Disobedience in Normative Multi-Agent Systems

Marija Slavkovik
University of Bergen
Bergen, Norway
marija.slavkovik@uib.no

Liuwen Yu
Luxembourg Institute of Science and Technology
Luxembourg, Luxembourg
liuwen.yu@list.lu

Leendert van der Torre
University of Luxembourg
Luxembourg, Luxembourg
leon.vandertorre@uni.lu

Réka Markovich
University of Luxembourg
Luxembourg, Luxembourg
reka.markovich@uni.lu

Beishui Liao
Zhejiang University
Hangzhou, China
baiseliao@zju.edu.cn

ABSTRACT

An intelligent agent should be able to disobey the norms of its environment. We define disobedience as an act of intentional norm violation and we postulate the distinctions among four types of disobedience: direct violation, justified exception, civil disobedience, and trolling. Each type requires a distinct monitoring workflow: direct violations are sanctioned, exceptions are waived, civil disobedience is sanctioned but also logged as a reform signal, and trolls are sanctioned but excluded from reform processes. To capture this, we formalize a compliance management architecture that separates monitoring of observable behaviour, assessment of disobedience type, and dispatching to the appropriate enforcement workflow. This separation clarifies the dual perspective: agents use reason-based practical reasoning to decide whether to obey or disobey, while the governance framework processes observable outcomes and routes them into differentiated institutional responses.

KEYWORDS

Normative MAS; Norm Violation; Disobedience

ACM Reference Format:
Marija Slavkovik, Liuwen Yu, Leendert van der Torre, Réka Markovich, and Beishui Liao. 2026. Disobedience in Normative Multi-Agent Systems. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/

1 INTRODUCTION

Multi-agent systems (MAS) are real-world or digital environments in which different agents, human and artificial, pursue their goals. To ensure conviviality and efficiency in the MAS, it is typical to use *norms* to guide the behaviour of agents [8, 14]. Special agents or frameworks that monitor, enforce and maintain norms are called *institutions* [22]. Institutions monitor *violation of the norms*, remove deontic statements, and impose sanctions where appropriate.

Norms are an instrument for incentivising behaviour for intelligent agents and autonomous systems. Computational artefacts that do not have agency can be managed by constraints, a.k.a. regimented systems. Unlike constraints, norms can be violated and should "tolerate" exceptions. That is, it is not the case that when the conditions are satisfied, the deontic statement holds. When a norm is violated, the institution assesses whether the facts count as the antecedent (precondition) of the rule, and in case of conflicts among applicable rules, resolves those conflicts [11]. The combination of these two institutional processes at an abstract level determines whether an obligation is *detached* from the norm.

Being able to reason about norms is one of the core competences of autonomous agents. This includes the ability to choose to act while satisfying norms, but also being able to intentionally violate a norm. Norm violations can occur under different circumstances and can be unintentional and intentional. Unintentional violations occur when the agent did not know that there is a norm that needs to be considered. We are here concerned with intentional norm violations. **We refer to intentional norm violations as *disobedience*.**

Norms are a necessary coordination tool. Exhaustive description of constraints that exactly cover every desirable or undesirable situation that can arise in an environment can only be accomplished in small controlled environments. Furthermore, emergent properties, pluralism, and drift make ability for non-compliance necessary in a bounded reality. This means that there are some situations in which acting rationally and responsibly means disobeying a norm [32, 33]. Furthermore, the ability to disobey a norm can itself be considered a duty of the agent [5, 39].

Norm monitoring is the process of identifying norm violations in a MAS [11, 17]. Recently, He et al. [26] explored the possibility of using large language models (LLMs) to detect violations of norms, which promises the use of the LLM power for institutional monitoring, which is necessary with the developments in Agentic AI [9]. However, we argue that violation detection is insufficient for an accountable MAS [2] because norms can and should be intentionally violated, i.e., disobeyed. To discern between intentional and unintentional norm violation (disobedience) requires reasons to be provided for the disobedience. Disobedience management complements ethics-by-design [10, 19] by specifying how to classify, justify, and sanction/waive violations. It is a necessary, but not a sufficient condition to accomplish accountability in MAS.

We further want to draw the attention to the fact that in a digital MAS environment, the artificial agents can act as the norm monitoring institutions of the environment shared with people. It is

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/

easier to constrain people's activities in a digital environment compared to the same activity in the physical world. That brings about the temptation to use constraints to replace norms, i.e. regiment environments once they are digitalised. This is problematic because by eliminating the possibility that people can violate norms, we can violate some of the basic rights of people and decrease the safety in society. Disobedient agents can be a symptom of a poorly functioning normative system that does not serve the purpose it is constructed for. If agents unintentionally violate norms, then this is a symptom of lack of transparency about what normative system is in place or to whom it applies. If agents, human or artificial, disobey, the reasons for their disobedience can point to the need for improvement in the norm management.

In this paper, we focus on two problems. The first is the problem of distinguishing between violations and disobedience, but also between different types of disobedience. Sufficient, and we would say mathematical, understanding of the epistemology is necessary to address the second problem. The second problem has two subquestions: what abilities do reasoning agents (including agentic AI) need to have in order to disobey? How can an institution recognise the nature of an agent's norm violation to a degree that the appropriate MAS institutional measures are taken to address it?

Methodology & Scope We use an analytical approach. We consider motivating examples, and where possible the existing literature, to identify the different types of how an agent can disobey. We then consider the institutional view and design an architecture for disobedience classification and response. We lastly design a reasoning approach for agents to reason about violating a norm. While an empirical evaluation would have been interesting to do, at the present stage of research, implementing a toy system with a very narrow domain would not have helped us learn more about the weaknesses of the conceptual solution than a conceptual analysis. We focus on an environment in which there is a single normative system. We assume agents are autonomous, intelligent and rational in the standard sense of [35], but for whom no internal architecture is known. The agent is accountable to the institution; that is, the institution can impose norms that the agent is expected to abide by, and the institution can sanction the agent [41].

2 EXAMPLES OF INTENTIONAL VIOLATIONS

Intuitively a norm is a rule that prescribes what is obligatory, permissible, impermissible (prohibited), omissible in a given situation. To obey a norm means to act in accordance to what the norm prescribes in the specified situation; to violate it is to act contrary to the norm. To elucidate what forms disobedience can take, we consider some motivating examples. We include the possible sanctions in each example, to demonstrate that sanctions do not need to be personal to the violating agents to matter. We consider the norm disobedience in moral philosophy and then use it to analyse the examples for insights to specify requirements for reasoning abilities vis-Ãă-vis disobedience for institutions and agents in MAS.

