Model and Mechanisms of Consent for Responsible Autonomy

Anastasia S. Apeiron
Utrecht University
Utrecht, Netherlands
a.s.apeiron@uu.nl

Davide Dell'Anna
Utrecht University
Utrecht, Netherlands
d.dellanna@uu.nl

Pradeep K. Murukannaiah
Delft University of Technology
Delft, Netherlands
p.k.murukannaiah@tudelft.nl

Pınar Yolum
Utrecht University
Utrecht, Netherlands
p.yolum@uu.nl

ABSTRACT

Socio-technical systems rely on human and software agents exercising their autonomy at the right time with the right limits. This requires each agent to know what they can do, when they need help or resources from others, and how they need to interact with others to obtain these resources. To facilitate responsible autonomy, we advocate for the use of *consent* as an abstraction. Although consent has been a part of the software ecosystem, there has been little work to understand its dynamics formally, and to devise mechanisms to use consent in facilitating autonomy. We propose a formal representation of consent based on its philosophical roots, a life-cycle to capture its evolution over interactions, and algorithms to express the consent mechanisms computationally. Following this, we demonstrate how this representation can model and detect various realistic autonomy violations on the web using a real-life example. Finally, we demonstrate a mechanism to dynamically enable consent to regulate the appropriate use of autonomy.

KEYWORDS

Consent; Socio-Technical Systems; Norms; Privacy

ACM Reference Format:
Anastasia S. Apeiron, Davide Dell'Anna, Pradeep K. Murukannaiah, and Pınar Yolum. 2025. Model and Mechanisms of Consent for Responsible Autonomy. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

1 INTRODUCTION

When a software agent and a human interact, we expect the agent to be autonomous and responsible for its actions [16, 38], which means taking ethical and moral consequences into consideration [15]. The agent might handle sensitive information the human needs (e.g., medical records), recommend actions to the human, disclose information about the human to others (agents or humans), and consult others when an action pertains to them. These interactions require regulating the level of autonomy for each agent, especially when making the choice to utilise another agent's resource [40].

Consent is an important construct to regulate autonomy in human interactions [34]. Rather than taking the liberty to act on behalf of others, humans interact first to obtain consent. The required consent varies based on context, the individuals involved, as well as existing norms. Similarly, when an agent is working with a human, it is necessary to identify what consent is needed when and from whom. Clearly, the agent should not request consent for all of its actions as many actions might not require consent. Computing if consent is required, and how to express and manage it, would enable the agent to act on behalf of humans responsibly.

Consent has been part of interactions with software systems for some time. Often, the idea of consent is used in conjunction with privacy. For example, when a human visits a website, they likely give consent for the system to collect personal information, share with other systems, and so on, as described by the privacy policy of that website. The General Data Protection Regulation (GDPR) regulates how such consent should be obtained in different situations. To comply with the GDPR, many websites employ a Consent Management Provider (CMP), which collects the users' consent information and shares it with other websites when necessary [21]. Generally, these systems pertain only to the usage of that specific website, which limits consent management to a single domain. Moreover, these CMPs operate as a bookkeeping service, where they record and lookup consent for different websites. Current literature has focused on important aspects such as realising CMPs securely over blockchains [2], checking if any actions against consent have been taken [9], and providing intuitive user interfaces to obtain consent [20]. However, existing CMPs are not equipped to handle the sort of interactions involving autonomous agents and multiagent systems. In such systems, consent must regulate not only data sharing, but also the reactive, proactive, and social behaviours of agents.

A related line of work is that of privacy assistants, which are autonomous agents that help users make decisions on giving consent [24, 39], for example when they are interacting with CMPs. These agents can learn the user's privacy preferences over time, make suggestions on how information should be shared [18, 23], or explain to the user why certain information should be kept private [29]. This body of work focuses on the important supporting concepts of consent but does not always provide mechanisms to understand whether one consent implies other consents, whether consent has been obtained using the right requirements, or how consent propagates from one actor to another in the system.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Accordingly, we propose a model of consent and outline its mechanisms for interactions between human and computational agents, as well as between computational agents that represent humans. Our proposed model considers essential aspects of consent as identified in the philosophy literature of consent between humans and couples that with the literature on consent management in information systems. We also provide a life-cycle of consent so that an agent—apart from requesting and granting consent—can keep track of the state of the consent it is associated with and can anticipate impending consent violations. Additionally, we express the workings of our model through a set of algorithms, and apply them to realistic scenarios of consent violations, and explore how a richer representation of consent facilitates responsible autonomy in human-AI collaborative systems [14].

Running example. To illustrate the components of our consent model, we introduce a running example: Case 122 from the AI Incidents Database [28]. A short description of Case 122 in the database is:

> "*Facebook's initial version of its Tag Suggestions feature, where users were offered suggestions about the identity of people's faces in photos, allegedly stored biometric data without consent, violating the Illinois Biometric Information Privacy Act*"

In this incident, Facebook stored the biometric data scanned from photos for their Tag Suggestions feature without user consent, which is prohibited by the Illinois Biometric Information Protection Act (BIPA). Barring malicious intent, how could Facebook (or an autonomous agent acting on its behalf) have avoided this incident? This turns out to be a non-trivial question. This leads to further questions such as: What was Facebook's goal? What resources did Facebook require to achieve that goal? Who had sovereignty over the required resources? What norms and sanctions governed the usage of the required resources? Answering these and similar questions requires a systematic understanding of the notion of consent (i.e., a model) and mechanisms to operationalise this understanding in a multiagent system.

Section 2 reviews related work. Section 3 provides a formalisation of consent and its life-cycle. Section 4 discusses automated mechanisms that enable autonomous agents to reason about important edge-cases of interactions that involve consent. Section 5 includes additional discussion and conclusions.

2 RELATED WORK

To situate our model in the current literature, we first explore the constituents of consent from its philosophical roots in human-human interaction followed by what current research on norms in socio-techinical systems (STS) have accomplished thus far.

2.1 Consent Management between Humans

To develop a model of consent for agents representing humans in a system, we must first understand how consent moderates interactions among humans. There are three prominent philosophical views on consent: the attitudinalist view, the communicationist view, and a hybrid view containing elements of the other two views.

The attitudinalist view argues that consent is a mental state that allows a specified legal or moral boundary-crossing [4], while the communicationist view argues that consent is a behavioural act derived from the mental act of consent by an agent [17]. The hybrid view states the necessity of, and distinction between, both a mental act and a behavioural act from an agent to constitute consent [19].

