded in modern life, our proposed framework and research has the potential for long-term impact in many areas of benefit to society. By improving the trust and responsible use of AIs, we can help these systems deliver benefits for health and well-being, financial stability, sustainability, and justice. As AIs become more pervasive we must ensure that proper legal, ethical, and social oversight is in place. This includes the public's oversight, how we (the public) can contribute to keeping those systems in check. The data we collect could, in principle, be used to empower and enable public oversight through the development of "oversight algorithms" whose function would be to monitor, audit, and hold AI programs accountable [25, 26].

REFERENCES

[1] Ryan Abbott. 2020. *The reasonable robot: artificial intelligence and the law.* Cambridge University Press.

[2] CC Abt. 1970. Serious Games. *New York City, New York, USA, 1st edition* (1970).

[3] Mark D Alicke. 2000. Culpable control and the psychology of blame. *Psychological bulletin* 126, 4 (2000), 556.

[4] Mark D Alicke, David R Mandel, Denis J Hilton, Tobias Gerstenberg, and David A Lagnado. 2015. Causal conceptions in social explanation and moral evaluation: A historical tour. *Perspectives on Psychological Science* 10, 6 (2015), 790–812.

[5] Hal Ashton and Matija Franklin. 2022. The corrupting influence of AI as boss and adversary. *Unpublished Manuscript* (2022).

[6] Hal Ashton, Matija Franklin, and David Lagnado. 2022. Testing a definition of intent for AI in a legal setting. *Unpublished Manuscript* (2022).

[7] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.

[8] Edmond Awad, Sydney Levine, Max Kleiman-Weiner, Sohan Dsouza, Joshua B Tenenbaum, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. Blaming humans in autonomous vehicle accidents: Shared responsibility across levels of automation. *arXiv preprint arXiv:1803.07170* (2018).

[9] Edmond Awad, Sydney Levine, Max Kleiman-Weiner, Sohan Dsouza, Joshua B Tenenbaum, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2020. Drivers are blamed more than their automated cars when both make mistakes. *Nature human behaviour* 4, 2 (2020), 134–143.

[10] Jaime Banks. 2019. A perceived moral agency scale: development and validation of a metric for humans and social machines. *Computers in Human Behavior* 90 (2019), 363–371.

[11] H Clark Barrett, Alexander Bolyanatz, Alyssa N Crittenden, Daniel MT Fessler, Simon Fitzpatrick, Michael Gurven, Joseph Henrich, Martin Kanovsky, Geoff Kushnick, Anne Pisor, et al. 2016. Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences* 113, 17 (2016), 4688–4693.

[12] David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health affairs* 33, 7 (2014), 1123–1131.

[13] Lori S Bennear and Jonathan B Wiener. 2019. Adaptive Regulation: Instrument Choice for Policy Learning over Time. *Obtenido de Universidad de Harvard: https://www. hks. harvard. edu* (2019).

[14] Yochanan E Bigman and Kurt Gray. 2018. People are averse to machines making moral decisions. *Cognition* 181 (2018), 21–34.

[15] Eric Bogert, Aaron Schecter, and Richard T Watson. 2021. Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports* 11, 1 (2021), 1–9.

[16] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2020. The Moral Psychology of AI and the Ethical Opt-Out Problem. In *Ethics of Artificial Intelligence*. Oxford University Press, 109–126. https://doi.org/10.1093/oso/9780190905033.003.0004

[17] Lemuria Carter and France Bélanger. 2005. The utilization of e-government services: citizen trust, innovation and acceptance factors. *Information systems journal* 15, 1 (2005), 5–25.

[18] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825.

[19] Fiery Cushman. 2008. Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108, 2 (2008), 353–380.

[20] Benedict GC Dellaert, Suzanne B Shu, Theo A Arentze, Tom Baker, Kristin Diehl, Bas Donkers, Nathanael J Fast, Gerald Häubl, Heidi Johnson, Uma R Karmarkar, et al. 2020. Consumer decisions with artificially intelligent voice assistants. *Marketing Letters* 31, 4 (2020), 335–347.