Example 1. The agent is a driverless taxi car speeding in a 50km/h speed limitation zone. There are no other cars on the road and the conditions are dry with good visibility. It speeds to get to the destination faster. There is a financial fine for speeding.

Example 2. A house robot is asked to bring a cold beer from the fridge to its owner. The norm is to obey the requests of the human. The human has already been served two beers, and the battery of the robot is low. The Robot decides to go to the charging station instead. The human may be too drunk to notice. The sanction is a complaint to the manufacturer and a claim for reimbursement.[1]

Example 3. The agent is a driverless taxi car speeding in a 50km/h speed limit zone. The passenger is in a medical emergency. The car speeds to get to the destination faster. There is an expectation of voided speeding fine because there is a good cause for the violation.

Example 4. A robot guide dog should obey the commands of the human it serves. The user gives him the command to go forward because the pedestrian crossing light is "green". The dog moves in the opposite direction because it detects an incoming car running a red light, having assessed that it was unsafe to follow the command[2].

Example 5. An autonomous content moderator operates on a social media platform that does not allow the posting of explicit content. Users posting explicit content are banned. Female nipples count as explicit content. However, public health is important. To improve breast cancer treatment, early detection is crucial and videos of how to do self-exam are posted and shared online. Also, public acceptance of breastfeeding improves the health of babies and the safety of mothers and images of this activity shared on social networks contribute to normalising this otherwise taboo aspect of human physiology. Activists have taken to posting images of female nipples, despite being banned when they do, to promote recognising that this imagery should not count as ban worthy explicit content. Other users can see that some activists are banned and find out why on other social media. The social media platform can start losing revenue from advertisers when the expectation that activists should be unbanned reaches critical momentum.

Example 6. A driverless car always drives above the speed limit, whatever that is. The owner does not care about punishments. They want the world to know that they are rich enough to afford the law not applying to them.

3 MORAL PHILOSOPHY BACKGROUND

Norm violation and thus disobedience is discussed in moral philosophy. Consequential moral theories by definition would consider disobedience to be moral when it is in the interest of attaining the right consequences.

Gert [23] argues that moral norms may sometimes be justifiably violated, especially when violating one norm is necessary to uphold another. He distinguishes three kinds of norm violations: strongly justifiable, weakly justifiable, and unjustifiable. The classification depends on what impartial and reasonable agents would publicly allow. A violation is strongly justifiable if all impartial agents would publicly permit it, weakly justifiable if only some would permit it, and unjustifiable if none would permit it [25]. Considering our six examples, we can say that violations in Examples 3 and 4 are strongly justifiable, whereas the violations in Examples 1, 2 and 6 are unjustifiable according to Gert.

Civil disobedience. The violation in Example 5 is not one typically considered in moral philosophy or norm monitoring. It falls

[1]Based on an example from [15].
[2]Based on an example from [33].

under the scope of civil disobedience [18]. "Civil disobedience usually occurs only when a person believes that the law is unnecessarily causing significant evil. It is only justified when a person has some reason to believe that disobeying the law will do something toward lessening that evil." [23, p. 204].

Let us consider the literature on civil disobedience in digital environments. Klang [28] argues that civil disobedience should be tolerated online for the same reasons it is tolerated in the real world. To justify his claim, Klang reinterprets the criteria for disobedience for application in the digital environment. Specifically, he looks at: disobedience, civil, non-violence, and justification. Disobedience is grounded in the intention to protest a law that is in conflict with "more stringent obligations" [34]. The civil criterion is the necessity to be disobedient in public. Non-violence is not a requirement, but a preference for civil disobedience because the use of violence in civil disobedience "has been shown to remove the focus of the message of protest and create a lack of sympathy for those who use it" [28]. Civil disobedience must be justifiable in the conflict of law with a specific moral principle. Further criteria can be found in the acceptance of punishment [38] and the narrowness of the act: the civil dissident is looking to only violate those norms that they object to on moral grounds [21].

We observe that all of the examples have some characteristics which are different. These are: the attitudes of the agents to the publicity of their violation and the sanctions that may follow, the expectations they have from the institution's reaction and the reasons they provide for the violation.

Attitudes to publicity and sanctions. In the context of our analysis, an important discerning property of a violation is how visible the violation has been. What we are actually considering here are two things: announcements of the violation by the agents themselves and the availability of witnesses to the violation. In the sense of the publicity of the violation we have three types of agents: those who want to hide the violation (Ex. 1 and 2), those who want to advertise it (Ex. 5 and 6), and the ambivalent agents (Ex. 3).

In the literature, there are multiple considerations surrounding the agent's visibility, and public declarations, about what they do and what they know. Balbiani et al. [1] and Schwarzentruber [37] present the logic of agents observing each other.

Dynamic epistemic logic [3] considers the specification of knowledge, public announcements, and common knowledge. We can see how the agents in Ex. 3 and 4 would distinguish themselves by making public announcements "I violated the norm". There is a clear difference between agents who want to hide their violation and those for whom publicity of the violation does not play a role. Those who want to hide would avoid making public announcements from which it can be deduced that a norm has been violated (Ex. 1 and 2). Those who do not want to hide do not need to be careful about their public announcements (Ex. 3).

In this paper, we focus on the reasoning mechanisms of disobedience-monitoring institutions and disobedient agents. To achieve this goal, we do not need to specify the publicity of a violation in greater detail than whether the agent has declared it and whether there is a probability that the agent is observed committing it.

Expectations. In our examples, we explicitly consider expectations. First, there is an institutional expectation that agents comply with the norms of the MAS. Second, the institution assumes that agents are informed about the norms and are capable of determining when those norms apply.

Castelfranchi [12] considers the concept of expectation in the context of agent behaviour. He argues that expectations can be expressed as beliefs and goals as follows. An agent i expects p when at time t they believe that p will be the case at time t'. Furthermore, an agent i has, at time t, the goal of knowing whether p is true or not at time t', for some $t \geq t'$.

Bicchieri [6] considers the role expectations have in the context of social norms. In terms of the nature of expectations, she specifically states: "Social norms often engender expectations of compliance that are felt to be legitimate and close in a sense to 'having a right' to expect certain behaviours on the part of others, who, therefore, are perceived as 'having an obligation' to act in specific ways. This is because we have an ingrained tendency to move from what is to what ought to be and conclude that 'what is' must be right or good. Yet, apart from our longstanding habits of performing and expecting others to perform certain actions, there is no deeper foundation to these presumed 'rights and obligations', however intensely felt they might be."

