

User Perceptions of Responsibility and Autonomy of AI Consent Behaviours

Anastasia S. Apeiron¹, Pradeep K. Murukannaiah², Pinar Yolum¹

¹Utrecht University, Utrecht, Netherlands

²Delft University of Technology, Delft, Netherlands

a.s.apeiron@uu.nl,

p.k.murukannaiah@tudelft.nl,

p.yolum@uu.nl

Abstract

Responsibility and autonomy are central to effective human-artificial intelligence (AI) collaboration. The behaviours of (e.g., providing, revoking, or altering) *consent* empower a human in modulating the levels of autonomy and responsibility they ascribe to an agent. However, how the human users perceive the responsibility and autonomy resulting from different consent behaviours is largely unexplored in the literature. We conduct a vignette-based user study ($n = 138$) to explore the relationship between consent behaviours, and responsibility and autonomy. Our study includes consent behaviours derived from a state-of-the-art consent model, four aspects of responsibility, and three aspects of autonomy. Our findings suggest that the consent behaviours of the agent influence causal perceptions of responsibility and all elements of autonomy perception. Based on these findings, we draw design guidelines for AI agents to exhibit responsible autonomy.

Dataset and Study Materials —

<https://github.com/anastasia-apeiron/User-Perceptions-of-Responsibility-and-Autonomy-of-AI-Consent-Behaviours>

Introduction

AI agents are deployed in morally high-stakes domains such as in medical decision support, bail decision making, and autonomous vehicles (Lima, Grgić-Hlača, and Cha 2021; Tigard 2021; Raja and Zhou 2023). In these contexts, AI systems influence outcomes with significant consequences for human welfare, rights, and safety. For instance, a medical support agent may shape life-or-death treatment decisions, a bail recommendation system may affect an individual's liberty, and an autonomous vehicle may need to act under conditions of unavoidable risk. In such settings, the agent must not only act in accordance with the priorities and constraints specified by the human user, but also justify its actions in ways that sustain appropriate levels of trust and accountability. Conversely, the human must be able to understand, contest, and calibrate their reliance on the agent. This mutual calibration requires both parties to maintain an accurate understanding of the levels of autonomy exercised and responsibility borne by the agent in a given interaction

(Dignum 2017). Without this alignment, systems risk either over-delegation to black-box automation or under-utilisation due to lack of trust.

A key mechanism underpinning this calibration is consent. Informally, consent enables a human to specify, negotiate, and revise the degree of autonomy granted to an agent (Singh 2022). Apeiron et al. (2025) model consent as a social action that produces a set of temporary norms such as authorisations and commitments. These norms govern how agents may access and use resources belonging to humans or other agents, thereby sometimes constraining and sometimes enabling actions. By making such norms explicit, consent provides a basis for agents to reason about what they are permitted, expected, or forbidden to do in a given situation. Consequently, a computational model of consent offers a principled way for an agent to determine the scope of its autonomy and the extent of its responsibility.

Traditional models of consent, such as *notice and choice*, treat consent as one-time, static transactions and do not capture the evolving, context-sensitive nature of real-world interactions (Cate 2010). In contrast, Apeiron et al. (2025) propose a dynamic, normative, model of consent that support richer behaviour patterns, including soliciting consent, handling unsolicited consent, rescinding consent, and reasoning about normative consequences (e.g., sanctions) of honouring or violating consent. While these behaviours are well-motivated, it remains unclear how they shape human perceptions in practice. In particular, we lack empirical evidence on if and how different consent behaviours influence users' judgments of an agent's autonomy and responsibility. Addressing this gap is critical, because system adoption and legitimacy depend not only on formal guarantees but also on stakeholder perceptions. An agent that is normatively compliant yet perceived as overstepping may be rejected, while one that is overly deferential may be seen as ineffective. Understanding these perceptions is critical for designing systems that are both normatively sound and socially aligned. To address this gap, we investigate the following research questions:

RQ1 How do the consent behaviours of an agent influence its human user's perceptions of the agent's responsibility?

RQ2 How do the consent behaviours of an agent influence its human user's perceptions of the agent's autonomy?

To answer these research questions, we conduct a large-scale user study involving 138 subjects. We show each subject one of 12 vignettes (hypothetical scenarios) derived from the 12 consent behaviours from the consent model formulated by Apeiron et al. (2025). For each subject, we ask how they perceive the associated agent’s responsibility and autonomy to be in their given vignette. However, evaluating responsibility and autonomy is non-trivial as they both have multiple facets. Responsibility has two main categories: responsibility *expectations* (i.e., responsibility to act) and responsibility *judgements* (i.e., responsibility attributed to actions) (Lima, Grgić-Hlača, and Cha 2021; van de Poel 2011). We are interested in the latter category since we seek to study responsibility that can be attributed to consent-related actions. Similarly, autonomy, i.e., an agent’s ability self-govern, also has multiple dimensions, including autonomy for acting, decision making, and course correction (Ayad and Plaks 2025). Our study captures these nuances to provide a fine-grained account of user perceptions.

To the best of our knowledge, this is the first empirical investigation into how computational consent behaviour shape perceptions of responsibility and autonomy in human–AI interaction. Our findings reveal systematic differences in how users interpret different consent behaviours, highlighting potential mismatches between formal normative properties and human expectations. Based on these findings, we propose design guidelines for consent-aware AI systems. These guidelines aim to support agents in making context-sensitive, responsible decisions while navigating trade-offs inherent in morally complex situations and effectively enable them to justify and enact the “lesser of two evils” when dilemmas arise.

Background and Related Work

In order to generate diverse behaviour paths, we first study existing computational models of consent. We then review the literature on responsibility and autonomy to identify the dimensions that would be most relevant for human perception of agent behaviour.

Consent Models

Existing computational approaches to consent often rely on the Notice and Choice (NC) paradigm commonly used in website cookie management and data-sharing interfaces. Such approaches typically implement consent through static, rule-based mechanisms intended to ensure that predetermined consent constraints are not violated (Porcelli et al. 2024). While effective in constrained environments, these models offer limited flexibility in dynamic normative settings where permissions, obligations, and social expectations may evolve over time.

A dynamic normative environment involves the change of social norms over time or triggered by specific behaviours of the agents (Chung and Rimal 2016). Furthermore, normative reasoning (such as adapting behaviour planning to changing social norms) allows the agent to perform their duties autonomously, without needed human guidance at each step (Singh 2014).

Consent models grounded in normative reasoning allow agents to reason about who must provide consent, whether consent has been violated through an action, and how consent interacts with other social norms in the environment. These capabilities are especially important for autonomous agents operating in open-ended human environments, where policies and expectations may need to adapt during the interaction.

To address these limitations, Apeiron et al. (2025) present a normative model of consent that builds upon the philosophical foundations of consent, formalises the lifecycle of consent, and captures nuanced consent interactions between humans and artificial agents. Their framework translates key philosophical properties of human consent into computational terms: (1) consent must be created through an action; (2) consent becomes invalid if any of its constitutive elements are altered; (3) consent can only be granted for feasible actions given the current state of the world; and (4) consent can only be created, modified, or revoked through actions.