We follow the hybrid view as the foundation for our model. This approach allows an agent to identify and manage consent based on a behavioural act (e.g., verbal or gesture) presented, and also capture a distinct mental act (e.g., goals). Additionally, a consent that is given through a behavioural act can only be modified or revoked through the same method, a behavioural act [19]. This prevents the consent being changed or revoked without both agents being aware of its modification. If both agents are not aware of changes in the consent, this can lead to deceptive practices, which diminishes the autonomy of the agent being deceived. Examples of deceptive practices that diminish autonomy are threats, where an agent can be threatened with violating the given consent while they were not aware of its modification [6].

2.2 Computational Consent Management

There has been research on building domain-specific consent management systems [30, 37]. For example, there is a rule-based consent management system to ease the cognitive load of cookie banners on users when navigating websites [30]. This approach of developing rule-based systems is useful to combat dark patterns such as cognitive overloading that 'nudge' users to make a decision that may not reflect their own desires or wishes. Such static approaches do not capture a generalisable model of consent with a well-defined life-cycle; thus, they cannot easily be applied to other domains or enable agents to exercise their autonomy responsibly.

Additionally, the rules of consent are generated by a human decision-maker, and do not allow for an agent utilising these rules to adapt to a changing normative environment. This limits an agent's autonomy and can lead to consent violations that could have been prevented if the agent had been able to adapt their policy accordingly [36]. Ideally, in an autonomous system, an agent is able to reason about the need for consent in relation to the current social norms and decide that a norm deviation is necessary [32]. This requires deliberation on norms [11, 31], which agent holds the resource they need, and if they can execute an action that produces a different output than what was consented for.

2.3 Consent and Social Norms

Consent has normative power [22, 26], and a domain where an agent can exercise normative power is their sovereignty, which includes their corpus, the goods they own, and knowledge they possess, as protected by legal and moral rights [4, 8]. Consent generally takes place between at least two distinct agents for a specified purpose, using specified resources. An agent cannot simultaneously give and receive consent for the same resource [17, 19], as a resource is either under the agent's sovereignty or not. This dichotomy also supports the two different kinds of consent: solicited and unsolicited consent. 'Solicited consent' refers to a situation where an agent requests consent for a resource that is not under their sovereignty, while 'unsolicited consent' occurs when an agent gives consent

for a resource that is under their sovereignty without being asked. By granting or denying consent, an agent can waive or enforce a norm that exists in their STS. For example, this can be the norm of respecting the sovereignty that an agent has over their resource, e.g., a property such as a garden, that can be waived specifically for the purposes of letting others join a garden party [22].

In this paper, we adopt the model of social norms (which includes the norm types of authorisations, commitments, prohibitions, power, and sanctions) from [33]. Although social norms and consent are related, consent is not merely a norm type. Instead, we formulate consent as an action that determines a set of norms that manipulate the normative consequences of social norms in an STS. We propose consent as a way for agents to negotiate and determine the norms applicable to them within the STS [7], as opposed to a more common approach involving the enforcement of norms by an authority external to the agents themselves—this allows for norms to emerge based on the interactions between the agents [1].

3 CONSENT MODEL

In this section, we describe our formal model of consent, beginning with the desiderata for a model of consent.

3.1 Desiderata for a Consent Model

Based on the literature in Section 2, we define consent as follows.

Consent: *An action that removes the normative consequences for the purposeful and conditional infringement of a social norm that protects an agent's sovereignty.*

We determine how to evaluate the suitability of our model to express consent from a human perspective by outlining four important nuances of human consent and their functions.

- **D1. Behavioural Necessity and Sufficiency** An action is required to instate the consent that alters the consequences of a social norm [26]. This act of consent does not directly indicate the mental phenomenon of consent, such as in insincere consent where a behavioural act of consent does not have to be indicative of the mental act of consent [27]. One example from [26] is the case of the abandoned suitcase at an airport. If the suitcase is taken by someone without explicit consent from the owner of the suitcase (who has mentally abandoned it), is it still a culpable offence?
- **D2. Uniqueness of Consent** A consent must be constrained to a specific agent, action, goal, and norms [26], and any changes to these components constitute a new consent. In the example from [25], an AI recommender system gives shopping recommendations based on the user's previous behaviour on the website, and deduces that the user is pregnant and suggests pregnancy-related recommendations before the user makes it explicit that they wish to receive these recommendations, violating the uniqueness property.
- **D3. Assumption of Feasibility** A consent must be feasible, such that, the agent must be able to fulfil their stated goal following the norms provided by the consent [35]. In another example from [26], consent is given by a motorcycle owner to their friend to ride their motorcycle knowing that the motorcycle is broken and the action can never be executed. This would violate the assumption of feasibility.
- **D4. Temporality of Consent** A consent can only be changed or revoked through the actions to do so, regardless of the passage of time. Coupled with **D1**, the example in [19] illustrates a case where a parent goes to sleep after giving consent for their child to drive their car. Since the consent was given via a behavioural act, was not explicitly modified, and the world has not changed to prevent the child from driving the car, the consent still holds, even though it is not continually given.

We propose a model of consent that allows for these nuances and can address the difficult areas of consent between humans, such as insincere consent.

3.2 Proposed Consent Model

We consider a STS in which human and artificial agents operate and interact to achieve their goals. Generally, an agent attempts to achieve a goal by means of a plan, which is a sequence of actions, that is expected to lead the STS to a state where the goal is satisfied.

The behaviour of agents in the STS is regulated by social norms. Some of these norms protect the agents' sovereignty over their resources, i.e. establish the resources over which agents have legal or moral rights. For example, in the US state of Illinois (the STS from Case 122 in our example), the BIPA establishes that people have sovereignty over their biometric data.

Notation. We represent a STS using the following notation:

- $\mathbb{A}$ is the set of agents that operate and interact in the STS.
- $\mathbb{L}$ is the set of norms that hold in the STS. $\mathbb{L}$ may include a set of norms that protects an agent's sovereignty (i.e., gives them exclusive rights) over their resources, such as prohibitions to infringe on another agent's resources. We assume that a priority mechanism is in place to ensure that the most recent norms are given priority in case of conflict with older ones.
- $\mathbb{R}_A$ is the set of resources under the sovereignty of an agent $A \in \mathbb{A}$. $\mathbb{R} = \{\mathbb{R}_A \mid A \in \mathbb{A}\}$ is the set of all the resources of agents in the STS. For simplicity, in this paper we assume that a single agent has sovereignty over a given resource.
- $\mathbb{G}$ is the set of propositional atoms used to represent goals of agents in $\mathbb{A}$. We call a goal that an agent has expressed to another agent a *stated goal*.
- Φ is a propositional language with the standard operators and a set of propositional atoms $\Omega \cup \mathbb{G}$.
- $\mathbb{T}_A$ indicates the set of possible actions that can be executed by an agent $A \in \mathbb{A}$. We represent an action $t \in \mathbb{T}_A$ as a tuple $\langle p, r \rangle$, where $p \subseteq \Omega$ is the (non-empty) post-condition (i.e., the effects) of t, i.e., the set of propositional atoms that becomes true in the state $\mathbb{S}$ of the STS after the execution of t, and $r \in \mathbb{R}$ is a resource that is required or affected by t (possibly none if the action does not affect any resource).
- $\mathbb{S}$ is the state of the STS, i.e., a propositional assignment. We use $\mathbb{S} \vDash \phi$ to indicate that a formula $\phi \in \Phi$ is true in $\mathbb{S}$.