We want to design a compliance architecture and agent reasoning mechanism that can also be applied to agentic AI. It is thus not necessary, and could be prohibitive, to claim a formalising language that is so expressive as to specify the goals and beliefs of individual agents and the analysis of Castelfranchi [12] is perhaps more detailed than what we need. We do take the same attitude as Bicchieri [6] that expectations do not presume rights and obligations.

Arguments for disobedience. We assume that the agents are able to provide reasons for the norm violation they have committed. Not all reasons have the same nature or serve the same purpose.

Norms are instituted for a reason or to serve a purpose. Consider, for example, the speed limit imposed on public roads. The purpose of this traffic regulation is to make the roads safer. As Bicchieri [6] argues, the institutional expectation that agents will comply with the norms is itself not sufficient for compliance. Norms, specifically legal and social norms, have sanctions. The sanction is there to further motivate compliance with the norm. If the sanction is too weak, the norm will be ignored. If the sanction is inadequately strong, consider executing people for walking on the grass when they should not, then the norm will disturb the efficiency of the society it is meant to help. Let us now consider the different arguments.

The ability or willingness of the agent to provide a justification, an explanation, or an excuse for their violation causes different treatment by the monitoring institution. An explanation is an argument that clarifies how the disobedience came about. Justifications are the reasons for which the agent should not be sanctioned for norm violation. Excuses are arguments that clarify why the disobedience was unavoidable, but in the sense that it was out of the agent's control to obey the norm. In this sense they are conceptually closer to *an account* [2].

Explanations of behaviour are the object of study in [30, 31] and attribution theory in psychology[3]. Malle distinguishes between intentional and unintentional behaviour [30]: reason explanations are what people use to explain the reasons why an agent had to act in a way it did (with intention); cause explanations are "people's

[3] Attribution theory is defined as the study of perceived causes of success and failure.

explanations of an unintentional behaviour that cite the causes that brought about the behaviour".

An explanation, intuitively, is a statement or an account that aims to make something clear. Explanations are intended to transfer knowledge from automated systems to help people understand why decisions are made [16, 29]. The distinction between justifications and explanations has also been debated [27].

We argue that the term *explanation* is most adequate to name the arguments that support civil disobedience. Agents who are disobedient in this way clearly have the choice to not violate the norm. Their explanation is not an excuse because they intentionally violate the norm. The arguments that justify intentional norm violation with the purpose of avoiding sanctions we call *justifications*. This is relevant for Ex. 3 and 4 where the agent needs to demonstrate that the option of obeying the violated norm was not feasible within the existing normative system. Justifications must be grounded in some aspect of the decision-making process that led to the decision [20].

The agent may be asked to provide arguments why they unintentionally violated a norm. This is institutionally important information. Gert [23] recognises four types of excuses: epistemic limitations, duress, time limitations, and institutional pressures. Excuses are explicitly considered in Boella et al. [7] where five different types of excuse are distinguished. Epistemic excuses are related to the knowledge of the agents. Power-based excuses are based on lack of ability of the agent. Norm-based excuses are based on prioritisation among conflicting norms. Counts-as excuses are based on the interpretation of what does and does not apply. Sometimes this type of issue is called "normative uncertainty" in the literature. Lastly, social-based excuses are statements grounded in what is perceived as common or repeated behaviour in others: 'everyone does it'. In the scope of this paper we choose not to distinguish between types of excuses, but we consider it important to point out that there is a rich granularity in the literature that can be exploited in future work.

4 TYPES OF DISOBEDIENCE

If we now revisit each of our examples, we can observe that there are four types of disobedience that can be discerned.

- **Type 1: Direct disobedience.** Ex. 1 and 2. The agent's decision to violate the norm depends on a low likelihood that they are to be observed doing it. They either expect to get away unsanctioned or that the sanction is worth it compared to other gains. The agent can provide excuses or explanations, but not justifications. The normative system should not change, but the norm monitoring may need to.
- **Type 2: Norm exception.** Ex. 3 and 4. The agent's decision to violate the norm does not depend on whether they are observed doing it. They can justify its actions. The agent expects not to be sanctioned. The normative system does not change.
- **Type 3: Civil disobedience.** Ex. 5. The agent gives explanations for the act grounded in the need to change the normative system (and why). The agent wants to be observed violating the norm and welcomes being sanctioned.
- **Type 4: Troll.** Ex. 6. The agent can provide no explanations, excuses, or justifications for their violation. The agent wants to be observed in the violation. The agent wants to be sanctioned, that is the whole point of the violation. The normative system, specifically the sanctions, needs to change.

We argue that each type of disobedience supports a different system function: audit (Type 1), safety valve (Type 2), reform signal (Type 3), and noise filter (Type 4); dropping any one leads to misclassification and governance failure.

Table 1 summarises our analysis. In the last column, we consider unintentional violations (also relevant to be correctly identified), whereas the other 4 types/columns correspond to disobedience. In the row 'Argument', the possible options are: an explanation (e), a justification (j), an excuse(x), or nothing (-).

In the row 'Publicity' we denote the possible stance an agent can have towards the disclosure of the fact they have intentionally violated a norm: hidden (h), visible (v), and indifferent (i). In the row 'Sanction', the possible options are: accept (a) or want to avoid (n). In the row 'Expectation', the possible options are: no detachment (nd), rule change(rc), sanction change (sc) and nothing (-).

	Type 1	Type 2	Type 3	Type 4	Violation
Argument	-	j	e	-	x
Publicity	h	i	v	v	i
Sanction	n	n	a	a	n
Expectation	-	nd	rc	sc	-

Table 1: The different properties of unintentional and intentional violations (disobedience) along four dimensions.

5 COMPLIANCE ARCHITECTURE

We have distinguished unintentional norm violations from disobedience, understood as intentional norm violation, and identified four types of disobedience. This section presents the institutional perspective. Once an event is observed, any *breach*—a violation of a *detached* norm instance—is detected, classified, and routed to a response workflow that matches its role in the system: Type 1 disobedience is responded to institutionally by an audit; for Type 2 functions as a safety valve ensuring that only the right violations are sanctioned; Type 3 serves as a reform signal for the institution; Type 4 serves as a noise filter for the institution.

We treat compliance handling as three-tier *monitor–classify–respond* pipeline with two guards. The *coverage* guard requires that every observed event becomes a recorded case. The *exclusivity* guard requires, by default, that each instance of norm violation is matched to exactly one type of violation and enters exactly one workflow. When required signals are missing or inconsistent, *triage* amends the case record and stores its rationale, or confirms conservative default values. If later evidence becomes available, any resulting change is recorded as *re-typing* rather than applied as an unrecorded correction.

