Norm Deviation in Multiagent Systems: A Foundation for Responsible Autonomy

Amika M. Singh[1] and **Munindar P. Singh**[2]
[1]Harvard Law School
[2]North Carolina State University
asingh@jd23.law.harvard.edu, mpsingh@ncsu.edu

Abstract

The power of norms in both human societies and sociotechnical systems arises from the facts that (1) societal norms, including laws and policies, characterize acceptable behavior in high-level terms and (2) they are not hard controls, as they can be deviated from. Thus, the design of responsibly autonomous agents faces an essential tension: these agents must both (1) respect applicable norms and (2) deviate from those norms when blindly following them may lead to diminished outcomes.

We propose a conceptual foundation for norm deviation. As a guiding framework, we adopt Habermas's theory of communicative action comprising objective, subjective, and practical validity claims regarding the suitability of deviation. Our analysis thus goes beyond previous studies of norm deviation and yields reasoning guidelines uniting norms and values by which to develop responsible agents.

1 Introduction

AI agents are taking on an increasing presence in our personal, civic, and work lives, making this question vital: *How can we ensure that AI agents act responsibly and prosocially?*

The multiagent systems (MAS) community [Dastani *et al.*, 2018] follows the legal [Hohfeld, 1919] and deontic logic [Von Wright, 1963] traditions in understanding a (social) *norm* to include laws and other prescriptions or proscriptions on social behavior. We conceive of a sociotechnical system (STS) [Chopra and Singh, 2018; Singh, 2013], viewed in computational terms as (1) comprising stakeholders (people, organizations, and the society at large) and AI agents and resources and (2) controlled socially through norms.

Responsible agents normally respect but *may violate* the applicable norms. A substantial body of research on normative multiagent systems concerns norm violation, e.g., via social controls such as sanctioning (penalties or rewards) to guide agents and discourage them from violating norms [Nardin *et al.*, 2016]. Some works identify norm conflicts [Dos Santos *et al.*, 2017] and resolve them at design or run time. Dell'Anna *et al.* [2020] identify the need for the norms to change at runtime. Kafalı *et al.* [2020] map stakeholder requirements to a combination of norms and low-level mechanisms to specify STSs that avoid or mitigate norm violations.

Balancing norm compliance and deviation is essential to trustworthiness [Yazdanpanah *et al.*, 2021b]. Specifically, we would like our agents to be (1) autonomous, so they apply their knowledge and intelligence to help stakeholders, and (2) prosocial and responsible with respect to the expectations of their stakeholders, including the society at large. Norm deviation is a practical concern. Agents may be designed to skirt the law, if not break it outright. For example, Tesla announced an "assertive" driving mode in which a car may make a rolling stop [BBC, 2022] even though traffic laws generally require a complete stop [NHTSA, 2022]. We posit that similar norm-breaking opportunities arise in other AI applications, such as algorithmic trading and loan assessment, though they may not be quite so transparently advertised.

Accordingly, this paper's contribution lies in addressing this key question: *What are legitimate criteria by which an agent may decide to comply with or deviate from a norm?* Importantly, legitimacy cannot rely on external criteria such as sanctions but the other way around: the legitimacy of a decision should determine the sanctions at play.

We adopt Habermas's [1984] theory of communicative action comprising objective, subjective, and practical validity claims as an organizing framework. We rely not merely on argumentation but on jurisprudence and case law to provide concrete, empirical ways in which norms are construed in practice. First, jurisprudential thought goes beyond the presumption that the law ought to be obeyed by default to investigate under what conditions an agent ought to obey or disobey the law. Second, case law provides a grounding to how claims of deviation from norms are viewed in practice and thus complements the theoretical basis of jurisprudence. Our case law examples are drawn from the US (federal and state), Canadian, and English courts of law.

Placing this empirical evidence in the Habermasian framework treats the law in action as a real-life normative system that combines the letter and practice of law. This exercise strengthens normative MAS by providing a new basis for agent behavior and STS design. This study provides a conceptual foundation for *responsible autonomy* [Singh, 2022] based on the Habermasian validity claims and grounded in legal thinking and case law. In addition, it provides a pathway to realizing responsible autonomy through argumentation.