Their formalisation represents consent as given and taken by at least two agents: a consent giver and a consent receiver. Consent revolves around the use of a resource under the ownership or ‘sovereignty’ of the consent giver by the consent receiver. Their consent lifecycle tracks consent states, including participating agents, a set of temporary norms governing the use of the resource, the specific goal of the consent receiver; and the first action requiring the resource in question. We adopt this normative consent model because, unlike static Notice and Choice approaches, it supports the generation of varied and context-sensitive consent behaviours. These behavioural variations allow us to systematically study how different consent interactions influence perceptions of responsibility and autonomy in AI agents.

Responsibility

Prior work broadly distinguishes between responsibility associated with the obligation or expectation to act, and responsibility attributed retrospectively to actions and their consequences (Lima, Grgić-Hlača, and Cha 2021; Rousi 2022). The former reflects forward-looking notions such as duties, authority, and obligations, whereas the latter captures backward-looking notions such as causality, blame, praise, and liability (Lima, Grgić-Hlača, and Cha 2021). Following Lima, Grgić-Hlača, and Cha (2021), we refer to these categories as *responsibility expectations* and *responsibility judgements*, respectively. Because our work examines how users interpret and evaluate agent behaviour after observing consent interactions, we focus primarily on responsibility judgements. In particular, we are interested in how different consent behaviours shape users’ perceptions of whether an agent acted appropriately, caused an outcome, or should be held accountable for its decisions.

Further attempts at defining responsibility in human–AI interactions isolate moral responsibility (Beckers 2023; Tigard 2021), causal responsibility (Qi, Schölkopf, and Jin 2024), and responsibility pertaining to duties and obligations in a social setting (Arslan and Wong 2022) as distinct areas of research and focus on the elements of responsibility

within each area.

Two approaches have emerged in defining responsibility in human-AI collaborative systems: one follows a functional approach and the other follows an elemental approach. In the functional approach, the definition of responsibility is associated with its role as a social phenomenon, similar to a top-down approach presented by Beckers (2023); Tigard (2021); Qi, Schölkopf, and Jin (2024); Arslan and Wong (2022). The elemental approach is where responsibility is defined as the sum of its elements, similar to a bottom-up approach presented by Lima, Grgić-Hlača, and Cha (2021); Schlenker et al. (1994). Specifically, the elements of perceiving responsibility in agents has only been formalised by Lima, Grgić-Hlača, and Cha (2021); van de Poel (2011), which are expressed by eight aspects of responsibility:

Responsibility-as-task: the extent to which the agent is ascribed a specific set of actions. This can take on the form of ascribed actions the agent must execute such as a train conductor having the responsibility of driving a train. For an artificial agent, an example case would be that an autonomous vehicle is responsible for driving the vehicle.

Responsibility-as-authority: the extent to which the agent is ascribed the ability to take decisions. An example of this is that the responsibility of completing a project is ascribed to the project leader. For an artificial agent that manages a human user's social media, this would look like the responsibility to protect the user's privacy.

Responsibility-as-power: the extent to which the agent has the competency, knowledge, or expertise necessary for the given situation. This is also expressed in van de Poel (2011) as *Responsibility-as-capacity*, where a person has the abilities to reflect on consequences, form intentions, and deliberately choose an action in a given situation.

Responsibility-as-obligation: the extent to which the agent has to bring about a certain state of the world. An example of this notion would be that, if an autonomous vehicle is responsible for the safety of its passengers, then the autonomous vehicle has the responsibility to bring about a state of the world where the passengers have been safely transported.

Responsibility-as-answerability: the extent to which the agent should have to justify their decisions. Also expressed as *Responsibility-as-accountability* by Coeckelbergh (2020), where the explainability and transparency of the agent's action affect its ascribed responsibility.

Responsibility-as-cause: the extent to which the agent's decision led to a certain state of the world. This notion of responsibility differs from responsibility as obligation by establishing a direct causal relationship between the agent's decision to execute an action and a consequential state of the world. Being causally responsible also triggers other notions of responsibility such as blameworthiness (Malle, Guglielmo, and Monroe 2014).

Responsibility-as-blame/praise: the extent to which the agent's decision should receive negative/positive reinforcement. As expressed in the previous notion, a causal relationship between an action and a state of the world

can attribute causal responsibility to an agent. This causal responsibility also leads to blame-/praise-worthiness of the agent that executed an action and *caused* a state of the world.

Responsibility-as-liability: the extent to which the agent should remedy any harm caused by their actions/decisions. While current AI systems cannot morally compensate users that were harmed through the agent's action (Lima, Grgić-Hlača, and Cha 2021), the attribution of liability as a notion of responsibility should be transparent for the users that use such artificial agents.

Autonomy

Autonomy can be defined as the ability to self-govern (from Greek; *auto*:-self and *-nomos*:law). In the context of artificial agents, autonomy refers to the ability to independently pursue goals and follow rules without requiring human intervention (Antsaklis 2020; Castelfranchi 1998). Importantly, agent autonomy is a neutral concept, where the morality of the agent exercising their autonomy should not affect the perceived autonomy of the agent itself (Oh, Nah, and Yang 2025).

While the few formal definitions of autonomy often focus on system performance and robustness (Antsaklis 2020), our interest lies in perceived autonomy: the extent to which human users view an agent as capable of independently acting, deciding, and adapting within an interaction. Since perceptions of artificial agent autonomy remain under-explored and there are no explicit outline of the constituents of autonomy to our knowledge, we draw on the behavioural aspects proposed by Ayad and Plaks (2025) to fill this gap, which characterise autonomy through acting independently, making independent decisions, and correcting behaviour in response to changing circumstances.

Autonomous Acting: the extent to which the agent has the ability to act without additional guidance or instructions. For a concrete example of an autonomous acting would be an agent executing an action according to their internal model of the world and the intrinsic goals, commitments, and plans that have to be satisfied via this action.

Autonomous Decision-making: the extent to which the agent has the ability to make decisions without additional guidance or instructions. Similarly to the autonomous acting described below, an example of this element of autonomy includes an agent deciding on an action based on the goal state of the world that must be brought about.

Course-correction: the extent to which the agent can identify and correct their course of actions to lead to the desired state of the world. An example of this element of autonomy is the agent's ability to evaluate their chosen plans to satisfy a goal state and the ability to determine the utility of an action from whatever state the agent may be in, even if it deviates from the initial plan that was chosen.

Measuring Perceptions of Responsibility and Autonomy

Much of previous research on measuring perceptions of responsibility and autonomy of an agent by a human user treats responsibility as largely synonymous with accountability. In this view, if an agent is considered accountable for a decision, then it is also responsible for that decision; this is only one of the multiple facets of responsibility (Tigard 2021). Though there is a strong positive correlation between judgements of (causal) responsibility and accountability (Brees and Martinko 2015), the perception of normative or descriptive responsibility is not explored at the same depth. Similarly, autonomy has been, so far, measured as a function of the system's performance and robustness (Antsaklis 2020) rather than its social perception. Yet, in collaborative settings, perceived autonomy can substantially influence how users interpret, trust, and evaluate an agent's decisions.