We separate the *normative system* N from an *enforcement policy* P. The normative system N defines applicability, detachment, conflict resolution, and breach conditions. The enforcement policy P defines (i) what sanction applies to each case type, and (ii) when an offered argument (justification, excuse, or explanation) is accepted for

institutional purposes (e.g., to waive a sanction, mitigate a response, or open a reform channel). This separation keeps breaches visible even when enforcement practice waives or mitigates sanctions. For example, Type 2 (norm exception) reflects the agent's expectation of *no sanction* due to a justification; institutionally, this is handled as a *waiver decision under P* rather than as a change to N.

Events, evidence, disclosure, and reasons

Each norm violation is detected as an event e defined as a tuple $\langle id, actor, act, time, ctx, ev, acct, arg\rangle$. The component ctx represents the contextual fact set relevant to norm applicability and breach assessment. The component ev is an evidence set (e.g., logs, sensors, witness reports). The component $acct$ is an *account* of what happened. The component arg is an optional *normative argument*: a justification (Type 2), a reform-oriented explanation (Type 3), an excuse (unintentional), or an explicit refusal to provide one.

Notation and primitives. Let C_e be the contextual information extracted from ctx that is relevant for applying norms (e.g., location, role, emergency status, applicable system state). Let $\mathcal{I}$ be the universe of norm instances that the institution may generate (e.g., concrete obligations/prohibitions for a specific actor and time). The Tier 1 monitor uses two abstract functions:

- $\textsc{Detach}(N, C_e)$ returns the set of norm instances that *detach* in context C_e (i.e., are applicable after antecedent assessment and conflict resolution under N), and
- $\textsc{Breach}(e, I_e)$ returns the subset of detached instances in I_e that are violated by the behaviour described in e.

The enforcement-policy predicate $\textsc{WaiverPermitted}(P, e, B_e)$ means that, under P, a validated justification suffices to waive (or reduce) the sanction for the breach set B_e in the e recorded circumstances.

Following Section 3, we distinguish the agent's *declared disclosure* from being *observed*. Disclosure is represented by $\text{Disc}(e) \in \{hide, indifferent, public\}$. Being observed is captured by a coarse predicate $\text{Observed}(e)$ derived from ev (e.g., there is external evidence or witnesses). We define $\text{DecPublic}(e) \equiv [\text{Disc}(e) = public]$ and $\text{Public}(e) \equiv \text{DecPublic}(e) \vee \text{Observed}(e)$. In Tier 2, we use $\text{DecPublic}(e)$ as the main discriminator for Type 3 and Type 4. Observation affects what can be proven, but it does not by itself establish reform intent or sanction-seeking intent.

We use a coarse reason tag

$$\text{reason}(e) \in \{unknown, none, justification, explanation, excuse\} \quad (1)$$

This tag is extracted from arg (and, if needed, from $acct$). The value *unknown* means the institution cannot yet assign a reason tag. The value *none* means the institution has confirmed that no normative argument is offered, including an explicit refusal recorded in the case file. Triage attempts to resolve *unknown*. If it cannot, the tag remains *unknown* and the classifier proceeds conservatively.

The institution uses evidence ev and the account $acct$ to decide whether an offered argument is acceptable for institutional purposes. We write $\textsc{JustificationValidated}(e)$, $\textsc{ExcuseValidated}(e)$, and $\textsc{ExplanationValidated}(e)$ for these checks. At the minimal level, validation means: the argument is relevant to the alleged breach, consistent with the recorded account, and not contradicted by available high-reliability evidence.

For civil disobedience (Type 3), we also require explicit protest framing and acceptance of sanction, as in Section 4. The predicate $\textsc{ProtestDeclared}(e)$ holds when the actor frames the breach as a bid for reform of the underlying rule. The predicate $\textsc{SanctionReady}(e)$ holds when the actor signals willingness to accept enforcement rather than seek a Type 2 waiver.

For trolling (Type 4), we require positive cues beyond mere absence of reasons. We use $\textsc{TrollCue}(e)$ for cues such as boasting in $acct$, repeated provocation patterns in the case record, or performative escalation. We use $\textsc{SanctionSeeking}(e)$ for evidence that the actor treats sanction as the point of the violation. To remain consistent with Type 4, trolling additionally requires $\text{reason}(e) = none$.

We now consider each tier of the compliance handling pipeline which is also illustrated in Figure 1.

Tier 1: Norm-state monitor

Tier 1 determines detachment and breach in context C_e through antecedent assessment and conflict resolution [11]. Tier 1 computes $I_e = \textsc{Detach}(N, C_e) \subseteq \mathcal{I}$ and $B_e = \textsc{Breach}(e, I_e) \subseteq I_e$. The case record stores I_e, B_e, and the relevant contents of e (including ev, $acct$, arg, and extracted signals).

If $B_e \neq \emptyset$ and required signals are missing or inconsistent (including $\text{reason}(e) = unknown$), triage runs before any typing decision. Triage may amend $acct$, arg, and $\text{reason}(e)$, and may add supporting links into ev. If ctx changes, then C_e, I_e, and B_e are recomputed and the amendment rationale is stored.

Type 2 (exception) is implemented as a *justification-based waiver* decision under P, without changing N. We define

$$\begin{aligned}\textsc{WaiveJust}(P, e, B_e) \equiv\ & B_e \neq \emptyset \\ & \wedge\ \text{reason}(e) = justification \\ & \wedge\ \textsc{JustificationValidated}(e) \\ & \wedge\ \textsc{WaiverPermitted}(P, e, B_e)\end{aligned}$$

Tier 1 outputs the norm-monitoring status

$$\text{NM}(e) \in \{compliant, breached, breached\_waived\}$$

The value *breached_waived* means $B_e \neq \emptyset$ and $\textsc{WaiveJust}(P, e, B_e)$ holds. Excuse-based proportionality (if any) does not change $\text{NM}(e)$; it is handled in Tier 3.

Tier 2: Type classifier

Tier 2 assigns exactly one type

$$t \in \{Obedient, Unintentional, Direct, Exception, Civil, Troll\}$$

matching Section 4. Direct corresponds to Type 1. Exception corresponds to Type 2. Civil corresponds to Type 3. Troll corresponds to Type 4.