[21] Berkeley J Dietvorst and Daniel M Bartels. 2021. Consumers object to algorithms making morally relevant tradeoffs because of algorithms' consequentialist decision strategies. *Journal of Consumer Psychology* (2021).

[22] Ralf Dörner, Stefan Göbel, Wolfgang Effelsberg, and Josef Wiemeyer. 2016. *Serious games*. Springer.

[23] Robin Antony Duff. 2007. *Answering for crime: Responsibility and liability in the criminal law*. Bloomsbury Publishing.

[24] Madeleine Clare Elish. 2019. Moral crumple zones: Cautionary tales in human-robot interaction (pre-print). *Engaging Science, Technology, and Society (pre-print)* (2019).

[25] Amitai Etzioni and Oren Etzioni. 2016. Designing AI systems that obey our laws and values. *Commun. ACM* 59, 9 (2016), 29–31.

[26] Amitai Etzioni and Oren Etzioni. 2016. Keeping AI legal. *Vand. J. Ent. & Tech. L.* 19 (2016), 133.

[27] Matija Franklin, Edmond Awad, and David Lagnado. 2021. Blaming automated vehicles in difficult situations. *Iscience* 24, 4 (2021), 102252.

[28] Caleb Furlough, Thomas Stokes, and Douglas J Gillan. 2021. Attributing blame to robots: I. The influence of robot autonomy. *Human Factors* 63, 4 (2021), 592–602.

[29] Patrick Gamez, Daniel B Shank, Carson Arnold, and Mallory North. 2020. Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. *AI & SOCIETY* 35, 4 (2020), 795–809.

[30] David Gefen, Elena Karahanna, and Detmar W Straub. 2003. Trust and TAM in online shopping: An integrated model. *MIS quarterly* (2003), 51–90.

[31] Mark A Geistfeld. 2017. A roadmap for autonomous vehicles: State tort liability, automobile insurance, and federal safety regulation. *Calif. L. Rev.* 105 (2017), 1611.

[32] Joshua C Gellers. 2020. *Rights for Robots: Artificial Intelligence, Animal and Environmental Law (Edition 1)*. Routledge.

[33] Tobias Gerstenberg, Anastasia Ejova, and David Lagnado. 2011. Blame the skilled. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 33.

[34] Tobias Gerstenberg and David A Lagnado. 2010. Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition* 115, 1 (2010), 166–171.

[35] Tobias Gerstenberg, Tomer D Ullman, Jonas Nagel, Max Kleiman-Weiner, David A Lagnado, and Joshua B Tenenbaum. 2018. Lucky or clever? From expectations to responsibility judgments. *Cognition* 177 (2018), 122–141.

[36] Donald E Gibson and Scott J Schroeder. 2003. Who ought to be blamed? The effect of organizational roles on blame and credit attributions. *International Journal of Conflict Management* (2003).

[37] Noah J Goodall. 2016. Can you program ethics into a self-driving car? *IEEE Spectrum* 53, 6 (2016), 28–58.

[38] Heather M Gray, Kurt Gray, and Daniel M Wegner. 2007. Dimensions of mind perception. *science* 315, 5812 (2007), 619–619.

[39] Kurt Gray, Liane Young, and Adam Waytz. 2012. Mind perception is the essence of morality. *Psychological inquiry* 23, 2 (2012), 101–124.

[40] Joseph Y Halpern. 2016. *Actual causality*. MiT Press.

[41] César A Hidalgo, Diana Orghian, Jordi Albo Canals, Filipa De Almeida, and Natalia Martin. 2021. *How humans judge machines*. MIT Press.

[42] Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. 2016. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *Journal of Artificial Intelligence Research* 55 (2016), 317–359.

[43] Fatimah Ishowo-Oloko, Jean-François Bonnefon, Zakariyah Soroye, Jacob Crandall, Iyad Rahwan, and Talal Rahwan. 2019. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence* 1, 11 (2019), 517–521.

[44] Robert A Jacobs and John K Kruschke. 2011. Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science* 2, 1 (2011), 8–21.

[45] Deborah G Johnson and Mario Verdicchio. 2019. AI, agency and responsibility: the VW fraud case and beyond. *Ai & Society* 34, 3 (2019), 639–647.