We say that an agent R (*consent receiver*) needs consent from an agent G (*consent giver*) if R intends to execute an action $\langle p, r \rangle$ that requires or affects a resource $r \in \mathbb{R}_G$ that is within the sovereignty of G, in order to achieve a stated goal $g_R \in \mathbb{G}$.

We assume that two agents G and R in the STS can interact and negotiate to agree upon the details of the consent[1]. We characterise the agreed details of consent as a *set of norms* $\mathbb{N}$ that affects the legal and/or moral relations between G and R, effectively amending the set of norms $\mathbb{L}$ that hold in the STS. When a negotiation begins, a *consent instance* (Def. 3.1) is initialised. A negotiation ends when G and R come to an agreement about the norms $\mathbb{N}$.

We define the set of negotiated and agreed upon norms as the set $\mathbb{N} = \{AU, CO\} \cup \mathbb{N}'$ composed of at least one authorisation AU and one commitment CO. The authorisation $AU = \langle G, R, c, \langle p, r \rangle \rangle$ authorises R to perform an action $\langle p, r \rangle \in \mathbb{T}_R$ that affects a resource $r \in \mathbb{R}_G$ under the sovereignty of G, given that a condition $c \in \Phi$ holds (i.e. $\mathbb{S} \vDash c$). We define a condition of an authorisation as a propositional formula s.t. $c = c_{det} \wedge \neg(c_{exp})$, with $c_{det} \in \Phi$ specifying the detachment condition of the authorisation (i.e., when the authorisation is active) and $c_{exp} \in \Phi$ specifying the expiration condition of the authorisation (i.e., when the authorisation is no longer active). The commitment $CO = \langle R, G, p, g_R \rangle$ describes a commitment made by the consent receiver R to G to bring about its stated goal g_R if the antecedent $p \in \Phi$ (same as the post-condition of $\langle p, r \rangle$ in AU) holds.

We impose the requirement of at least one commitment alongside the authorisation to encourage the expression of the stated goal, enabling situationally accurate norms in the consent instance. However, in certain cases—such as unsolicited consent, where an agent pre-emptively authorises another agent to perform an action—a commitment may not be necessary and, therefore, may not be part of the consent instance.

The set $\mathbb{N}'$ may include additional norms including any other authorisation, commitment, or prohibition agreed between G and R during the negotiation phase. For instance, the agents could agree on an additional prohibition that explicitly prohibits performing specific non-authorised actions that might affect resource r or to use resource r for a stated goal different than g_R. For instance, an agent may consent to another agent borrowing their car for the next few days to move their belongings from one house to another, but may prohibit the agent from parking the car in front of the current house (e.g., because R may currently live in a dangerous neighbourhood).

Given the set of norms $\mathbb{N}$ and a stated goal g_R, we can now formally define a *consent instance* as follows.

Definition 3.1 (Consent Instance). A consent instance is a tuple $\langle G, R, \mathbb{N}, g_R, t \rangle$ s.t. $G, R \in \mathbb{A}$ are, respectively, the consent giver and consent receiver agents, $\mathbb{N}$ is the set of norms negotiated between G and R that affect their (legal and/or moral) relationship, and $g_R \in \mathbb{G}$ is R's stated goal for executing an action t that requires consent from G.

A consent instance may change over time only through the progression of the states of the norms (e.g., they are violated) as well as events in the MAS. We characterise the life cycle of a consent instance in terms of its key states and transitions. Figure 1 illustrates these. Over time, an instance of consent can be in one of seven states: $\Sigma = \{\sigma_n, \sigma_d, \sigma_a, \sigma_u, \sigma_r, \sigma_h, \sigma_v\}$.

[1]We use the term negotiation to refer to the interaction phase that makes consent active without restricting such interaction to a specific form. The negotiation may, for example, resemble a standard consent request via a web interface or dialogue.

- State σ_n (Negotiating): the agents G and R have started but not concluded the process of negotiating the norms $\mathbb{N}$, i.e., the set $\mathbb{N}$ is not finalised yet and the norms in $\mathbb{N}$ are not active yet.
- State σ_d (Deferred): a terminal state where agents G and R could not reach an agreed set of norms $\mathbb{N}$.
- State σ_a (Active): the negotiating phase between agents G and R is concluded. The set of agreed norms $\mathbb{N}$ are active.
- State σ_u (Unrealised): a terminal state where the action t authorised by $AU \in \mathbb{N}$ has not been executed and the expiration condition has been reached.
- State σ_r (Renegotiate): a terminal state where an agent has withdrawn their agreement to the consent norms.
- State σ_h (Honoured): a terminal state where all norms in $\mathbb{N}$ are fulfilled. This also implies that commitment $CO \in \mathbb{N}$ is fulfilled, i.e., that the stated goal g_R is achieved.
- State σ_v (Violated): a terminal state where at least one of the norms in $\mathbb{N}$ has been violated.

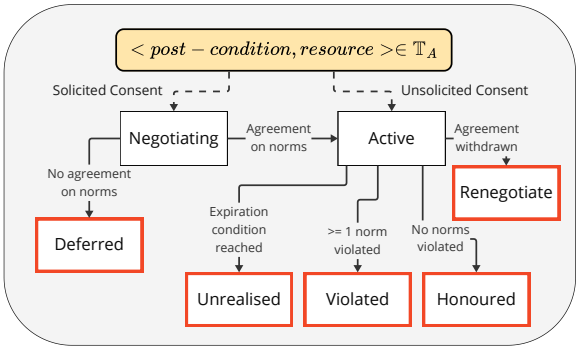

Figure 1: The life-cycle of a consent instance. The dashed lines indicate if the agent is soliciting consent to execute the action represented in the topmost box, or giving unsolicited consent to execute the action. Following from this, the consent progresses through the states until it reaches the terminal states with thicker borders.