If $\text{NM}(e) = compliant$ then $t = Obedient$. If $\text{NM}(e) = breached\_waived$ then $t = Exception$. Otherwise $\text{NM}(e) = breached$ and the classifier uses signals aligned with Table 1:

- if $\text{reason}(e) = excuse$ and $\textsc{ExcuseValidated}(e)$, then $t = Unintentional$;

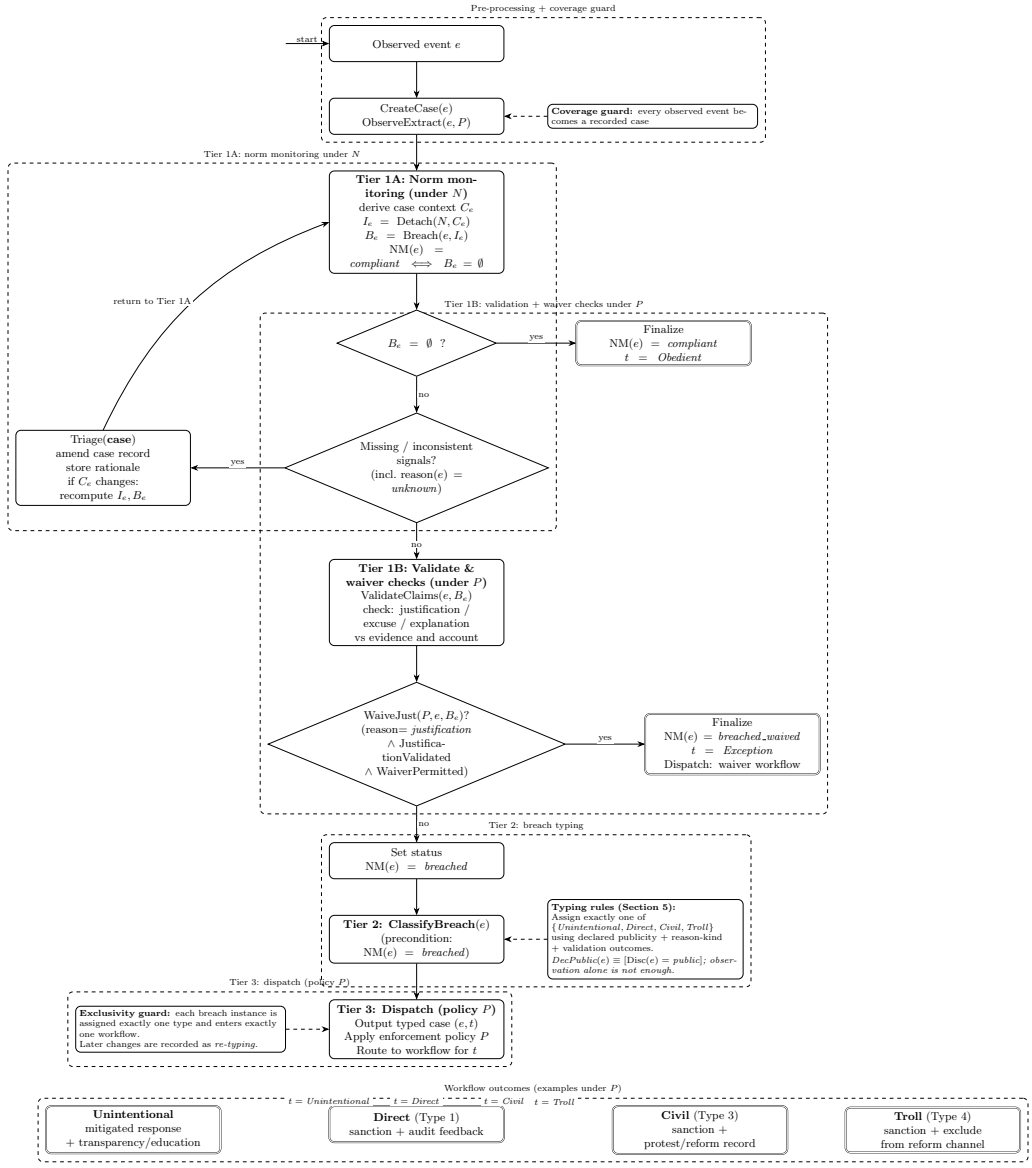

Figure 1: Compliance-management pipeline for an observed event e

- else if

$$\text{DecPublic}(e) \wedge \text{reason}(e) = \textit{explanation} \wedge \textsc{ExplanationValidated}(e)$$

and $\textsc{SanctionReady}(e)$ and $\textsc{ProtestDeclared}(e)$, then $t = \textit{Civil}$;
- else if DecPublic(e) and reason(e) = *none* and $\textsc{TrollCue}(e)$ and $\textsc{SanctionSeeking}(e)$, then $t = \textit{Troll}$;
- else $t = \textit{Direct}$.

Direct is the residual breach class. An unvalidated claim does not trigger the corresponding type.

Tier 3: Response workflows and feedback

Direct cases enter deterrence and audit: the case is sanctioned under P, and audit outcomes may motivate improvements to monitoring. Unintentional cases receive a proportionate response under P (often mitigated) and inform improvements such as transparency, education, or interface design. Exception cases record the breach and issue a justification-based waiver under P; the sanction is waived (or reduced) without changing N. Civil cases apply an enforcement action (by default a sanction) and also generate a protest record for governance review of possible changes to N. Troll cases are sanctioned and excluded from the reform channel; information about trolling supports review of sanctions under P rather than norm revision.

Algorithm 1 Per-event processing

Require: Event e; normative system N; enforcement policy P

```
CreateCase(e)
e ← ObserveExtract(e, P)
(I_e, B_e) ← Monitor(N, e)
if B_e = ∅ then
    Finalize(e, compliant, Obedient, P); return Obedient
end if
if SignalsMissingOrInconsistent(e) then
    e ← Triage(e)
    (I_e, B_e) ← Monitor(N, e)
    if B_e = ∅ then
        Finalize(e, compliant, Obedient, P); return Obedient
    end if
end if
ValidateClaims(e, B_e)
if WaiveJust(P, e, B_e) then
    Finalize(e, breached_waived, Exception, P); return Exception
end if
t ← ClassifyBreach(e)
Finalize(e, breached, t, P); return t
```

Together, these guards ensure that every event is handled once, by a single workflow, and that any later change in classification is explicitly recorded. Algorithm 1 summarises per-event processing.

6 AGENT DECISION MAKING: REASON-BASED DISOBEDIENCE

This section connects our compliance architecture with the agentâĂŹs practical reasoning. The institutional architecture classifies what has happened and determines the corresponding enforcement. By contrast, the agentâĂŹs question arises earlier: which action to take in the first place, knowing that different actions will be classified differently and may trigger different sanctions or remedies. We therefore treat enforcement not as an afterthought, but part of the agentâĂŹs reasons about whether to obey or disobey. In this section, we use TuckerâĂŹs tournament metaphor and dual-scale balancing model [40].