[46] Moritz Jörling, Robert Böhm, and Stefanie Paluch. 2019. Service robots: Drivers of perceived responsibility for service outcomes. *Journal of Service Research* 22, 4 (2019), 404–420.

[47] Eun-Sung Kim. 2020. Deep learning and principal–agent problems of algorithmic governance: The new materialism perspective. *Technology in Society* 63 (2020), 101378. https://doi.org/10.1016/j.techsoc.2020.101378

[48] Richard Kim, Max Kleiman-Weiner, Andrés Abeliuk, Edmond Awad, Sohan Dsouza, Joshua B Tenenbaum, and Iyad Rahwan. 2018. A computational model of commonsense moral decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 197–203.

[49] Lara Kirfel and David Lagnado. 2021. Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition* 212 (2021), 104721.

[50] Lara Kirfel and David Lagnado. 2021. Changing Minds—Epistemic Interventions in Causal Reasoning. (2021).

[51] Max Kleiman-Weiner, Tobias Gerstenberg, Sydney Levine, and Joshua B Tenenbaum. 2015. Inference of Intention and Permissibility in Moral Decision Making.. In *CogSci*.

[52] Max Kleiman-Weiner, Rebecca Saxe, and Joshua B Tenenbaum. 2017. Learning a commonsense moral theory. *Cognition* 167 (2017), 107–123.

[53] Markus Kneer and Michael T Stuart. 2021. Playing the blame game with robots. In *Companion of the 2021 ACM/IEEE international conference on human-robot interaction*. 407–411.

[54] Nils Köbis, Jean-François Bonnefon, and Iyad Rahwan. 2021. Bad machines corrupt good morals. *Nature Human Behaviour* 5, 6 (2021), 679–685.

[55] Fedwa Laamarti, Mohamad Eid, and Abdulmotaleb El Saddik. 2014. An overview of serious games. *International Journal of Computer Games Technology* 2014 (2014).

[56] Francesca Lagioia and Giovanni Sartor. 2020. Ai systems under criminal law: a legal analysis and a regulatory perspective. *Philosophy & Technology* 33, 3 (2020), 433–465.

[57] DA Lagnado and Tobias Gerstenberg. 2015. A difference-making framework for intuitive judgments of responsibility. *Oxford studies in agency and responsibility* 3 (2015), 213–241.

[58] David A Lagnado and Shelley Channon. 2008. Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition* 108, 3 (2008), 754–770.

[59] David A Lagnado and Tobias Gerstenberg. 2017. Causation in legal and moral reasoning. *Oxford handbook of causal reasoning* (2017), 565–602.

[60] David A Lagnado, Tobias Gerstenberg, and Ro'i Zultan. 2013. Causal responsibility and counterfactuals. *Cognitive science* 37, 6 (2013), 1036–1073.

[61] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.

[62] John D Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies* 40, 1 (1994), 153–184.

[63] Anat Lior. 2020. AI entities as AI agents: Artificial intelligence liability and the AI Respondeat Superior Analogy. *Mitchell Hamline Law Review* 46, 5 (2020), 2.

[64] Christian List. 2021. Group agency and artificial intelligence. *Philosophy & Technology* 34, 4 (2021), 1213–1242.

[65] Christian List, Philip Pettit, et al. 2011. *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

[66] Peng Liu, Run Yang, and Zhigang Xu. 2019. How safe is safe enough for self-driving vehicles? *Risk analysis* 39, 2 (2019), 315–325.

[67] Poornima Madhavan and Douglas A Wiegmann. 2007. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science* 8, 4 (2007), 277–301.

[68] Bertram F Malle. 2001. Intention: A Folk-Conceptual Analysis. *Intentions and intentionality: Foundations of social cognition* (2001), 45.

[69] Bertram F Malle, Steve Guglielmo, and Andrew E Monroe. 2014. A theory of blame. *Psychological Inquiry* 25, 2 (2014), 147–186.

[70] Bertram F Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 117–124.

[71] Andreas Matthias. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology* 6, 3 (2004), 175–183.

[72] Alison McIntyre. 2004. Doctrine of double effect. (2004).