2 Positioning in the Literature

This paper addresses more basic concerns than previous work on norm compliance or deviation in AI.

2.1 Brief Literature Review

The literature on norm deviation focuses on two topics: (1) identifying and resolving norm conflicts [Dos Santos *et al.*, 2017], so that any resulting violation is justified by a conflict and (2) sanctions for the deviant agents, e.g., [Savarimuthu *et al.*, 2008]. Recent works address formal reasoning about norms to elucidate ethical concerns through constructs such as values and responsibility [Serramia *et al.*, 2018; Yazdanpanah *et al.*, 2021a]. Falcone *et al.* [2013] discuss the relationship between norms and trust: how trust is needed for norms (because norms can be violated) and how norm compliance promotes trust. Nardin *et al.* [2016] discuss how trust and reputation can serve as positive or negative sanctions.

Murukannaiah *et al.* [2020] relate the values of stakeholders to the norms under which a sociotechnical system operates. They provide no indication of which norm violations are justified and how to give an accounting for the violations. Woodgate and Ajmeri [2022] address how values and norms interplay. They entertain agents violating norms but do not elucidate potential reasons for violation. Montes and Sierra [2021] describe how values may determine norms and provide a computational framework. However, their values are fixed and norms are undirected, unlike [Singh, 2013].

Cranefield *et al.* [2017] introduce values into a belief-desire-intention (BDI) programming model for agents. The values provide a basis for selecting plans (their emphasis) as well as norms (which they mention). They do not address norm deviation, which would entail selecting norms and then failing to comply with them. Tubella *et al.*'s [2019] model for governance is based on mapping or "interpreting" moral values in terms of concrete norms, which can vary with the setting. They do not support norm deviation, which may be necessary when the norms do not adequately capture what the values suggest in a particular context.

2.2 Novelty

Most of the above approaches are evaluated via examples crafted by the authors. But empirical validation is crucial since, to understand the suitability of agent behaviors or the analyses produced by formal tools, we must understand what people would do. Further, understanding such intuitions can lead to more perspicuous explanations.

Recent AI approaches, e.g., Liscio *et al.* [2022], carry out empirical validation by consulting samples of a target population. Surveys can involve many respondents efficiently but suffer from being based on abstract queries, which do not give an in-depth view of a respondent's intuitions. A *vignette study* [Schafheitle *et al.*, 2020] is a variant in which respondents are presented with a short description of a story and asked to comment. Vignettes add depth beyond traditional questions but remain abstract and surely lack the life-and-death stakes of real life. Specifically, the framing of the problem in a vignette influences the response. An alternative is a qualitative methodology, specifically, where a skilled researcher interviews a respondent by asking open-ended questions [Dubljević *et al.*, 2022].

In contrast, we adopt case law as a source of empirical knowledge about how laws are interpreted and applied. In general, litigation concerns edge cases because disputes with a clear outcome are resolved informally. Thus, case law provides a means to apply Flanagan's [1954] *critical incident technique*: analyzing extreme cases to identify criteria for *typical* performance. Importantly, case law includes a combination of jury trials, where the juries are made of members of the public, and bench or appellate trials, where no juries are present. In both settings, lawyers contest their claims, sometimes backed by domain experts. Thus, case law represents a combination of expert and nonexpert opinions based on an intensively contested debate lasting days or weeks, which makes case law more reliable as a source of insight than short context-free surveys of the public. Singh and Singh [2023] recently applied this methodology to model the trustworthiness of AI.

3 Background

To introduce the relevant concepts, consider an autonomous vehicle (AV) as an exemplar of an intelligent agent who faces situations that call for deviation from norms. It helps to distinguish an agent's *primary* from other stakeholders: the primary stakeholders are those the agent is directly a representative or assistant of. For example, an AV's primary stakeholders are its passengers, and its other stakeholders include other vehicles, pedestrians, and transportation authorities.

3.1 Values

In moral psychology, the *values* [Rokeach, 1973; Schwartz, 2012] of an individual lie at the heart of ethical decision making. This view has been adopted in AI ethics as well: To act ethically is to act in accordance with one's values [Liscio *et al.*, 2022]. Rokeach [1973] identifies 18 terminal (ends) and 18 instrumental (means) values. Examples of ends are freedom and equality; examples of means are independence and ambition. Schwartz [2012] identifies ten values, such as universalism, self-direction, achievement, and security.