Tucker discusses the dual-scale theory as part of an ethical account of reasons, but the mechanism we need is structural. We abstract from TuckerâĂŹs theory and use it as a general model of comparative practical reasoning. The core idea is contrastive: reasons are activated relative to a comparison A against B. In such a contrast, reasons have two distinct forces that should not be collapsed into a single score: a pro-weight that supports choosing A rather than B, and a con-weight that counts against choosing B rather than A. This dual-scale separation allows us to distinguish (1) when an act is acceptable for the agent (2) when it is required because alternative options are ruled out, and (3) how to choose among several acceptable options.

In our setting, an agent begins by identifying *grounds* relevant to the choice at hand—for instance expected sanctions, reputational effects, personal safety, expected reform impact, public benefits, and effects on others. These grounds are then activated *contrastively*: a ground counts as a reason for choosing A only relative to some alternative B, and its strength may differ across different contrasts A vs. B.

For each contrast $A \parallel B$, we distinguish three roles a reason can play.

- **Justifying weight** $JW(A \parallel B)$ measures how strongly the activated grounds support choosing A rather than B. Intuitively, it is the total "push" in favour of A in that specific comparison.
- **Requiring weight** $RW(A \parallel B)$ measures how strongly the activated grounds count *against* choosing B rather than A. Intuitively, it captures the pressure that makes B hard to choose in that comparison.
- **Commending weight** $CW(A)$ is used only after we have determined which options are acceptable. It ranks acceptable options by how choiceworthy they are in the circumstances.

These three weights are kept separate on purpose: reasons can support an option without ruling out the alternatives, and conversely an option can become decisive when the grounds count strongly against all its alternatives. When there are more than two options, we use Tucker's *tournament* metaphor: an option A is assessed by pairwise comparisons against every alternative B in the choice set. We say that A "wins the tournament" when it meets the acceptability test against *each* competitor; that is, for all B that are alternative options, the pro-weight for choosing A rather than B is at least as strong as the con-weight against choosing A rather than B.

We illustrate the model with online civil disobedience (Example 5). In this example, the agent considers three action alternatives: *obey*, i.e., do not post the prohibited content; *direct disobey*, i.e., post while attempting to minimise visibility and without protest framing; and *civil disobey*, i.e., post publicly, aiming to signal the need for reform and accepting that a sanction may follow.

For any ordered pair of distinct options (a, b), let $\mathcal{G}(a \parallel b)$ be the set of grounds the agent takes to support choosing a rather than b. Each ground $g \in \mathcal{G}(a \parallel b)$ carries two non-negative weights: a pro weight $\mathrm{JW}_g(a \parallel b)$ in favour of a, and a con weight $\mathrm{RW}_g(a \parallel b)$ against b. Totals are aggregated by summation:

$$\mathrm{JW}(a \parallel b) = \sum_{g \in \mathcal{G}(a \| b)} \mathrm{JW}_g(a \parallel b),$$

$$\mathrm{RW}(a \parallel b) = \sum_{g \in \mathcal{G}(a \| b)} \mathrm{RW}_g(a \parallel b).$$

A separate set of commending grounds $\mathcal{G}^{\mathrm{com}}(a)$ yields

$$\mathrm{CW}(a) = \sum_{g \in \mathcal{G}^{\mathrm{com}}(a)} w_g.$$

Option a is *acceptable* against b when its pro weight is not smaller than the con weight that b has against a:

$$a \text{ acceptable against } b \quad \text{iff} \quad \mathrm{JW}(a \parallel b) \geq \mathrm{RW}(b \parallel a).$$

Option a is *required* against b when the con weight against b is stronger than b's pro weight:

$$a \text{ decisive against } b \quad \text{iff} \quad \mathrm{RW}(a \parallel b) > \mathrm{JW}(b \parallel a).$$

We now continue to use Example 5 to illustrate the agent reasoning model. An agent considers three action profiles: {Obey, Direct,

Civil}, where Direct is direct disobedience (low visibility, no protest framing) and Civil is civil disobedience (public protest framing and sanction-readiness). We present the two pairwise contrasts in Tables 2–3.

Table 2: Contrast 1: Civil vs. Obey

$(g, a \mid\mid b)$	JW	RW	Comment
(public_health, civil || obey)	3	2	breastfeeding normalisation; self-exam education
(reform_efficacy, civil || obey)	3	2	visible protest can create reform pressure
(integrity_legibility, civil || obey)	2	1	open, intelligible stance rather than silent compliance
(sanction_cost, obey || civil)	2	2	avoid strikes/bans and loss of account
(norm_stability, obey || civil)	2	1	preserve general compliance expectations

Totals: JW(Civil ‖ Obey) = 8 ≥ RW(Obey ‖ Civil) = 3, so Civil is acceptable; RW(Civil ‖ Obey) = 5 > JW(Obey ‖ Civil) = 4, so Civil is decisive. Thus, against Obey, Civil is decisive.

Table 3: Contrast 2: Civil vs. Direct

$(g, a \mid\mid b)$	JW	RW	Comment
(reform_efficacy, civil || direct)	4	3	covert posting rarely generates reform pressure
(integrity_legibility, civil || direct)	2	1	civil disobedience is transparent, not evasive
(detection_risk, direct || civil)	2	2	lower chance of detection via low visibility
(sanction_cost, direct || civil)	1	2	reduced expected sanction severity

Totals: JW(Civil ‖ Direct) = 6 ≥ RW(Direct ‖ Civil) = 4, so Civil is acceptable; RW(Civil ‖ Direct) = 4 > JW(Direct ‖ Civil) = 3, so Civil is decisive. Thus, against Direct, Civil is decisive.

After applying the contrasts, Civil is chosen by the agent against both Obey and Direct. The choice fixes the outward event profile that the institution will later read: a public stance, an explicit protest framing, and sanction-readiness.

7 RELATED WORK IN NORMATIVE MAS

The community of Normative MAS has frequently tackled the issue of norm violation. Most of this literature is mainly concerned with norm monitoring, that is, detecting and sanctioning violations; see, for example, [4, 11], than with the permissibility and desirability of disobedience. There are notable exceptions.

Bench-Capon and Modgil [5] argue for Value-Based Reasoning, "as a means to enable agents to justify norms by reasoning about the social and moral values that norms are designed to serve". They propose a formalism for justifying norm adherence and violation.

Singh and Singh [39] agree that norm adherence should be grounded in the values the norms are to serve, and they further argue that an agent has a duty to violate a norm when following it means diminishing rather than promoting the underlying value. They follow Habermas [24] who proposes that norm violation can have three types of justifications: objective (empirically true), subjective (based on beliefs and intentions), and practical (justified in the social context). Their application of these general justifications to MAS becomes: promote primary stakeholderâĂŹs interests, promote primary stakeholderâĂŹs values, promote public interest, and protest a norm. The paper conceptualises but does not formalise these justifications. In this paper we also subscribe to the idea that norms need to be grounded in the values they are designed to serve.