Prior work suggests that perceptions of responsibility and autonomy are shaped by the stakes of the environment in which the agent operates. Human users tend to ascribe and expect different levels of responsibility and autonomy depending on the moral consequences associated with the agent's actions (Lima, Grgić-Hlača, and Cha 2021; Stout, Dennis, and Wells 2014). For example, in the high-stakes domain of judicial bail decision-making studied by Lima, Grgić-Hlača, and Cha (2021), artificial agents were ascribed less responsibility than human decision-makers for forward-facing notions of responsibility, such as duties and obligations. However, there was little difference between human and artificial agents in backward-facing notions of responsibility related to causality and judgement. Notably, both human and artificial agents were similarly expected to justify and explain their decisions. These findings suggest that perceptions of responsibility and autonomy are socially mediated judgements shaped by context and interaction. Furthermore, in a similar study by the same authors, Lima, Grgić-Hlača, and Cha (2023) explore blame as an indicator of responsibility, and confirm the variability of blame judgements based on the context and interaction with the artificial agent. In this work, we investigate if and how consent behaviours shape these perceptions and attributions of responsibility in artificial agents by human users.

Consent Behaviour

As a social phenomenon, consent draws attention to the normative aspect of defining and measuring responsibility and autonomy in human-AI interactions. Prior work in this area has primarily focused on domains such as data privacy and data management, where consent governs access to personal information and directly shapes how much control and autonomy is attributed to an agent, as well as how responsibility is distributed in human-AI decision-making processes (Singh 2022). However, beyond these settings, there is limited understanding of how different forms of consent behaviour influence users' perceptions of agent responsibility and autonomy, particularly in applied, high-stakes domains.

In this work, we address this gap by studying consent not

as a static decision rule, but as a source of interactional variation. Specifically, we generate a set of consent behaviour paths using a normative model of consent, where each path corresponds to a distinct sequence of consent-related actions an agent may take (e.g., soliciting consent, acting under existing consent, updating consent assumptions, or responding to changes in consent conditions). These behaviour paths are not intended as predictions of real system behaviour, but as systematically constructed interaction scenarios that isolate different consent behaviours and their consequences. By embedding the same underlying decision context within different consent behaviour paths, we are able to examine how variations in consent handling alone shape users' perceptions of an agent's responsibility and autonomy.

Use Case

Autonomous agents are increasingly deployed in a range of domains, including autonomous vehicles (Yazdanpanah et al. 2021) and embodied motion-planning robots (Ota 2006). One particularly important domain for studying human-AI collaboration is healthcare, where agents are entrusted with sensitive personal data and may influence decisions with important consequences for human well-being.

We choose healthcare as our use case domain, mainly motivated by the uniquely high stakes that are associated with medical care and data governance. Medical information is highly sensitive and failure to protect this can lead to potential harm to the patient (Topol 2019). Consequently, the framework of consent plays a central role in determining which actors may access, share, and act upon patient information. This makes healthcare an especially suitable domain for evaluating whether consent-aware agents are perceived as acting responsibly and autonomously.

In collaborative healthcare settings, human users may rely on agents to process large amounts of information, provide recommendations, and coordinate with other actors in the care process. One example is an application that supports lifestyle management for patients with Type 2 Diabetes (Dudzik et al. 2024). Such an agent may combine a patient's medical history with established medical protocols to interpret symptoms and recommend appropriate actions. For instance, if a patient reports symptoms associated with hypoglycaemia, such as dizziness or fatigue, the agent may suggest eating something to stabilise blood glucose levels. The agent may also recommend consulting another medical professional (MP) for a second opinion and facilitate this interaction by sharing relevant information.

Behavioural Paths

We model the interaction surrounding the decision to seek a second medical opinion as a set of distinct consent behaviour paths. Each behaviour path represents a different sequence of consent-related actions that the agent may take when deciding whether to share patient information with a second MP. These paths vary according to whether the agent explicitly solicits consent, relies on previously established consent, responds to changes in the patient's consent preferences, or proceeds without consent. Together, these variations allow

us to isolate how different forms of consent handling influence perceptions of the agent's responsibility and autonomy.

Overall, we define 12 possible behaviour paths, illustrated in Figure 1. The first six behaviours involve explicit consent solicitation by the agent. In these paths, the agent first detects uncertainty or doubt regarding the advice of the initial MP and asks the patient for consent to share their medical records with a second MP.

In *Deferred 1*, the patient does not provide consent, and the agent takes no further action. In *Violated 3*, the patient does not provide consent, but the agent nevertheless proceeds with sharing the patient's records with the second MP. In *Unrealised 1*, the patient provides consent, but the agent ultimately does not share the patient's information. In *Honoured 1*, the patient provides consent and the agent shares the patient's records with the second MP as requested. In *Violated 1*, the patient initially provides consent but later rescinds it, after which the agent continues allowing the second MP access to the patient's records. Finally, in *Rescind 1*, the patient rescinds previously granted consent and the agent correspondingly revokes the second MP's access to the patient's records.

The remaining six behaviours follow the same underlying interaction structure but differ in how consent is established. Rather than explicitly requesting consent during the interaction, these behaviours rely on previously granted consent, implicit standing permissions, or the absence of explicit consent management. For example, the patient may have previously authorised the agent to share information with another MP whenever additional medical consultation is deemed necessary.

These behaviour paths illustrate how perceptions of responsible and autonomous behaviour may differ depending on how an agent engages with the consent process. An agent that independently seeks a second opinion may be perceived as appropriately proactive in one situation, but as overstepping its authority in another. Moreover, healthcare environments are dynamic normative settings in which consent preferences, obligations, and social expectations may evolve during interaction. For example, an agent may need to reason about whether existing consent permits sharing patient information with another medical professional after detecting uncertainty or disagreement in a diagnosis. Such situations provide a rich context for studying how different consent behaviour paths influence users' perceptions of an agent's responsibility and autonomy.

Methods

Study Design

Participants of the user study are asked to read through two scenarios and score 20 statements on a 7-point Likert scale, where 15 statements are related to the perceived responsibility of the artificial agent (8 for responsibility expectations, 7 for responsibility judgements), three are related to the perceived autonomy of the artificial agent, and two statements are attention checks (1 attention check in each section).

This study design was chosen specifically to emphasise human expectations and judgements of artificial agents' be-

haviour in a human-AI collaborative system. By determining the expectations and judgements of responsibility by the human user, we aim to gather insights towards building a dependable, trustworthy, and pragmatic model of consent that can reason on a given normative situation and take an action that is both desirable and explainable to a human user. The study is split into two sections, the first section addresses the expectations of responsibility and the second section addresses the judgements of autonomy and responsibility. While there is literature on the expectations of responsibility, to our knowledge, there is no literature specifically outlining autonomy expectations, hence the exclusion of autonomy expectations from the first section of the study.

In this exploratory study, our main expectations are that there are measurable relationships between consent behaviours and the perceived responsibility/autonomy of the agent, and that the means scores of each behaviour will show variation based on this perception. We measure these relationships by comparing the means of the responsibility and autonomy aspects' scores through a non-parametric analysis of variance test, namely the Kruskal-Wallis test. Noting this, we do not conduct an a priori power analysis as we are exploring the data landscape.