From the Negotiating state σ_n, the consent instance can transition into either the Active state σ_a or to the Deferred state σ_d depending on their agreement on the norms or not, respectively. From the Active state σ_a, the consent instance can transition (i) into the Unrealised state σ_u if the expiration condition of the authorisation is reached but the authorised action has never been executed, (ii) into the Renegotiate state σ_r if the consent instance is terminated by an agent (e.g., request to change some norms), (iii) into the Honoured state σ_h if all the norms have been fulfilled, and (iv) into the Violated state σ_v if at least one norm has been violated. In line with Singh [33], we say that an authorisation $AU = \langle G, R, c, \langle p, r \rangle \rangle$ is violated in a state of the MAS $\mathbb{S}$ if agent R executes the action $\langle p, r \rangle$ even though the authorisation's detachment conditions c_{det} in c remain false or after the authorisation's expiration condition becomes true (i.e., if $\mathbb{S} \nvDash c_{det} \vee \mathbb{S} \vDash c_{exp}$); while a commitment $CO = \langle R, G, p, g_R \rangle$ is violated when the consequent g_R does not

hold after the antecedent p holds (i.e., $\mathbb{S} \models p$ and $\mathbb{S} \not\models g_R$). Violation of other types of norms, such as prohibitions or obligations, depends on their definition (e.g., [3, 33]).

3.3 Modelling the Running Example

We illustrate our model of consent using the running example. We characterise the STS as a Multiagent System (MAS) involving two agents $\mathbb{A} = \{F, U\}$, representing Facebook and a User, respectively. We represent the User's photo and account (the user's profile on Facebook) as resources, i.e., $\mathbb{R}_U = \{photo, account\}$. We assume, for the sake of this example, that the Facebook agent does not have resources, i.e., $\mathbb{R}_F = \{\varnothing\}$.

We model Facebook's objective to offer Tag Suggestions as the stated goal *suggestTag* from the set $\mathbb{G} = \{suggestTag, collectBioData\}$. We model Facebook's possible actions as the set

$$\mathbb{T}_F = \begin{Bmatrix} \langle scannedPhoto, photo \rangle, \\ \langle storedBioData, photo \rangle \\ \langle deletedAccount, account \rangle \end{Bmatrix}$$

with $\{scannedPhoto, storedBioData, deletedAccount\} \subseteq \Omega$.

These characterise, respectively, the action of scanning a photo, the action of storing biometric data obtained from the photo, and the action of deleting a personal account from Facebook.

We model the User's possible actions as the set

$$\mathbb{T}_U = \begin{Bmatrix} \langle uploadedPhoto, photo \rangle, \\ \langle deletedAccount, account \rangle \end{Bmatrix}$$

with $uploadedPhoto \in \Omega$. These characterise, respectively, the action of uploading a photo, and the action of deleting a personal account from Facebook.

Finally, we model the prohibitions enforced in the Facebook's STS as the set of prohibitions $\mathbb{L} = \{P(A, t) | A \in \mathbb{A}, t = \langle p, r \rangle \in \mathbb{T}_A, r \notin \mathbb{R}_A\}$ that prohibits any agent $A \in \mathbb{A}$ that does not own a resource r (i.e., $r \notin \mathbb{R}_A$) to execute any action t that requires or affects r under the sovereignty of another agent $A' \neq A$. Specifically, given the two considered sets of actions $\mathbb{T}_F$ and $\mathbb{T}_U$, the set $\mathbb{L}$ corresponds to the following set of prohibitions for Facebook:

$$\mathbb{L} = \begin{Bmatrix} P(F, \langle deletedAccount, account \rangle), \\ P(F, \langle scannedPhoto, photo \rangle), \\ P(F, \langle storedBioData, photo \rangle) \end{Bmatrix}$$

These prohibitions express that Facebook is prohibited from deleting a User's account, from scanning a photo, and from storing biometric data obtained from a photo. These norms contain both the prohibition expressed by the BIPA (storing biometric data) and other prohibitions exemplifying expectations from the use of the Facebook platform. These actions can only be done if there is consent from U as the owner of *photo*.

In this example, the consent instance would be expressed as:

$$ci_{122} = \langle U, F, \mathbb{N}, suggestTag, \langle scannedPhoto, photo \rangle \rangle,$$

where U is the consent giver, F is the consent receiver, and $\mathbb{N} = \{AU, CO\}$ is the set of norms of the consent negotiated and agreed between F and U, such that

$$AU = \left\langle \begin{matrix} U, F, uploadedPhoto \wedge \neg deletedAccount, \\ \langle scannedPhoto, photo \rangle \end{matrix} \right\rangle$$

is an authorisation from the User to Facebook to scan the photo once uploaded, until the user's account is not deleted, and

$$CO = \langle F, U, scannedPhoto, suggestTag \rangle$$

is a commitment from Facebook to the User to suggest a tag once they have scanned the photo.

We note that the prohibition for Facebook to store biometric data obtained from a photo expressed in the norms $\mathbb{L}$ is still valid, since the consent agreement did not amend it, i.e. there is no authorisation in $\mathbb{N}$ that amends $P(F, \langle storedBioData, photo \rangle)$. From this modelling of Case 122 from the AI Incidents Database [13], we can see that the consent violation comes from Facebook storing biometric data obtained from the photo of the user even though this was not explicitly authorised (consented) by the user.

4 CONSENT MECHANISMS

Given the formal model of consent, we describe the mechanisms necessary to operationalise our proposed model in a STS through a set of algorithms. Section 4.1 describes how agents can determine whether consent is needed to execute an action (Algorithm 1). Section 4.2 discusses how agents can solicit consent and create an instance of consent (Algorithm 2). Section 4.3 describes consent-based reasoning (Algorithm 3), and how to monitor and update the state of active consent instances (Algorithm 4). To illustrate the mechanisms presented in the algorithms we utilise nine use cases that build on top of the running example modelled in Section 3.3.

4.1 "Do I Need Consent?"

An agent can determine whether it needs consent to execute an action by determining a resource's owner and by examining the norms in $\mathbb{L}$ that protect agent sovereignty. Algorithm 1 describes a mechanism to determine who should give consent for an action $t = \langle p, r \rangle$ that an agent R intends to execute. The function determines the owner G of the resource r by invoking function findResourceOwner, which determines the agent $A \in \mathbb{A}$ s.t. $r \in \mathbb{R}_A$. If R is the owner itself of the resource or if R has previously received consent by G to perform the action t (i.e., the boolean function hasConsent(R,G,$\mathbb{L}$,t) determines that the norms $\mathbb{L}$ contain an authorisation obtained by R from G to perform the action t), then the function returns R. We illustrate this in the use cases below.