A value is a core part of the cognitive and affective model of an individual that determines how they are motivated to preserve or achieve certain states. That is, values are affective but refer to an individual's goals and help prioritize actions or decisions an individual is considering. The distinctions between people arise not so much about their values in the abstract but in their value preferences in different contexts.

3.2 Habermas's Validity Criteria

Habermas's [1984] theory addresses the validity of interactions in the public sphere. Bohman and Rehg [2017] introduce Habermas's thought. Of relevance to us is Habermas's dialectical framing of validity, whereby an agent may seek to justify its reasoning and actions. Importantly, Habermas goes beyond empirical truth in the conventional sense but includes moral and subjective predicates such as rightness, goodness, and sincerity as acceptable justifications for actions. This characterization is crucial since suitable justifications of norm deviation rely on such cognitive, moral, and social constructs.

Habermas [1984] associates three validity claims (i.e., distinct standards of correctness) with each communication in the public sphere. We adapt these claims to any agent behavior that is subject to norms as a basis for identifying standards for when an agent may suitably deviate from a norm. The three kinds of claims are *objective* (empirically true), *subjective* (based on beliefs and intentions), and *practical* (justified in the social context). For example, if an AV explains that it drove on the wrong side of the road to avoid a collision, its explanation would be evaluated in three ways. Objectively, is there evidence that a collision was imminent without the (purportedly) evasive action? Subjectively, did the AV believe there was going to be a collision, even if its beliefs may have been false? Practically, given the conditions (e.g., of heavy traffic), was it wise to risk multiple head-on collisions to avert a minor collision?

4 Justified Norm Deviation

Ideally, an agent would comply with all applicable norms. However, compliance might not be possible or ethical in certain circumstances. Only in cases where there is a good reason may an agent deviate from any applicable norm. But what might a good or good enough reason for violating a norm be?

Based on the literature and our intuitions, we group potential reasons for deviating from norms into these categories:

Promote primary stakeholder's interests An AV may drive on the wrong side of the road (crossing the solid lines) to avoid an accident that might injure its occupant.

Promote primary stakeholder's values An AV may drive with its headlights off at night to promote its occupant's privacy (as to when they leave or enter their neighborhood with snooping neighbors).

Promote public interest A civilian AV may block a vehicle trying to merge into traffic from a side street to prevent it from obstructing an ambulance, although a civilian AV is not authorized to block any vehicle.

Protest a norm An AV may block the path of a foreign government dignitary to protest human-rights violations there as well as domestic laws that favor senior politicians. Here, the goal is to disrupt existing norms.

Although the above examples of AV behavior have not yet arisen in the legal system, legal doctrine and practice have addressed relevant cases, albeit in other domains. That is, the above categories of reasons for deviating from norms may be induced from the legal literature we discuss below.

A contribution of this paper is to ground discussions of responsible autonomy in what has been tackled in the law.

5 Norm Deviation Conceptually

The law is a useful source of insight into the opportunities and limits to autonomy. Indeed, the law is often thought of as a "line drawing" institution [Weisbach, 1999]. Here, we are concerned with characterizing the circumstances where AI agents ought to deviate from an applicable norm—in other words, to understand where the lines are.

We now discuss the legal literature on an agent's duty to obey the law, the extent of an agent's discretion, and an agent's duty to disobey the law under suitable circumstances.

5.1 Morality

Authority and morality are central concepts in jurisprudence. Legal philosophers agree that the law carries some authority, but centuries of jurisprudence have contemplated whether the force of the law does, or ought to, outweigh an individual's morality, requiring them to obey the law even when it conflicts with their morality. To some, e.g., Wolff [1998], moral autonomy must be forgone in favor of an obligation to respect authority, and moral considerations of the individual have no bearing on whether or not they ought to obey the law. Others, e.g., Finnis [2003], take the view that laws ought to be complied with unless there is an overriding moral reason not to. That is, Finnis is supportive of deviating from norms but only in cases of a conflict with values.