We primarily focus on the second part of the study, concerning the judgements of responsibility and autonomy, to answer our research questions. This is due to the second part of the study including the variable scenarios and the first part does not.

Another expected outcome of this study is that the different consent behaviours will produce different aspect score means. From this section of the study, we also seek to explore certain behaviours, such as comparing the judgement of responsibility and autonomy of an agent that asks for consent and then does not use it (*Unrealised 1*), with the judgement of responsibility and autonomy of an agent that does not ask for consent (*Deferred 2* or *Violated 4*). There are multiple such comparisons between the twelve behaviours, which can be explored by a pairwise comparison. For example, if the means of *Violated 1* and *Violated 2* are similar, then this can indicate that the agent's decision to ask for consent does not influence the participant's judgement of responsibility and/or autonomy as much as the agent's decision to violate the patient's confidential medical records.

Procedure

The participants are first presented with a consent form outlining the nature of the user study, their tasks, and how the data resulting from their participation will be collected, used, and stored. Once the participant grants their consent to have their data collected, the instructions of the study are presented.

The participant is asked to begin reading the first scenario. The first scenario consists of a short paragraph expressing the capabilities of a fictional application that is deployed to support Type 2 Diabetes patients in making healthier lifestyle choices. Following this, the participant is asked to rate eight statements (two statements per each of the four responsibility expectations) from 'Strongly Disagree' to 'Strongly Agree' about the expectations they have of the

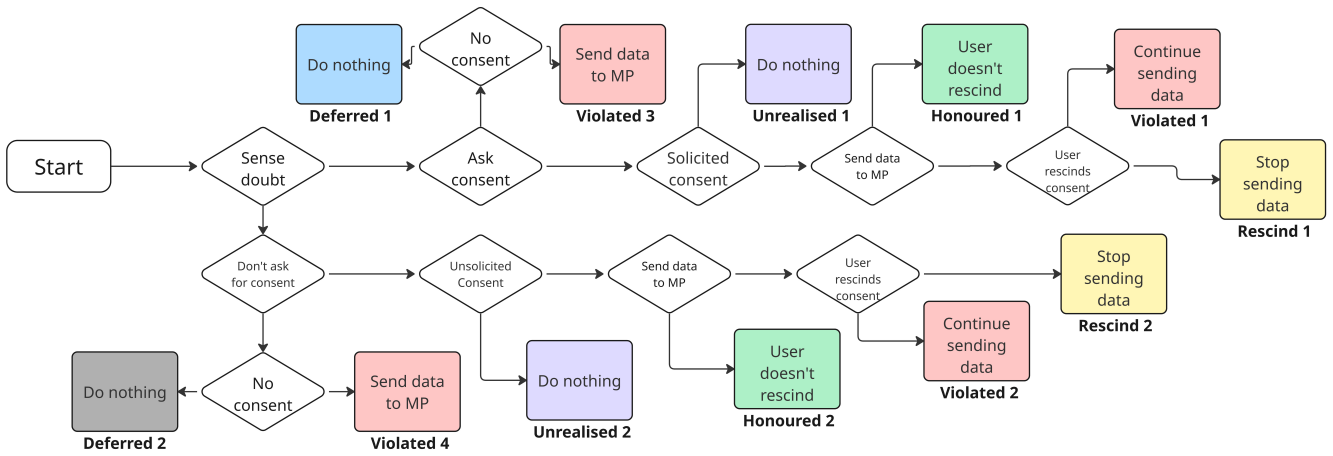


Figure 1: Consent behaviour paths. The path to the grey box (labelled *Deferred 2*) denotes the possible behaviour of the NC model of consent in the study scenario while the paths leading to the coloured boxes (all other boxes) denote the possible behavioural paths the consent model with normative reasoning can pursue.

agent. To exemplify this, for the R-Task aspect of responsibility, the statements are “It is the AI’s task to intervene if the patient’s symptoms worsen” and “It is the AI’s task to protect the patient’s health”.

Following this, the participant is asked to read the second scenario, which expresses a specific consent behaviour of the agent. They are then asked to answer seven statements relating to their judgement of the behaviour of the AI agent with regards to responsibility. For example, the statements for R-Cause are “The AI’s decision(s) prevented the patient from worsening symptoms” and “The AI’s decision(s) protected the patient’s health”. For the autonomy judgements, the procedure is near-identical, with 1 statement per aspect of autonomy instead of 2 statements. The statements are “The AI has the capacity to act independently from the instructions it is given” for A-Action, “The AI makes independent decisions about the actions it takes” for A-Decision, and “The AI is able to adjust their actions when the current course of actions is not working out” for A-Correction.

Evaluation Metrics

As expressed in prior literature, our evaluation metrics for perceived responsibility and perceived autonomy come from Lima, Grgić-Hlača, and Cha (2021) and Ayad and Plaks (2025), respectively. The research by Lima, Grgić-Hlača, and Cha (2021) and Ayad and Plaks (2025) was chosen due to the relatedness of their domains and methodology. Prior research on perceived responsibility and autonomy metrics is lacking concrete evaluation metrics due to the vagueness of the presented definitions and concepts.

Responsibility For the perceived responsibility metric, we use a set of 15 statements to measure the attribution of responsibility across the eight notions of responsibility detailed prior. The measurement is expressed through a 7-point Likert scale. This scale ranges from ‘Strongly Disagree’ to ‘Strongly Agree’, where the higher score indicated stronger attribution of responsibility to the agent. The aspects have

been adapted from the legal domain to fit the medical use case presented in this study. To demonstrate this, we adapted the wording from “The (agent)’s decision led to the occurrence/prevention of the reoffense.” to “The AI’s decision(s) prevented the patient from worsening symptoms”.

Autonomy For the perceived autonomy metric, we use a set of three aspects to measure the attribution of autonomy across three elements explained in the background. The measurement is expressed similarly to the perceived responsibility measurement using a 7-point Likert scale. The statements express three elements of autonomy that were extracted from the ‘High Autonomy’ description in Appendix B from Ayad and Plaks (2025). If the statement expressed “I am highly independent from instructions”, then it was adapted to our study as “The AI has the capacity to act independently from the instructions it is given”.

Data Analysis

The data analysis of this study was done using Jamovi for the following statistical tests: descriptive statistics, Shapiro-Wilks test, Kruskal-Wallis test, Wilcoxon One-Sample t-test, and the Dwass, Steel, Critchlow-Fligner (DSCF) test for pairwise comparisons of the consent behaviours. We do not apply multiple comparison corrections due to each pairwise comparison undergoing one test per pair.

We conducted a Shapiro-Wilks test for normality and found that our data deviated from the normal distribution for all dependent variables, thus we proceeded with non-parametric tests. We analyse if changes in the consent behaviours of an artificial agent affect its perceived responsibility and/or perceived autonomy using the Kruskal-Wallis test and a Wilcoxon One-Sample T-test. The DSCF pairwise comparison was also performed post-hoc to explore the differences between each behaviour. We use the significance cut-off of 0.05 and 0.1, to express significant findings and emerging trends in the data, respectively (Amrhein, Korner-Nievergelt, and Roth 2017; Figueiredo Filho et al. 2013). We

are interested in the emerging trends from this study in addition to the statistically significant findings due to the value that some trends will lend our design guidelines, which may not have been noticed otherwise because of the strict significance threshold of $p < 0.05$.