Algorithm 1 Determining the agent who needs to give consent to perform an action

Input: agent R invoking the function, action $t = \langle p, r \rangle$, set of norms $\mathbb{L}$, set of agents $\mathbb{A}$
Output: the agent to which consent should be asked for consent to execute the action t (or the agent itself, R, if consent is not required)

```
1: function getConsentGiver(R,t,L,A)
2:     G ← findResourceOwner(r, A)
3:     if G = R or hasConsent(R,G,L,t) then
4:         return R
5:     return G
```

Use Case 1. (Non-necessity of consent) *U deletes their account before uploading a photo.*

If an agent wishes to execute an action that requires a resource under their sovereignty, then they do not require consent. Executing

the function getConsentGiver(U,$\langle deletedAccount, account\rangle$,$\mathbb{L}$,$\mathbb{A}$) from Algorithm 1 returns the value U, indicating that the agent does not require consent from another agent to execute the action because the account is under their sovereignty, i.e., $account \in \mathbb{R}_U$, determined by function findResourceOwner.

Use Case 2. (Using another agent's resource) *F intends to scan a photo uploaded by U.*

Executing the function getConsentGiver(F,t,$\mathbb{L}$,$\mathbb{A}$), with action $t = \langle scannedPhoto, photo\rangle$, returns the value U because invoking findResourceOwner($photo$, $\mathbb{A}$) outputs U and because there is no norm in $\mathbb{L}$ that authorises F to execute t, i.e., the function hasConsent(F, U, $\mathbb{L}$, t) returns *False*. If the action t was to be executed by F without consent from U, then F would be violating the norms in $\mathbb{L}$.

Unless permitted by norms in $\mathbb{L}$, an agent can have consent to execute an action that affects some other agent resource only in two cases: if another agent has previously given *unsolicited consent* (discussed below), or because of a previous consent solicitation (discussed in Section 4.2).

Use Case 3. (Giving unsolicited consent) *When U's account is hacked, U communicates to F that F can delete U's account because they won't use it any more.*

In our model, unsolicited consent can be achieved via updating the MAS norms $\mathbb{L}$ with a new instance of consent. In this use case, an agent U can create a consent instance $\langle U, F, \{AU\}, \top, t\rangle$, where $AU = \langle U, F, c, t\rangle$ with authorisation condition $c = c_{exp} = \neg deletedAccount$ and authorised action $t = \langle deletedAccount, account\rangle$, and $\top$ indicates that the consent instance does not require F to state a goal. The instance indicates that U gives consent to F to delete U's account. By updating the norms in $\mathbb{L}$ with the new consent instance (see function update described in Section 4.2), the action $\langle deletedAccount, account\rangle$ is no longer prohibited. As a consequence, after U gives unsolicited consent to F to delete their account, executing Algorithm 1 the function getConsentGiver(F,$\langle deletedAccount, account\rangle$,$\mathbb{L}$,$\mathbb{A}$) returns value F.

4.2 Soliciting Consent

Algorithm 2 describes the mechanism for the solicitation of consent by an agent R from an agent G. The function first creates an instance of consent ci, initialised with the Negotiating state (σ_n). Then it begins a negotiation between agents G and R to negotiate the terms of consent for executing the action t (function negotiate). To negotiate for consent, the agent R is required to state its goal g_R for executing the action t, clarifying the purpose for requesting consent. If the negotiation succeeds ($nres = success$), the function negotiate returns a set of norms $\mathbb{N}$ agreed by agents R and G that regulate the conditions of consent, and the (possibly revised) goal agreed between the agents.

The instance ci of consent is then updated with the negotiated norms, and its state ci_{state} transitions to the Active state σ_a. If the negotiation does not succeed, the instance of consent is left at its initial state (with an empty set of norms), and the state of the consent instance moves to Deferred (σ_d).

Use Case 4. (Active consent instance) *F is aware of BIPA and wants to avoid a violation of the law when scanning the photo of the user. F solicits consent from U to scan the photo to suggest a tag and collect biometric data. U authorises F to scan the photo for the goal of suggesting a tag, but not for collecting biometric data. Both parties agree with the conditions.*

Algorithm 2 Soliciting consent from agent G to execute the action t for the stated goal g_R

Input: agent R soliciting the consent, stated goal g_R, action t, agent G to which consent is solicited
Output: pairing of the consent instance and its state

```
1:  function solicitConsent(R,g_R,t,G)
2:      ci ← ⟨R, G, ∅, g_R, t⟩
3:      ci_state ← σ_n
4:      ⟨ℕ, nres, g'_R⟩ ← negotiate(R, G, g_R, t)
5:      if nres = success then
6:          ci ← ⟨R, G, ℕ, g'_R, t⟩
7:          ci_state ← σ_a
8:      else
9:          ci_state ← σ_d
10:     return ⟨ci, ci_state⟩
```

The solicitConsent function initiates an instance of consent and a negotiation between F and U for the execution of the action $t = \langle scannedPhoto, photo\rangle$ with the stated goal $g_R = suggestTag \wedge collectBioData$. Since F agrees to the norms $\mathbb{N}$, which does not allow $collectBioData$ to become true, the negotiation yields a consent instance $ci = \langle F, U, \mathbb{N}, g'_R, t\rangle$, with the set of norms $\mathbb{N}$ defined in Section 3.3, where $g'_R = suggestTag \wedge \neg collectBioData$. The state ci_{state} is updated into the Active state σ_a, and the pair $\langle ci, ci_{state}\rangle$ is returned.

Use Case 5. (Deferred consent instance) *F is aware of BIPA and wants to avoid a violation of the law when scanning the photo of the user. F solicits consent from U to scan the photo to suggest a tag and collect biometric data. U authorises F to scan the photo to suggest a tag, but not to collect biometric data. However, F refuses the latter condition set forth by U not to collect biometric data.*

Given that all of the norms must be agreed upon before the consent becomes active, failure to negotiate (i.e., $nres \neq success$, due to F's refusal of the conditions set by U) results in the state of the consent instance ci_{state} being updated into σ_d (Deferred), and the function solicitConsent returns the pair $\langle ci, ci_{state}\rangle$, where the consent instance ci has an empty set of norms $\mathbb{N} = \emptyset$. As a consequence, no norm in $\mathbb{L}$ is amended, and F is still not allowed to scan U's photo.

Algorithm 3 puts together Algorithms 1 and 2 as the procedure consentBasedReasoning, where an agent R can reason about the need for consent and solicit consent from another agent for executing an action t to achieve a goal g. Line 5 of Algorithm 3 invokes a function determineStatedGoal, which determines which goal the agent will state during the negotiation with the other agent in order to solicit consent. This function supports the fact that agents may publicly communicate a different goal than their internal one. Line 8 invokes a function update, which updates $\mathbb{L}$ with the newly created instance of consent, in case of a successful negotiation. This enables the agents to consider the norms they agreed upon in the consent instance in their next consent-based reasoning, and to

monitor the state of consent over time. Based on the state of the consent given as input to the function UPDATE (e.g., if the consent instance is created, or revoked), the function may add or remove norms to $\mathbb{L}$, respectively.