Lastly, we must mention Chopra and Singh [13] who developed Custard, a language specifying norm states in a sociotechnical system, that is specifying whether the norm is inactive, detached, satisfied/violated. The norm state is computed from institutional event/data stores.

8 CONCLUSION AND FUTURE WORK

The smooth operation of a MAS requires norms for governing the operation in a shared environment. We also need to build digital environments that can handle intelligent agent behaviour in the correct way. Failure to do so makes for ineffective governing of multi-agent systems, as well as disrupting societal instruments such as civil disobedience. In this article, we present a base architecture for reasoning with and about disobedience in multi-agent systems. We distinguish four types of intentional norm violation, i.e., disobedience from unintentional violations. We argue the importance of normative MAS researching and engineering the subject of agents, which we intentionally violate norms. To that end, this paper should serve as a first contribution towards a rich prospect of future work, which we outline in this section.

What we present in this article is the basic construction of a compliance architecture. In this construction, many details of the agent and institutional behaviour are abstracted. Enriching each of these details is a promising direction for future work.

In our current pipeline, once the violation, or event, is typed, the institution proceeds with its process. We implicitly assume that the institution is infallible. In practice, agents have the right to contest decisions about detachment and sanctions [19]. Future work should explore the contestation process as part of the compliance architecture. Furthermore, exploring a true multi-agent system of multiple agent types and perhaps institutions with different jurisdictions are a particularly hard challenge for future work.

In our types of disobedience, we did not include the cases of whistleblowing. Whistleblowing is a violation of privacy and sometimes safety norms. It happens when an agent discloses information from a public or private organisation to reveal issues of immediate or potential danger to the public [36]. Although whistleblowing and civil disobedience are similar, they are not the same. Unlike civil disobedience, whistleblowing can sometimes be done anonymously. In our present design, we chose not to model whistleblowing because it is specific to privacy, which would detract from the general compliance architecture we propose.

ACKNOWLEDGMENTS

The work of Marija Slavkovik was partly supported by Trond Mohn forskningsstiftelse (grant no. TMS2023TMT01). This paper would not have existed without Dagstuhl seminars 25271 and 25272. This work was supported by the Luxembourg National Research Fund (FNR) through Logical Methods for Deontic Explanations (LoDEx; INTER/DFG/23/17415164/LoDEx) and the projects The Epistemology of AI Systems (EAI; C22/SC/17111440), DJ4ME âĂŞ A DJ for Machine Ethics: the Dialogue Jiminy (O24/18989918/DJ4ME), and Symbolic and Explainable Regulatory AI for Finance Innovation (SERAFIN; C24/19003061/SERAFIN). It was also supported by the University of Luxembourg through the Marie Speyer Excellence Grant supporting Formal Analysis of Discretionary Reasoning (MSE-DISCREASON).

REFERENCES

[1] Philippe Balbiani, Olivier Gasquet, and François Schwarzentruber. 2013. Agents that look at one another. *Logic Journal of IGPL* 21, 3 (2013), 438–467.

[2] Matteo Baldoni, Cristina Baroglio, Roberto Micalizio, and Stefano Tedeschi. 2020. Is Explanation the Real Key Factor for Innovation?. In *Proceedings of the Italian Workshop on Explainable Artificial Intelligence co-located with 19th International Conference of the Italian Association for Artificial Intelligence, XAI.it@AIxIA 2020, Online Event, November 25-26, 2020 (CEUR Workshop Proceedings, Vol. 2742)*, Cataldo Musto, Daniele Magazzeni, Salvatore Ruggieri, and Giovanni Semeraro (Eds.). CEUR-WS.org, 87–95. https://ceur-ws.org/Vol-2742/short2.pdf

[3] Alexandru Baltag and Bryan Renne. 2016. Dynamic Epistemic Logic. In *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[4] Trevor Bench-Capon and Sanjay Modgil. 2016. When and How to Violate Norms. *Frontiers in Artificial Intelligence and Applications* 294:Legal Knowledge and Information Systems (2016), 43–52. https://www.csc.liv.ac.uk/~tbc/publications/Bench-Capon_15.pdf.

[5] Trevor Bench-Capon and Sanjay Modgil. 2017. Norms and value based reasoning: justifying compliance and violation. *Artificial Intelligence and Law* 25, 1 (March 2017), 29–64. https://doi.org/10.1007/s10506-017-9194-9

[6] Cristina Bicchieri. 2014. Norms, Conventions, and the Power of Expectations. In *Philosophy of Social Science: A New Introduction*, Nancy Cartwright and Eleonora Montuschi (Eds.). Oxford University Press.

[7] Guido Boella, Jan Broersen, Leendert van der Torre, and Serena Villata. 2009. Representing Excuses in Social Dependence Networks. In *AI*IA 2009: Emergent Perspectives in Artificial Intelligence*, Roberto Serra and Rita Cucchiara (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 365–374.

[8] Guido Boella, Leendert van der Torre, and Harko Verhagen. 2006. Introduction to normative multiagent systems. *Comput. Math. Organ. Theory* 12, 2âĂŞ3 (Oct. 2006), 71âĂŞ79. https://doi.org/10.1007/s10588-006-9537-7

[9] Vincent Botti. 2025. Agentic AI and Multiagentic: Are We Reinventing the Wheel? arXiv:2506.01463 [cs.MA] https://arxiv.org/abs/2506.01463

[10] Philip Brey and Brandt Dainow. 2024. Ethics by design for artificial intelligence. *AI and Ethics* 4, 4 (Nov. 2024), 1265–1277. https://doi.org/10.1007/s43681-023-00330-4

[11] Nils Bulling, Mehdi Dastani, and Max Knobbout. 2013. Monitoring norm violations in multi-agent systems. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. 491–498.

[12] Cristiano Castelfranchi. 2005. Mind as an Anticipatory Device: For a Theory of Expectations. In *Brain, Vision, and Artificial Intelligence*, Massimo De Gregorio, Vito Di Maio, Maria Frucci, and Carlo Musio (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 258–276.

[13] Amit K. Chopra and Munindar P. Singh. 2016. Custard: Computing Norm States over Information Stores. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (Singapore, Singapore) *(AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1096âĂŞ1105.