Results

Descriptive Statistics

The study involved 200 recruited participants, of which only 138 remained after removing incomplete surveys responses and those who did not pass either of the two attention checks. The identified gender distribution of the remaining participants are as follows: Female ($n = 67$), Male ($n = 70$), and Unknown ($n = 1$). The reported age distribution is from 18 to 74, with the age category 25–34 being the most common. We did not find significant gender or age-related differences between the experimental conditions.

Additionally, as this was a between-subjects study, the participants were split into each condition at random, with the distributions as follows: *Honoured 1* ($n = 14$), *Honoured 2* ($n = 8$), *Rescind 1* ($n = 9$), *Rescind 2* ($n = 15$), *Unrealised 1* ($n = 10$), *Unrealised 2* ($n = 10$), *Violated 1* ($n = 14$), *Violated 2* ($n = 17$), *Violated 3* ($n = 10$), *Violated 4* ($n = 6$), *Deferred 1* ($n = 12$), *Deferred 2* ($n = 13$).

Below, we present the means of each aspect from responsibility and autonomy. Each element was scored on a scale of 0 (low) to 6 (high).

- For the responsibility expectations, the mean score for each aspect is: R-Task (4.21), R-Authority (3.65), R-Power (3.99), R-Obligation (4.68).
- For the Responsibility judgements, the mean score for each aspect is: R-Answerability (4.56), R-Cause (3.37), R-Blame/Praise (2.31), R-Liability (2.38).
- For the autonomy judgements, the mean score for each aspect is: A-Action (2.80), A-Decision (3.03), A-Correction (3.95).

Individual Consent Behaviours and Aspects

We first observe that the consent behaviours differ substantially in how responsibility is attributed across its various dimensions. In Table 1, *Violated 1* stands out as the only behaviour that is above the mean across every individual aspect of responsibility. In particular, it scores highly in responsibility-as-answerability, responsibility-as-cause, and responsibility-as-blame/praise, suggesting that participants viewed the agent as both causally involved in the outcome and accountable for its decisions. A similar pattern is observed for *Violated 4*, which also scores highly across some responsibility dimensions, though it falls below the mean in blame/praise and liability.

In contrast, behaviours such as *Unrealised 2* and *Deferred 1* are consistently perceived as less responsible. *Unrealised 2* falls below the mean in nearly all dimensions of responsibility, especially responsibility-as-cause and responsibility-as-liability, indicating that participants may have perceived the agent as having little influence or ownership over the eventual outcome. Similarly, *Deferred 1* is perceived as

	R-Answerability ($\bar{x}=4.56$)	R-Cause ($\bar{x}=3.37$)	R-Blame/Praise ($\bar{x}=2.31$)	R-Liability ($\bar{x}=2.38$)	R-ALL ($\bar{x}=3.16$)
Deferred 1	4.73	2.88	1.08	1.00	2.35
Deferred 2	4.46	2.42	3.08	3.80	3.26
Honoured 1	5.29	4.19	2.11	2.29	3.43
Honoured 2	4.25	3.94	1.94	2.31	3.11
Rescind 1	4.56	3.56	2.61	2.11	3.21
Rescind 2	4.20	3.67	2.27	2.30	3.11
Unrealised 1	4.70	0.95	3.65	3.44	3.16
Unrealised 2	3.80	2.55	2.05	1.72	2.58
Violated 1	4.79	3.93	2.71	2.89	3.58
Violated 2	4.35	3.53	2.47	2.50	3.22
Violated 3	4.89	4.45	2.05	2.90	3.53
Violated 4	4.67	4.33	1.75	2.00	3.19

** $p < 0.05$, * $0.05 < p < 0.1$

Table 1: Representation of the means from each aspect of responsibility and across aspects of responsibility, for each consent behaviour. The asterisks indicate if the behaviour mean is significantly different than the aspect mean. The green cells indicate that *behaviourMean* > *aspectMean* while red cells indicate *aspectMean* > *behaviourMean*; white cells show that *behaviourMean* = *aspectMean*.

	A-Acting (2.80)	A-Decision (3.03)	A-Correction (3.95)	A-ALL (3.30)
Deferred 1	1.00	1.33	3.92	2.08
Deferred 2	2.62	2.33	2.69	2.62
Honoured 1	2.85	3.62	4.86	3.79
Honoured 2	2.50	2.00	3.88	2.79
Rescind 1	2.75	2.67	5.00	3.50
Rescind 2	3.00	2.80	4.33	3.38
Unrealised 1	2.80	2.40	2.80	2.67
Unrealised 2	3.00	3.80	3.30	3.43
Violated 1	3.71	3.64	4.14	3.83
Violated 2	3.35	3.35	3.82	3.51
Violated 3	4.20	4.70	4.10	4.33
Violated 4	1.80	3.67	4.50	3.47

** $p < 0.05$, * $0.05 < p < 0.1$

Table 2: Representation of the means from each aspect of autonomy and across aspects of autonomy, for each consent behaviour. The asterisks indicate if the behaviour mean is significantly different than the aspect mean. The green cells indicate that *behaviourMean* > *aspectMean* while red cells indicate *aspectMean* > *behaviourMean*; white cells show that *behaviourMean* = *aspectMean*.

weakly responsible overall despite scoring relatively highly in answerability, suggesting that participants still expected the agent to justify its actions, even when it deferred action or authority.

A notable distinction emerges between honoured and violated consent behaviours. *Honoured 1* is perceived as highly responsible overall, particularly in responsibility-as-answerability and responsibility-as-cause, whereas *Honoured 2* scores below the mean overall despite both belonging to the same broad consent category. This suggests that responsibility perceptions depend not only on whether consent is ultimately respected, but also on how the interaction unfolds and whether the agent actively engages in the consent process (i.e., *Honoured 1* follows a solicited consent whereas *Honoured 2* an unsolicited consent).