Algorithm 3 Simple consent-based reasoning

Input: agent R performing consent-based reasoning, R's goal g, set of norms $\mathbb{L}$, action t, set of all agents $\mathbb{A}$
Output: *True* if the R has consent to execute the action t, *False* otherwise

1: **function** CONSENTBASEDREASONING(R,g,t,$\mathbb{L}$,$\mathbb{A}$)
2: $has\_consent\_to\_exec \leftarrow True$
3: $G \leftarrow$ GETCONSENTGIVER($R, t, \mathbb{L}, \mathbb{A}$)
4: **if** $(G \neq R)$ **then**
5: $g_R \leftarrow$ DETERMINESTATEDGOAL(R, g, t, G)
6: $\langle ci, ci_{state} \rangle \leftarrow$ SOLICITCONSENT(R, g_R, t, G)
7: **if** $ci_{state} = \sigma_a$ **then**
8: UPDATE($\mathbb{L}, \langle ci, ci_{state} \rangle$)
9: **else**
10: $has\_consent\_to\_exec \leftarrow False$
11: **return** $has\_consent\_to\_exec$

Algorithm 4 Monitoring and updating consent instances states

Input: state of the MAS $\mathbb{S}$, set of norms $\mathbb{L}$
Output: updated state of the consent instance and state of the MAS $\mathbb{S}$

1: **function** UPDATECONSENTINSTANCES($\mathbb{S}$,$\mathbb{L}$)
2: **for** $\langle ci, \sigma_a \rangle \in \mathbb{L}$ s.t. $ci = \langle R, G, \mathbb{N}, g_R, t \rangle$ **do**
3: $\langle G, R, c, \langle p, r \rangle \rangle \leftarrow$ GETAUTHORISATIONFROM($\mathbb{N}$)
4: $c_{exp} \leftarrow$ GETEXPCOND(c)
5: $new\_consent\_state \leftarrow \sigma_a$
6: **if** $(\mathbb{S} \vDash c_{exp}) \wedge \mathbb{S} \nvDash p$ **then**
7: $new\_consent\_state = \sigma_u$ ▷ Unrealised
8: **else if** $\exists n \in \mathbb{N} \mid$ VIOL($n, \mathbb{S}$) **then**
9: $new\_consent\_state = \sigma_v$ ▷ Violated
10: **else if** $\forall n \in \mathbb{N}$: SAT($n, \mathbb{S}$) **then**
11: $new\_consent\_state = \sigma_h$ ▷ Honoured
12: **else if** $stated\_withdrawal\_ci \in \mathbb{S}$ **then**
13: $new\_consent\_state = \sigma_r$ ▷ Renegotiate
14: UPDATE($\mathbb{L}, \langle ci, new\_consent\_state \rangle$)

4.3 Monitoring Consent

An agent can monitor and update the state of an active consent instance by monitoring the state of the STS and the state of the norms agreed upon for the consent instance.

Algorithm 4 illustrates a mechanism that can be employed in a MAS to update, when needed, the state of an active consent instance, based on monitored changes in the state of the STS. The algorithm describes how a consent instance can transition from the Active state into the Unrealised (σ_u), Violated (σ_v), Honoured (σ_h), or Renegotiate (σ_r) states, as described in Section 3. Function GETEXPCOND(c) returns the expiration condition c_{exp} of the authorisation's condition c. The withdrawal of agents from their previous agreement for consent (necessary to transition into the Renegotiate state (σ_r)) is expressed in lines 12-13 via the presence of a propositional atom *stated_withdrawal_ci* in the STS state (e.g., resulting from a communication from one agent to another).

USE CASE 6. (Consent expires) *F received consent from U to scan U's photo to suggest a tag according to Use Case 4, but the user deletes its account before F scans the photo.*

The deletion of the account from U results in the proposition *deletedAccount* becoming true in the state of the MAS, i.e., $\mathbb{S} \vDash deletedAccount$. When monitoring the state of active consent instances, Algorithm 4 retrieves the authorisation AU, whose condition is $c = uploadedPhoto \wedge \neg(deletedAccount)$, with expiration condition $c_{exp} = deletedAccount$. The if-statement in line 5 is satisfied since both the authorisation has expired ($\mathbb{S} \vDash deletedAccount$) and the photo has never been scanned ($\mathbb{S} \nvDash scannedPhoto$). This indicates that the authorisation has expired. The state of the consent instance is updated to σ_u.

USE CASE 7. (Consent violation) *F received consent from U to scan U's photo to suggest a tag according to Use Case 4. F scans the photo of U but does not suggest a tag.*

After executing the action $\langle scannedPhoto, photo \rangle$, F is expected (due to the commitment $CO \in \mathbb{N}$) to suggest a tag. Function VIOL in line 8 of Algorithm 4 determines that commitment $CO \in \mathbb{N}$ is violated because $\mathbb{S} \vDash scannedPhoto$ and $\mathbb{S} \nvDash suggestTag$. Since a norm in $\mathbb{N}$ is violated, the state of the consent instance is updated to the Violated state σ_v.

USE CASE 8. (Honouring the consent) *F received consent from U to scan U's photo to suggest a tag according to Use Case 4. F scans the photo of U, suggests a tag, and does not store biometric data from U's photo.*

Since all norms in $\mathbb{N}$ are satisfied and the stated goal is achieved ($\mathbb{S} \vDash (suggestTag \wedge \neg collectBioData)$), the consent instance is honoured. Algorithm 4 determines this by invoking function SAT for all norms in $\mathbb{N}$ to verify if the norm is satisfied in a state of the MAS. The state of the consent instance is then updated to the Honoured state σ_h.

USE CASE 9. (Revoking consent) *F received consent from U to scan U's photo to suggest a tag according to Use Case 4, but the user edits its preference in the Facebook's profile Setting page before F scans the photo.*

Whenever an agent retracts their agreement or changes the norms of a consent instance like in the current use case, we assume that this results in a proposition *stated_withdrawal_ci* becoming true in the MAS state $\mathbb{S}$. Accordingly, in lines 12-13 Algorithm 4 updates the state of the consent instance to the Renegotiate state σ_r and updates the norms in $\mathbb{L}$ by revoking the consent instance (function UPDATE, line 14). We call this state Renegotiate because we support the possibility that the norms previously agreed upon can be renegotiated. F also holds the same power to revoke consent, as both agents must be in agreement with respect to the consent at all times until the consent instance terminates.

5 DISCUSSION

The model and mechanisms we propose are a rich representation of consent for an STS involving human and artificial agents. In this section, we reflect on the strengths and limitations of our work.