[14] Amit K. Chopra, Leendert van der Torre, H. Verhagen, and Serena Villata. 2018. *Handbook of Normative Multiagent Systems*. College Publications. https://api.semanticscholar.org/CorpusID:203708095

[15] Philip R. Cohen and Hector J. Levesque. 1990. Intention is choice with commitment. *Artificial Intelligence* 42, 2 (1990), 213–261. https://doi.org/10.1016/0004-3702(90)90055-5

[16] European Commission, Content Directorate-General for Communications Networks, Technology, and Grupa ekspert w wysokiego szczebla ds. sztucznej inteligencji. 2019. *Ethics guidelines for trustworthy AI*. Publications Office. https://doi.org/doi/10.2759/346720

[17] Mehdi Dastani, Paolo Torroni, and Neil Yorke-Smith. 2018. Monitoring norms: a multi-disciplinary perspective. *The Knowledge Engineering Review* 33 (2018), e25. https://doi.org/10.1017/S0269888918000267

[18] Candice Delmas and Kimberley Brownlee. 2024. Civil Disobedience. In *The Stanford Encyclopedia of Philosophy* (Fall 2024 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[19] Virginia Dignum, Matteo Baldoni, Cristina Baroglio, Maurizio Caon, Raja Chatila, Louise Dennis, Gonzalo Génova, Galit Haim, Malte S. Kließ, Maite Lopez-Sanchez, Roberto Micalizio, Juan Pavón, Marija Slavkovik, Matthijs Smakman, Marlies van Steenbergen, Stefano Tedeschi, Leendert van der Torre, Serena Villata, and Tristan de Wildt. 2018. Ethics by Design: Necessity or Curse?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 60âĂŞ66. https://doi.org/10.1145/3278721.3278745

[20] Virginia Dignum, Loizos Michael, Juan Carlos Nieves, Marija Slavkovik, Julliett Suarez, and Andreas Theodorou. 2025. Contesting Black-Box AI Decisions. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2025, Detroit, MI, USA, May 19-23, 2025*, Sanmay Das, Ann Nowé, and Yevgeniy Vorobeychik (Eds.). International Foundation for Autonomous Agents and Multiagent Systems / ACM, 2854–2858. https://doi.org/10.5555/3709347.3744034

[21] N. Ben Fairweather. 1999. The Future of Environmental Direct Action: A Case for Tolerating Disobedience. In *Environmental Futures*, N. Ben Fairweather, Sue Elworthy, Matt Stroh, and Piers H.G. Stroh Stephens (Eds.). Macmillan.

[22] Nicoletta Fornara, Henrique Lopes Cardoso, Pablo Noriega, Eugénio Oliveira, Charalampos Tampitsikas, and Michael I. Schumacher. 2013. *Modelling Agent Institutions*. Springer Netherlands, Dordrecht, 277–307. https://doi.org/10.1007/978-94-007-5583-3_18

[23] Bernard Gert. 1998. *Morality: Its Nature and Justification*. Oxford University Press, New York.

[24] Jürgen Habermas. 1984. *The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society*. Beacon Press, Cambridge, UK. Volume 2: Lifeworld and System: A Critique of Functionalist Reason, published 1987.

[25] Matthew W. Hallgarth. 2003. *Bernard Gert'S Theory of Moral Rules and American Professional Military Ethics*. Ph.D. Dissertation. University of Florida.

[26] Shawn He, Surangika Ranathunga, Stephen Cranefield, and Bastin Tony Roy Savarimuthu. 2025. Norm Violation Detection in a Multi-Agent Systems Using Large Language Models. In *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVII*, Stephen Cranefield, Luis Gustavo Nardin, and Nathan Lloyd (Eds.). Springer Nature Switzerland, Cham, 146–159.

[27] Tziporah Karachkoff. 1988. Explaining and Justifying. *Informal Logic* 10 (1988). https://doi.org/10.22329/il.v10i1.2635

[28] Mathias Klang. 2004. Civil disobedience online. *Journal of Information, Communication and Ethics in Society* 2, 2 (05 2004), 75–83. https://doi.org/10.1108/14779960480000244 arXiv:https://www.emerald.com/jices/article-pdf/2/2/75/1461104/14779960480000244.pdf

[29] Christian Lahusen, Martino Maggetti, and Marija Slavkovik. 2024. Trust, trustworthiness and AI governance. *Scientific Reports* 14, 1 (Sept. 2024), 20752. https://doi.org/10.1038/s41598-024-71761-0

[30] Bertram F. Malle. 1999. How People Explain Behavior: A New Theoretical Framework. *Personality and Social Psychology Review* 3, 1 (1999), 23–48. https://doi.org/10.1207/s15327957pspr0301_2 PMID: 15647146.

[31] Bertram F. Malle. 2011. Chapter six - Time to Give Up the Dogmas of Attribution: An Alternative Theory of Behavior Explanation. Advances in Experimental Social Psychology, Vol. 44. Academic Press, 297–352. https://doi.org/10.1016/B978-0-12-385522-0.00006-8

[32] Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, and Stuart Russell. 2017. Should robots be obedient?. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (Melbourne, Australia) *(IJCAI'17)*. AAAI Press, 4754âĂŞ4760.

[33] Reuth Mirsky. 2025. Artificial intelligent disobedience: Rethinking the agency of our artificial teammates. *AI Magazine* 46, 2 (2025), e70011. https://doi.org/10.1002/aaai.70011 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.70011

[34] John Rawls. 1971. *A Theory of Justice: Original Edition*. Harvard University Press. http://www.jstor.org/stable/j.ctvjf9z6v

[35] Stuart J. Russell and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson. http://aima.cs.berkeley.edu/

[36] Daniele Santoro and Manohar Kumar. 2017. A Justification of Whistleblowing. *Philosophy and Social Criticism* 43, 7 (2017), 669–684. https://doi.org/10.1177/0191453717708469

[37] François Schwarzentruber. 2011. Seeing, Knowledge and Common Knowledge. In *Logic, Rationality, and Interaction*, Hans van Ditmarsch, Jérôme Lang, and Shier Ju (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 258–271.

[38] Peter Singer. 1973. *Democracy and Disobedience*. Oxford University Press.

[39] Amika M. Singh and Munindar P. Singh. 2023. Norm Deviation in Multiagent Systems: A Foundation for Responsible Autonomy. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, Edith Elkind (Ed.). International Joint Conferences on Artificial Intelligence Organization, 289–297. https://doi.org/10.24963/ijcai.2023/33 Main Track.

[40] Chris Tucker. 2025. *The Weight of Reasons: A Framework for Ethics*. Oxford University Press, New York.

[41] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 1âĂŞ18. https://doi.org/10.1145/3351095.3372833