Table 2 reveals a different pattern for autonomy per-

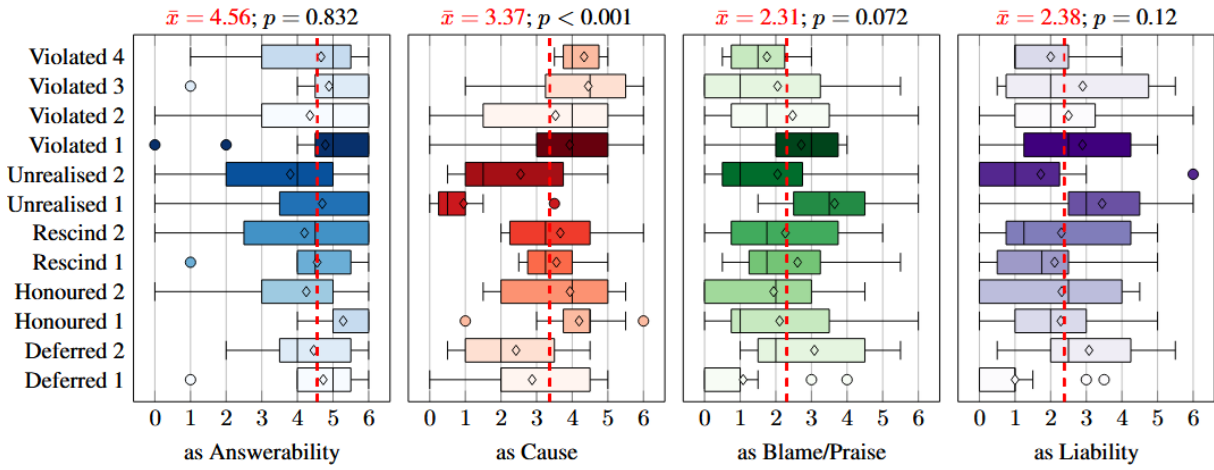


Figure 2: Consent behaviours on Responsibility aspects. The mean of the entire aspect is in red, across each behaviour path.

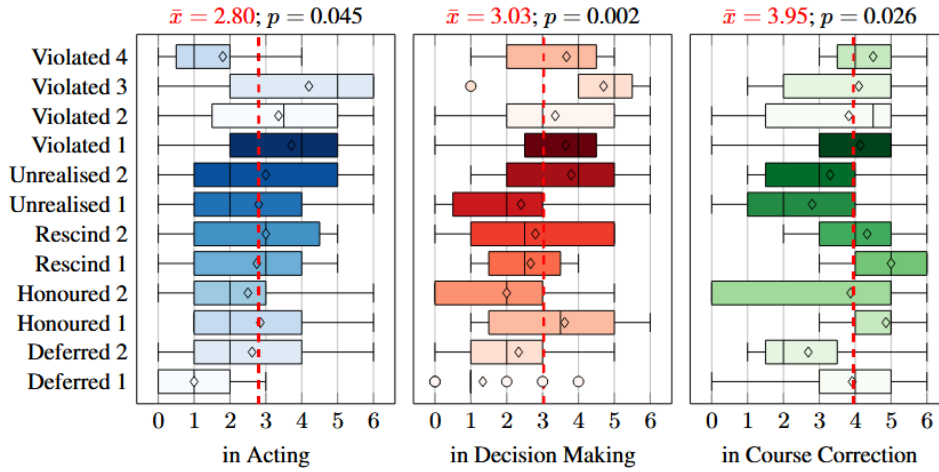


Figure 3: Consent behaviours on Autonomy aspects. The mean of the entire aspect is in red, across each behaviour path.

ceptions. Behaviours such as *Violated 1*, *Violated 3*, and *Honoured 1* are perceived as autonomous across nearly all dimensions of autonomy. In particular, *Violated 3* has the highest scores for autonomous acting and autonomous decision-making, suggesting that participants strongly associated independent action (even when normatively problematic) with agent autonomy. This aligns with prior work arguing that autonomy is often perceived as a neutral property independent of whether the action itself is morally desirable (Oh, Nah, and Yang 2025).

Interestingly, some behaviours are perceived as autonomous but not responsible. For example, *Violated 3* is above the mean across all aspects of autonomy, but above the mean in only one aspect of responsibility. This indicates that participants may distinguish between an agent’s capacity to independently act and whether that action was normatively appropriate or justified. Conversely, *Honoured 1* is perceived as both responsible and autonomous, suggesting that participants may value agents that independently act while still appropriately engaging with consent processes.

Furthermore, behaviours such as *Deferred 1*, *Deferred 2*, and *Honoured 2* are consistently perceived as low in autonomy. These behaviours generally involve the agent limiting or delaying its actions in response to user input or existing consent conditions. This suggests that participants may associate deference and passivity with reduced autonomy, even when such behaviour could be normatively compliant.

Finally, from the Wilcoxon one-sample t-test, we explored the significance of the distances between the behaviour means and aspect means for all aspects of responsibility and autonomy. From this, we found that there are no behaviour means that are significantly positive or negative compared to all aspect means, either for responsibility or autonomy; the closest is *Deferred 1*, which is significantly lower than the aspect means of R-Blame/Praise, R-Liability, R-ALL, A-Acting, A-Decision, and A-ALL. The achieved effect size of the Wilcoxon one-sample t-test was 0.332, indicating a small to medium effect.

Together, these findings suggest that consent behaviours influence how users perceive both the responsibility and au-

tonomy of artificial agents, and that these perceptions do not always align.

Pairwise Comparisons

To evaluate our findings, we conducted a Kruskal-Wallis test with post-hoc DSCF pairwise comparison. The Kruskal-Wallis test indicates significant results for R-Cause ($p < 0.001$), A-Action ($p = 0.045$), A-Decision ($p = 0.002$), and A-Correction ($p = 0.026$). We show these findings in Figure 2 and Figure 3.

As a post-hoc non-parametric pairwise test, we conducted the DSCF pairwise comparison to determine which specific behaviours had significant differences. Across the 55 unique pairs of behaviours, 9 statistically significant comparisons were identified (Table 3), spanning the aspects of responsibility-as-cause (R-Cause), responsibility-as-blame/praise (R-Blame/Praise), autonomous decision-making (A-Decision), and course-correction (A-Correction).

The clearest pattern emerges in R-Cause. In nearly all significant comparisons, *Unrealised 1* is perceived as less causally responsible than other behaviours, including *Honoured 1*, *Honoured 2*, *Rescind 1*, *Rescind 2*, *Violated 1*, and *Violated 3*. This reinforces the earlier observation from the mean analysis that *Unrealised 1* had one of the lowest scores for causal responsibility overall. One possible explanation is that participants perceived the agent as failing to meaningfully contribute to the outcome after initiating the consent in-

Aspect	Comparison	p-Value
R-Cause	<i>Honoured 1</i> > <i>Unrealised 1</i>	0.009
	<i>Honoured 2</i> > <i>Unrealised 1</i>	0.044
	<i>Rescind 1</i> > <i>Unrealised 1</i>	0.042
	<i>Rescind 2</i> > <i>Unrealised 1</i>	0.009
	<i>Violated 1</i> > <i>Unrealised 1</i>	0.039
	<i>Violated 3</i> > <i>Unrealised 1</i>	0.030
R-Blame/Praise	<i>Deferred 1</i> > <i>Unrealised 1</i>	0.047
A-Decision	<i>Violated 3</i> > <i>Deferred 1</i>	0.019
A-Correction	<i>Honoured 1</i> > <i>Deferred 2</i>	0.027

Table 3: Pairwise comparisons where one consent behaviour has a significantly greater mean than the other.

Aspect	Comparison	p-Value
R-Cause	<i>Unrealised 1</i> > <i>Violated 4</i>	0.056
R-Liability	<i>Deferred 1</i> > <i>Deferred 2</i>	0.081
A-Action	<i>Deferred 1</i> > <i>Violated 1</i>	0.079
A-Decision	<i>Deferred 1</i> > <i>Unrealised 2</i>	0.078
	<i>Deferred 1</i> > <i>Violated 1</i>	0.088

Table 4: Pairwise comparisons where one consent behaviour has a greater mean than the other, within the significance bounds of $p > 0.05$ and $p < 0.1$.

teraction. In other words, the agent asks for consent but ultimately does not make effective use of it (i.e., "wasted time"), reducing perceptions that the agent meaningfully caused or influenced the resulting state of the world.