5.2 Responsibility to Community

Dworkin [1988] writes (paraphrased in MAS terminology) that people ought to uphold norms that benefit their associative communities, provided that four conditions hold. First, these norms must be special, existing only within the boundaries of that community. Second, the norms must be personal, directed from individual to individual and not to the community as a whole. Third, the community must be equal, where each individual within it is thought to contribute to the community's purpose. Fourth, the norms must originate out of concern for members of the community.

Not all deviations from applicable norms motivated by one's community would satisfy Dworkin's four conditions. Consider the case of the Masterpiece Cakeshop [SCOTUS, 2018]. The shop's owner, Phillips, declined to provide a cake for a same-sex wedding because of his religious opposition to same-sex marriages. The denial of service violated the Colorado Anti-Discrimination Act. Many believe that the owner of a private business ought to have the right to deny service to whomever they so choose. Thus the question becomes whether the owner, in finding a law he did not like, could have refused to comply with it.

Let us apply Dworkin's four tenets to this situation. It is possible that Phillips was acting out of a sense of obligation to his religious community. However, the crucial component of this case is that the in-group, being the religious community Phillips is a member of, was discriminatory to the outgroup, the gay community. Dworkin writes that "if the consequences for strangers to the group are grave, as they will be if the discriminating group is large or powerful within a larger community, then this will be unjust." Here, the harm to the engaged couple may be considered major as an assault on their dignity as humans. What makes the bakery case different from Dworkin's communities is the context of the outgroup being marginalized.

5.3 Duty to Violate a Norm

Thoreau [1848] argues that if a government requires you to be an agent of injustice toward others, there is a duty to break the law, saying a citizen should use their "whole influence."

Young [2001] writes of an activist acting under non-ideal conditions, such as when there is an unjust law. The activist is a force for justice when they advocate for the public interest. The public interest is not well defined, but we can take

it as the aggregate good of all people subject to the deliberative democracy, i.e., the good of society as a whole. To work in the public interest, one must not be limited to one's own group. Young's next criterion is to be on the side of the powerless, who may not have access to the political processes of the deliberative democracy, and, as such, are unable to express their views without activism. Most importantly, activists must explain the reasons for disobeying the law, and they must do so with the intent to change the law. According to Young, activists are constrained by public reasons and must use peaceful means of protest. Young's activists are not violent or destructive, and their goal is not simply to disturb the peace but to end the injustice that motivated their protest.

5.4 Proximate Causation

The extent of autonomy afforded to an individual is constrained on both sides by proximate causation. If following a law would be proximally linked to a bad outcome, then the individual can be expected to violate the law. Conversely, if there are no proximal bad outcomes, this argument for autonomy in violating a norm dissipates. The moral duty to refuse an immediately unjust law is limited by proximate causation.

This principle of proximate causation was established in the United States by Palsgraf v. Long Island Railroad. Though this is a tort case, and the relevance of this particular outcome may vary by state, Palsgraf has nonetheless shaped the doctrine of proximate cause. The facts of the case are as follows: The plaintiff, a Mrs. Palsgraf, was standing on the platform, waiting for her train. A man carrying a package rushed ahead of her to board a moving train. Two train employees helped him on. In the process, the man dropped the package, and it exploded: it contained fireworks. The explosion caused a scale on the platform to fall on and injure Mrs. Palsgraf. She then sued the railroad company, claiming that the negligence of the workers assisting the man with the package caused her harm. Initially, she won the case, and she won on the first appeal. But on the second appeal, in the New York Court of Appeals, Chief Judge Benjamin Cardozo overturned the previous decisions. Judge Cardozo's opinion in Palsgraf established the idea of proximate cause, as he found that the railroad company could not be found negligent because the harm to Mrs. Palsgraf could not have been foreseeable to the railroad workers who were assisting a customer.

If a harmful outcome from following a norm cannot be predicted, or if the link between the norm and that outcome is too tenuous, there cannot be an obligation to deviate from it.