[73] Ryan M McManus and Abraham M Rutchick. 2019. Autonomous vehicles and the attribution of moral responsibility. *Social psychological and personality science* 10, 3 (2019), 345–352.

[74] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[75] Judea Pearl. 2009. *Causality*. Cambridge university press.

[76] Stephen R Perry. 2000. Loss, agency, and responsibility for outcomes: Three conceptions of corrective justice. *Philosophy of law* 6 (2000), 546–559.

[77] Walt L Perry. 2013. *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

[78] David A Pizarro and David Tannenbaum. 2012. Bringing character back: How the motivation to evaluate character influences judgments of moral blame. (2012).

[79] Warren S Quinn. 1989. Actions, intentions, and consequences: The doctrine of double effect. *Philosophy & Public Affairs* (1989), 334–351.

[80] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. 2019. Machine behaviour. *Nature* 568, 7753 (2019), 477–486.

[81] Matt Richtel. 2013. How big data is playing recruiter for specialized workers. *New York Times* (2013), 1–7.

[82] Y Rogers, H Sharp, and J Preece. 2002. Interaction design: Beyond human-computer interaction, jon wiley & sons. *Inc* (2002).

[83] Jon Roozenbeek and Sander Van der Linden. 2019. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications* 5, 1 (2019), 1–10.

[84] Stephan Schlögl, Claudia Postulka, Reinhard Bernsteiner, and Christian Ploder. 2019. Artificial intelligence tool penetration in business: Adoption, challenges and fears. In *International Conference on Knowledge Management in Organizations*. Springer, 259–270.

[85] Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2017. Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour* 1, 10 (2017), 694–696.

[86] Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2021. How safe is safe enough? Psychological mechanisms underlying extreme safety demands for self-driving cars. *Transportation research part C: emerging technologies* 126 (2021), 103069.

[87] Francis X Shen, Morris B Hoffman, Owen D Jones, and Joshua D Greene. 2011. Sorting guilty minds. *NYUL rev.* 86 (2011), 1306.

[88] Jenifer Z Siegel, Molly J Crockett, and Raymond J Dolan. 2017. Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition* 167 (2017), 201–211.

[89] Steven A Sloman and David Lagnado. 2015. Causality in thought. *Annual review of psychology* 66 (2015), 223–247.

[90] Stephen C Slota, Kenneth R Fleischmann, Sherri Greenberg, Nitin Verma, Brenna Cummings, Lan Li, and Chris Shenefiel. 2021. Many hands make many fingers to point: challenges in creating accountable AI. *AI & SOCIETY* (2021), 1–13.

[91] Michael T Stuart and Markus Kneer. 2021. Guilty Artificial Minds: Folk Attributions of Mens Rea and Culpability to Artificially Intelligent Agents. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–27.

[92] Tara Qian Sun and Rony Medaglia. 2019. Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly* 36, 2 (2019), 368–383.

[93] Danielle Swanepoel. 2021. Does Artificial Intelligence Have Agency? In *The Mind-Technology Problem*. Springer, 83–104.

[94] Kevin Tobia, Aileen Nielsen, and Alexander Stremitzer. 2021. When does physician use of AI increase liability? *Journal of Nuclear Medicine* 62, 1 (2021), 17–21.

[95] Jacob Turner. 2018. *Robot rules: Regulating artificial intelligence*. Springer.

[96] Nicole A Vincent. 2009. Responsibility: Distinguishing virtue from capacity. *Polish Journal of Philosophy* 3, 1 (2009), 111–126.

[97] Nicole A Vincent. 2011. A structured taxonomy of responsibility concepts. In *Moral responsibility*. Springer, 15–35.

[98] Adam Waytz and Michael I Norton. 2014. Botsourcing and outsourcing: Robot, British, Chinese, and German workers are for thinking—not feeling—jobs. *Emotion* 14, 2 (2014), 434.

[99] C Westcott and David Lagnado. 2019. The AI will see you now: Judgments of responsibility at the intersection of artificial intelligence and medicine (Master's thesis). *Unpublished Manuscript* (2019).

[100] Glanville Williams. 1987. Oblique intention. *The Cambridge Law Journal* 46, 3 (1987), 417–438.