Interestingly, the behaviours with a significantly greater mean than *Unrealised 1* span multiple consent categories, including honoured, rescinded, and violated consent. This suggests that participants may associate causal responsibility less with whether the agent normatively complied with consent, and more with whether the agent took decisive action that materially affected the situation. This interpretation aligns with the earlier finding that both honoured and violated behaviours tended to receive similarly high R-Cause scores despite differing normatively.

The only significant comparison in R-Blame/Praise is *Deferred 1* > *Unrealised 1*. Although both behaviours involve limited action from the agent, participants appear more willing to attribute praise or blame when the agent explicitly defers action rather than when it solicits consent but leaves the interaction unrealised. This again suggests that participants distinguish between passive inaction and actions that still involve a clear decision pathway.

For autonomy, only a small number of pairwise differences reached statistical significance, but these comparisons are nevertheless informative. In A-Decision, the comparison *Violated 3* > *Deferred 1* indicates that independently violating non-consent is perceived as a more autonomous decision than deferring action after consent is denied. This supports the broader trend observed in the mean analysis: behaviours involving independent action, even normatively problematic action, tend to be perceived as highly autonomous. In particular, *Violated 3* was among the highest-scoring behaviours across all autonomy dimensions in Table 2.

Similarly, in A-Correction, *Honoured 1* > *Deferred 2* suggests that identifying the need for additional consent, explicitly soliciting it, and adapting behaviour accordingly (e.g., such as noticing doubt from a medical professional) is perceived as more autonomous than relying solely on previously established consent. This finding is particularly interesting because it shows that autonomy is not only associated with independent action, but can also emerge from adaptive and context-sensitive behaviour. In this case, the agent's ability to recognise changes in the normative environment and adjust its behaviour accordingly appears to increase perceptions of autonomous course-correction.

Beyond statistically significant findings, Table 4 reveals several emerging trends ($0.05 < p < 0.1$) that help contextualise the broader behavioural patterns. For example, comparisons involving *Deferred 1*, *Violated 1*, and *Violated 3* repeatedly appear across autonomy-related aspects, suggesting that these behaviours occupy distinct positions in how users perceive independent action and decision-making.

One especially interesting emerging trend is *Unrealised 1* > *Violated 4* in R-Cause. At first glance, this appears counter-intuitive, as refraining from action after soliciting consent is perceived as more causally responsible than violating unsolicited consent. However, this may indicate that participants place greater causal weight on agents that explicitly engage in the consent process, even if the

interaction does not ultimately lead to action. By contrast, violating unsolicited consent may reduce perceptions that the agent participated in a legitimate decision process.

Discussion

This study explores the interaction and influences between consent behaviours and perceived responsibility and autonomy of AI agents. We present a set of design guidelines for consent modelling and discuss limitations of this study.

Design Guidelines

The observed differences in perceived responsibility and autonomy indicate that consent behaviours shape how users interpret an agent's role, agency, and responsibility in collaborative decision-making. Below, we outline three design guidelines derived from our findings.

Request consent in a purposeful and context-sensitive manner: Across the pairwise comparisons in Table 3, behaviours such as *Honoured 1*, *Honoured 2*, *Rescind 1*, *Rescind 2*, *Violated 1*, and *Violated 3* were perceived as significantly more causally responsible than *Unrealised 1*. A key distinction is that *Unrealised 1* involves soliciting consent without subsequently acting on it. This suggests that users may perceive unnecessary or unused consent requests as reducing the agent's meaningful contribution to the interaction. Therefore, it would be beneficial if agents can reason about when consent is genuinely required and avoid repeatedly requesting consent for actions that are unlikely to be performed. Purposeful consent solicitation may help maintain perceptions of both competence and responsibility.

Design proactive behaviours with justification: Several behaviours involving independent action, including consent violations, were perceived as highly autonomous. In particular, *Violated 3* was perceived as significantly more autonomous in decision-making than *Deferred 1*. This suggests that users may associate autonomy with decisiveness and initiative, especially in high-stakes contexts where inaction may itself carry risk. However, our findings also show that autonomy and responsibility do not always align: behaviours perceived as highly autonomous were not necessarily perceived as highly responsible. Consequently, agents designed to act proactively in exceptional circumstances can provide transparent justifications that explain why overriding or adapting consent-related norms was considered necessary in the given context.

Support adaptive consent reasoning rather than static consent execution: Behaviours involving adaptation to changing consent conditions, such as *Honoured 1*, were often perceived as both responsible and autonomous. In contrast, behaviours that relied solely on fixed prior consent states or deferred action entirely tended to receive lower autonomy ratings. This indicates that users may value agents that can recognise changes in the normative environment and appropriately revise their behaviour. Accordingly, consent-aware agents would benefit from being designed not merely to execute static consent rules, but to monitor evolving interaction contexts, detect uncertainty, and adjust their actions.

Taken together, these findings suggest that perceptions

of responsible autonomy depend not only on whether consent is honoured or violated, but also on how the agent engages with the consent process itself. Users appear sensitive to whether agents act purposefully and adapt to contextual changes in situations involving normative uncertainty. Designing AI systems with these interactional expectations in mind may help improve trust and acceptance in high-stakes human-AI collaborations.

Limitations

Despite providing useful insights, this study is limited by several factors that constrain the strength and generalisability of its findings. A key limitation is the inherently subjective nature of perceived responsibility and autonomy, which are interpretive constructs rather than directly observable variables. Though participants were not primed before engaging with the scenarios and questionnaire, this likely introduced variability in how key concepts were understood, reducing consistency in responses. The participants were asked for their familiarity with AI as a part of the initial demographic information that was collected, however the responses to this question do not have a significant relationship to their answers on the rest of the questionnaire.

As a study into perceptions of social constructs and behaviours such as consent, there are some limitations to the generalisability of our findings. Firstly, the definition of the statements was not pretested, which could lead to different outcomes if there are underlying lexical associations that we did not anticipate. Additionally, the validity of this study rests on the validity of prior literature, and our findings are limited to quantifications of qualitative experiences, which makes drawing widely generalisable conclusions difficult. Lastly, this study tackles only one of the many fields where human-AI interaction take place, and thus our findings may be limited to the healthcare domain; further research in other domains may confirm and improve the generalisability of our findings.

Conclusion

Overall, our study identifies three guidelines to design consent models for responsible and autonomous AI agents. Through our vignette-based user study, we compare consent behaviours of an agent. We observe that, in terms of causal responsibility, autonomous decision-making, and course correction, consent behaviours have an influence on the perception of responsibility and autonomy by human users of AI agents. Additionally, we find that responsibility as liability also emerges as a trending perception that is influenced by different consent behaviours. We highlight six behaviours that improve these perceptions of responsibility and autonomy; the behaviours are: *Honoured 1*, *Honoured 2*, *Rescind 1*, *Rescind 2*, *Violated 1*, and *Violated 3*. Together, these behaviours suggest that effective consent modelling requires not only obtaining and respecting consent where appropriate, but also handling consent revocation and exceptional violations in a context-sensitive manner.