5.1 Revisiting the Consent Desiderata

In Section 3.1, we formulated the desiderata for a consent model. By satisfying these desiderata, our model identifies consent violations and anticipate different states of consent based on the actions of an agent. Our model allows an agent to check if consent exists or is needed, identify and solicit the necessary consent, adapt to changes that affect consent, and determine if and why a consent violation has occurred. Additionally, our model enables to identification of cases involving insincere consent and deception, and determine culpability of the agents involved. We argue that taking into account these nuances, an agent can identify and reason about consent in a way that fosters responsible autonomy.

5.1.1 D1: Behavioural Necessity and Sufficiency. Our model satisfies the desideratum **D1** by having an explicit set of norms $\mathbb{N}$ expressed in a consent instance. Satisfying **D1** holds agents accountable to only the explicitly stated norms, and prevents an agent from relying on a mental act of modifying/revoking consent to avoid culpability.

In Use Case 2, only relying on the mental act would mean that if $photo \in \mathbb{R}_U$ and F executed the action $\langle scannedPhoto, photo \rangle$, we have no way of verifying the existence of consent that permits the action, or determining that there was a consent violation.

5.1.2 D2: Uniqueness of Consent. The uniqueness desideratum **D2** is satisfied by explicitly identifying the consent constituents expressed in the tuple $\langle G, R, \mathbb{N}, g_F, t \rangle$, and allowing only instance with that combination of elements to exist within the STS. Satisfying the uniqueness desideratum prevents the use of one consent instance by different agents, or producing an outcome that was not specified in the consent instance.

In Use Case 4, we state that a successful solicitation of consent produces a consent instance with the consent giver, consent receiver, consent norms, and stated goal and action specified. If any of these elements of the consent instance ci_{122} are changed (e.g., revoking some authorisation, as in Use Case 9) then the agreement on the norms of ci_{122} no longer holds (i.e. transitions the state of ci_{122} to σ_r). A change in a consent instance may lead, in some STSs, to producing a new instance ci'_{122}, which based on the types of interactions put in place for the modification of a consent instance, may or may not require a new negotiation.

5.1.3 D3: Assumption of Feasibility. Desideratum **D3** is expressed as the necessity that all norms are satisfied and that the stated goal is achieved for the consent to transition to the Honoured state σ_h. The feasibility of the consent norms (i.e. the stated goal follows from the satisfaction of the norms) follows from the negotiation phase of the consent, and allows the agent to be sure that the given norms will facilitate the achievement of their stated goal.

From the running example (Section 3.3), we can see that achieving the stated goal follows from the authorised actions and commitments. If the stated goal did not follow from the authorised actions and commitments, then the consent instance can never transition to the Honoured state σ_h. In the worst case, an agent who attempts to achieve their stated goal with consent norms that render this action infeasible may inadvertently violate the consent instance.

5.1.4 D4: Temporality of Consent. Lastly, we introduce the states of consent Σ and their transitions to capture that the time spent in a specific state does not have an effect on the outcome of the consent, satisfying the fourth desideratum **D4**. By instating a temporal aspect to consent, we can represent change or stability over time and expiration of the consent.

As stated in Use Case 6, an authorisation does not expire until a specific expiration condition holds (i.e. $\mathbb{S} \vDash c_{exp}$). As outlined in Algorithm 4, if there is (i) no executed action $\langle p, r \rangle$ where $p = c_{exp}$, (ii) no revocation of consent $\mathbb{S} \vDash stated\_withdrawal\_ci$, (iii) no violation of the consent norms, and (iv) no achieved stated goal, then the consent is active until one of these conditions is met.

5.2 Conclusions and Directions

We illustrated the expressiveness of our model and mechanisms by leveraging various realistic examples based on a real-world case extracted from the AI Incident Database. In doing so, we showcased different use cases for the proposed algorithms, and the enhanced ability of agents to deal with different realistic and complex situations, such as those involving insincere agents.

The main limitations of the current work are integrating our proposal into a real MAS and conducting user studies, e.g., on the generalisability and the usability of the model and mechanisms. Additional directions for future work include exploring the effects of consent mechanisms on norm conflict resolution and norm emergence, and effects of group ownership of a resource in our proposed model.

Further exploration into formalising human-like consent can also be useful in understanding accountability and its mechanisms in STS [12]. There are five rules of expressing accountability in a STS, as expressed in [5]. Two of these rules, the Alert and Treatment rules, are applicable in developing a robust system to ascribe blame when an instance of consent is violated. From this, the violation account can be used to investigate and ascribe blame to the culpable party. Rethinking these rules in the context of our model and integrating them would further strengthen our model's normative applicability.

From other areas relating to our work, we also suggest further exploration of the developmental aspect of understanding consent to extend our proposed model. Similar to the development of social norm awareness in children [10], a computational agent may also have a similar trajectory of understanding consent over time. There have been no attempts to realise this developmental process in a computational agent, which may produce a more explainable computational model of consent.

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 thank Munindar P. Singh and the anonymous reviewers for valuable feedback.

REFERENCES

[1] Stéphane Airiau, Sandip Sen, and Daniel Villatoro. 2014. Emergence of Conventions Through Social Learning: Heterogeneous Learners in Complex Networks. *Autonomous Agents and Multi-Agent Systems* (2014), 779–804.

[2] Giuseppe Albanese, Jean-Paul Calbimonte, Michael Schumacher, and Davide Calvaresi. 2020. Dynamic Consent Management for Clinical Trials via Private Blockchain Technology. *Ambient Intelligence and Humanized Computing* (2020), 4909–4926.

[3] Natasha Alechina, Brian Logan, and Mehdi Dastani. 2018. Modeling Norm Specification and Verification in Multiagent Systems. *FLAP* (2018), 457–490.

[4] Larry Alexander. 2014. The Ontology of Consent. *Analytic Philosophy* (2014), 14–137.

[5] Matteo Baldoni, Cristina Baroglio, Roberto Micalizio, and Stefano Tedeschi. 2023. Accountability in Multi-agent Organizations: From Conceptual Design to Agent Programming. *Autonomous Agents and Multi-Agent Systems* (2023), 7.

[6] Vera Bergelson. 2014. The Meaning of Consent. *Ohio State Journal of Criminal Law* (2014), 171.

[7] Guido Boella and Leendert Van Der Torre. 2007. Norm Negotiation in Multiagent Systems. *International Journal of Cooperative Information Systems* (2007), 97–122.

[8] Renée Jorgensen Bolinger. 2019. Moral Risk and Communicating Consent. *Philosophy & Public Affairs* (2019), 179–207.

[9] Dino Bollinger, Karel Kubicek, Carlos Cotrini, and David Basin. 2022. Automating Cookie Consent and {GDPR} Violation Detection. In *Proceedings of the 2022 USENIX Security Symposium*. 2893–2910.