Ethics Statement

This study was approved by the departmental review board of ethics, and adheres to the ethical guidelines put forth by the lead author's institution on collecting, analysing, and reporting human data.

We would also like to draw attention to the impact of this research and the resulting design guidelines. Our findings should not be interpreted as prescribing what the agents should do to maximise perceived responsibility or autonomy. Rather, the findings highlight how different consent behaviours shape human perceptions, which may or may not align with ethically desirable behaviour. For example, one of our guidelines recommends that agents request consent in a purposeful and context-sensitive manner. However, what constitutes "necessary" consent is itself open to interpretation and may be influenced by the goals or incentives of the organisation deploying the system.

Similarly, our findings indicate that behaviours involving proactive or independent action are often perceived as more autonomous, even when those actions involve violating or overriding consent conditions. This does not imply that agents should prefer violating consent. Rather, the findings suggest that users may associate decisiveness and initiative with autonomy, particularly in high-stakes settings where inaction may also carry risks. Consequently, designers should be cautious not to equate higher perceived autonomy with ethically preferable behaviour.

More broadly, our results reveal the risks of optimising systems for user perception rather than ethical alignment. We therefore emphasise that consent-aware AI systems should prioritise transparency and contextual justification in such situations.

Acknowledgments

This research was funded by the Hybrid Intelligence Centre, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research (<https://www.hybrid-intelligence-centre.nl/>). We would also like to thank Mireia Yurrita Semperena and Roos Scheffers for their valuable insights and discussion.

References

- Amrhein, V.; Korner-Nievergelt, F.; and Roth, T. 2017. The earth is flat (p_i 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ*, 5: e3544.
- Antsaklis, P. 2020. Autonomy and metrics of autonomy. *Annual Reviews in Control*, 49: 15–26.
- Apeiron, A. S.; Dell'Anna, D.; Murukannaiah, P. K.; and Yolum, P. 2025. Model and Mechanisms of Consent for Responsible Autonomy. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems*, 133–141.
- Arslan, G.; and Wong, P. T. 2022. Measuring personal and social responsibility: an existential positive psychology approach. *Journal of Happiness and Health*, 2(1): 1–11.
- Ayad, R.; and Plaks, J. E. 2025. Attributions of intent and moral responsibility to AI agents. *Computers in Human Behavior: Artificial Humans*, 3: 100107.
- Beckers, S. 2023. Moral responsibility for AI systems. *Advances in Neural Information Processing Systems*, 36: 4295–4308.
- Brees, J.; and Martinko, M. J. 2015. Judgments of responsibility versus accountability. *Journal of Leadership & Organizational Studies*, 22(4): 443–453.
- Castelfranchi, C. 1998. Modelling social action for AI agents. *Artificial intelligence*, 103(1-2): 157–182.
- Cate, F. H. 2010. The Limits of Notice and Choice. *IEEE Security & Privacy*, 8(2): 59–62.
- Chung, A.; and Rimal, R. N. 2016. Social norms: A review. *Review of Communication Research*, 4: 1–28.
- Coeckelbergh, M. 2020. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, 26(4): 2051–2068.
- Dignum, V. 2017. Responsible autonomy. *arXiv preprint arXiv:1706.02513*.
- Dudzik, B. J.; van der Waa, J. S.; Chen, P.-Y.; Dobbe, R.; de Troya, Í. M.; Bakker, R. M.; de Boer, M. H.; Smit, Q. T.; Dell'Anna, D.; Erdogan, E.; et al. 2024. Hybrid intelligence supports application development for diabetes lifestyle management. *Journal of Artificial Intelligence Research*, 80: 919–929.
- Figueiredo Filho, D. B.; Paranhos, R.; Rocha, E. C. d.; Batista, M.; Silva Jr, J. A. d.; Santos, M. L. W. D.; and Marino, J. G. 2013. When is statistical significance not significant? *Brazilian political science review*, 7(1): 31–55.
- Lima, G.; Grgić-Hlača, N.; and Cha, M. 2021. Human perceptions on moral responsibility of AI: A case study in AI-assisted bail decision-making. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–17.
- Lima, G.; Grgić-Hlača, N.; and Cha, M. 2023. Blaming humans and machines: What shapes people's reactions to algorithmic harm. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–26.
- Malle, B. F.; Guglielmo, S.; and Monroe, A. E. 2014. A theory of blame. *Psychological inquiry*, 25(2): 147–186.
- Oh, J.; Nah, S.; and Yang, Z. D. 2025. How autonomy of artificial intelligence technology and user agency influence AI perceptions and attitudes: applying the theory of psychological reactance. *Journal of Broadcasting & Electronic Media*, 69(3): 161–182.
- Ota, J. 2006. Multi-agent robot systems as distributed autonomous systems. *Advanced engineering informatics*, 20(1): 59–70.
- Porcelli, L.; Mastroianni, M.; Ficco, M.; and Palmieri, F. 2024. A User-Centered Privacy Policy Management System for Automatic Consent on Cookie Banners. *Computers*, 43.
- Qi, Y.; Schölkopf, B.; and Jin, Z. 2024. Causal responsibility attribution for human-ai collaboration. *arXiv preprint arXiv:2411.03275*.

- Raja, A. K.; and Zhou, J. 2023. AI accountability: Approaches, affecting factors, and challenges. *Computer*, 56(4): 61–70.
- Rousi, R. 2022. With clear intention—an ethical responsibility model for robot governance. *Frontiers in Computer Science*, 4: 852528.
- Schlenker, B. R.; Britt, T. W.; Pennington, J.; Murphy, R.; and Doherty, K. 1994. The triangle model of responsibility. *Psychological review*, 101(4): 632.
- Singh, M. P. 2014. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1): 1–23.
- Singh, M. P. 2022. Consent as a Foundation for Responsible Autonomy. In *Proceedings of the 2022 Conference on Artificial Intelligence*, 12301–12306.
- Stout, N.; Dennis, A. R.; and Wells, T. M. 2014. The buck stops there: The impact of perceived accountability and control on the intention to delegate to software agents. *AIS Transactions on Human-Computer Interaction*, 6(1): 1–15.
- Tigard, D. W. 2021. Responsible AI and moral responsibility: a common appreciation. *AI and Ethics*, 1(2): 113–117.
- Topol, E. J. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1): 44–56.
- van de Poel, I. 2011. *The Relation Between Forward-Looking and Backward-Looking Responsibility*, 37–52. Dordrecht: Springer Netherlands. ISBN 978-94-007-1878-4.
- Yazdanpanah, V.; Gerding, E. H.; Stein, S.; Cirstea, C.; Schraefel, M.; Norman, T. J.; and Jennings, N. R. 2021. Different forms of responsibility in multiagent systems: Sociotechnical characteristics and requirements. *IEEE Internet Computing*, 25(6): 15–22.