[10] Johannes L Brandl and Frank Esken. 2017. The Problem of Understanding Social Norms and What It Would Take for Robots to Solve It. *Sociality and Normativity for Robots: Philosophical Inquiries into Human-Robot Interactions* (2017), 201–215.

[11] Cristiano Castelfranchi, Frank Dignum, Catholijn M Jonker, and Jan Treur. 2000. Deliberative Normative Agents: Principles and Architecture. In *Proceedings of the 2000 Agent Theories, Architectures, and Languages Conference*. 364–378.

[12] Amit K Chopra and Munindar P Singh. 2018. Sociotechnical Systems and Ethics in the Large. In *Proceedings of the 2018 Conference on AI, Ethics, and Society*. 48–53.

[13] AI Incidents Database. 2024. https://incidentdatabase.ai/ Last Accessed: 17/10/2024.

[14] Davide Dell'Anna, Pradeep K Murukannaiah, Bernd Dudzik, Davide Grossi, Catholijn M Jonker, Catharine Oertel, and Pınar Yolum. 2024. Toward a Quality Model for Hybrid Intelligence Teams. In *Proceedings of the 2024 International Conference on Autonomous Agents and Multiagent Systems*. 434–443.

[15] Virginia Dignum. 2017. Responsible Autonomy. In *Proceedings of the 2017 International Joint Conference on Artificial Intelligence*. 4698–4704.

[16] Virginia Dignum, Matteo Baldoni, Cristina Baroglio, Maurizio Caon, Raja Chatila, Louise Dennis, Gonzalo Génova, Galit Haim, Malte S Kließ, Maite Lopez-Sanchez, et al. 2018. Ethics by Design: Necessity or Curse?. In *Proceedings of the 2018 Conference on AI, Ethics, and Society*. 60–66.

[17] Tom Dougherty. 2015. Yes Means Yes: Consent as Communication. *Philosophy & Public Affairs* (2015), 224–253.

[18] Ricard L Fogues, Pradeep K Murukannaiah, Jose M Such, and Munindar P Singh. 2017. Sosharp: Recommending Sharing Policies in Multiuser Privacy Scenarios. *Internet Computing* (2017), 28–36.

[19] Robert E Goodin. 2023. Consent as an Act of Commitment. *European Journal of Philosophy* (2023), 194–209.

[20] Hana Habib, Megan Li, Ellie Young, and Lorrie Cranor. 2022. "Okay, whatever": An Evaluation of Cookie Consent Interfaces. In *Proceedings of the 2022 Conference on Human Factors in Computing Systems*. 1–27.

[21] Maximilian Hils, Daniel W Woods, and Rainer Böhme. 2020. Measuring the Emergence of Consent Management on the Web. In *Proceedings of the 2020 Internet Measurement Conference*. 317–332.

[22] Heidi M Hurd. 2015. The Normative Force of Consent. *Forthcoming in the Routledge Handbook on The Ethics of Consent* (2015), 15–36.

[23] Nadin Kökciyan and Pınar Yolum. 2022. Taking Situation-Based Privacy Decisions: Privacy Assistants Working with Humans.. In *Proceedings of the 2022 International Joint Conference on Artificial Intelligence*. 703–709.

[24] A Can Kurtan and Pınar Yolum. 2021. Assisting Humans in Privacy Management: An Agent-Based Approach. *Autonomous Agents and Multi-Agent Systems* (2021), 7.

[25] Jens-Erik Mai. 2016. Big Data Privacy: The Datafication of Personal Information. *The Information Society* (2016), 192–199.

[26] Neil C Manson. 2016. Permissive Consent: A Robust Reason-Changing Account. *Philosophical Studies* (2016), 3317–3334.

[27] Neil C Manson. 2022. Autonomy and Consent. *The Routledge Handbook of Autonomy* (2022), 357–367.

[28] Sean McGregor. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. In *Proceedings of the 2021 Conference on Artificial Intelligence*. 15458–15463.

[29] Francesca Mosca and Jose M Such. 2021. ELVIRA: An Explainable Agent for Value and Utility-Driven Multiuser Privacy.. In *Proceedings of the 2021 International Conference on Autonomous Agents and Multi-Agent Systems*. 916–924.

[30] Lorenzo Porcelli, Michele Mastroianni, Massimo Ficco, and Francesco Palmieri. 2024. A User-Centered Privacy Policy Management System for Automatic Consent on Cookie Banners. *Computers* (2024), 43.

[31] Marc Serramia, Manel Rodriguez-Soto, Maite Lopez-Sanchez, Juan A Rodriguez-Aguilar, Filippo Bistaffa, Paula Boddington, Michael Wooldridge, and Carlos Ansotegui. 2023. Encoding Ethics to Compute Value-Aligned Norms. *Minds and Machines* (2023), 761–790.

[32] Amika M Singh and Munindar P Singh. 2023. Norm Deviation in Multiagent Systems: A Foundation for Responsible Autonomy.. In *Proceedings of the 2023 International Joint Conference on Artificial Intelligence*. 289–297.

[33] Munindar P. Singh. 2013. Norms as a Basis for Governing Sociotechnical Systems. *Transactions on Intelligent Systems and Technology* (2013), 21:1–21:23.

[34] Munindar P Singh. 2022. Consent as a Foundation for Responsible Autonomy. In *Proceedings of the 2022 Conference on Artificial Intelligence*. 12301–12306.

[35] Roseanna Sommers. 2019. Commonsense Consent. *Yale Law Journal* (2019), 2232.

[36] Anna Cinzia Squicciarini, Dan Lin, Smitha Sundareswaran, and Joshua Wede. 2014. Privacy Policy Inference of User-Uploaded Images on Content Sharing Sites. *Transactions on Knowledge and Data Engineering* (2014), 193–206.

[37] Shukun Tokas and Olaf Owe. 2020. A Formal Framework for Consent Management. In *Proceedings of the 2020 International Federated Conference on Distributed Computing Techniques*. 169–186.

[38] Jessica Woodgate and Nirav Ajmeri. 2024. Macro Ethics Principles for Responsible AI Systems: Taxonomy and Directions. *Computing Surveys* (2024), 1–37.

[39] Mengwei Xu, Louise A Dennis, and Mustafa A Mustafa. 2024. Safeguard Privacy for Minimal Data Collection with Trustworthy Autonomous Agents. In *Proceedings of the 2024 International Conference on Autonomous Agents and Multi-Agent Systems*. 1966–1974.

[40] Vahid Yazdanpanah, Enrico H Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M Jonker, and Timothy J Norman. 2021. Responsibility Research for Trustworthy Autonomous Systems. In *Proceedings of the 2021 International Conference on Autonomous Agents and Multi-Agent Systems*. 57–62.