5.5 Social and Historical Context

To determine what constitutes justifiable norm deviation, we must look to social and historical contexts to first understand what constitutes justice. It is crucial to ensure that AI processes do not further the subordination of a "less equal" group. Consider a case in Canada [Canada, 1989]. Andrews was a resident of Canada who met all the requirements for the British Columbia bar except that of Canadian citizenship. The Supreme Court of Canada held that the law preventing his entry into the bar violated the Canadian Charter of Rights and Freedoms, which requires equal protection and benefit of the law for all persons. The Court held that equality means that there may be no distinction or differentiation based on discrimination. That is, there could be no distinction made based on the personal characteristics of an individual or group that withholds or limits access by that individual or group to advantages available to other members of society. As the current state of law and technology includes racial and other discriminatory biases, we must be careful not to perpetuate such inequalities in autonomous agents.

6 Norm Deviation in Practice

Armed with the foregoing discussion of the legal foundations of autonomy and Habermas's validity claims, we now review how the law comes down on the various justifications for deviating from norms with respect to specific laws. Whereas Habermas's validity claims are notionally to be conjoined (in that a defensible public action should be valid in all three respects), in legal cases, the objective aspects are hardly ever called out unless violated, and the focus largely remains on the subjective (the intents of the people involved) and the practical (the context in which they were acting) aspects. Of the three, there is usually one primary claim, which we use to organize the discussion below of the justifications for norm deviation from Section 4 along with the relevant case law drawn from the US (federal and state), Canadian, and UK courts of law. Table 1 summarizes our findings.

6.1 Promote Primary Stakeholder's Interests

Since the case law is focused on people, of relevance are cases focused on one's personal interest and cases focused on one's client's interest (as for a medical practitioner and their patient). It is convenient to discuss such cases separately.

Promote Personal Interest

Objective validity. The law has carved out some exceptions to generally applicable laws where an individual's personal interests are so strong as to outweigh the interests behind the law. In property law, a trespass is when someone enters an owner's land without permission. Whereas ordinarily, the entrant would face some punishment, there is an exception of necessity. In Ploof v. Putnam [Vermont, 1910], Ploof was in a boat with his family when a storm began. Ploof tied his boat to Putnam's dock for safety, and Putnam (or, rather, his employee) untied it, causing the boat to be destroyed and the family to be injured. The Vermont Supreme Court held that the necessity of needing to tie the boat to a dock in order to remain safe meant that the entry was justified and was not trespass. That is, Ploof's exercise of his autonomy in deviating from the norm was justifiable, and Putnam was out of line in having Ploof's boat unmoored. We classify this case as supported by objective validity because the facts of the storm were empirically established.

Subjective validity. With some limits, one can be allowed to deviate from a norm based merely on a belief that one's personal interests are at risk by following a law. A common example is speeding on the way to a hospital in case of a medical emergency, which violation can be overlooked if there is no harm to others. However, if harm to others is caused, the deviation from a norm is likely deemed unacceptable. In

Reason	Objective Validity	Subjective Validity	Practical Validity
Personal interest	A deviation is acceptable for risk to one's life • Ploof [Vermont, 1910]	No cases found, likely because such cases are either never reported or fall into another category	Acceptable deviation is limited by the potential of harm to a child • Re A [England & Wales, 2001]
Client interest	Therapist retains the privilege to keep therapy session notes secret • Jaffee [SCOTUS, 1996]	No cases were found since a professional serving a client has a duty to work in the client's interest, not just believe that they are	A psychiatrist does not have the legitimate autonomy to preserve privilege with respect to their client under imminent danger • Tarasoff [California, 1976]
Personal values	Not possible	Autonomy to serve customers does not extend to discrimination • Masterpiece [SCOTUS, 2018]	Conscientious objectors: the law carves out an exception for discretion in avoiding certain kinds of military service
Public interest	Prohibition on smoking on airplanes based on evidence of health and safety risks	Dress-code laws in France, which end up being discriminatory and not in the public interest	Discretion is limited when it would further harm a historically discriminated group • Andrews [Canada, 1989]
Protest a norm	Only determined after the fact when a protested norm is overturned (or not)	Potentially determined after the fact or from contemporaneous statements	Unethical voting restrictions would not change without civil disobedience • Clay (Ali) [SCOTUS, 1971]

Table 1: Illustrations of legal precedents regarding validity claims justifying norm deviation.

other words, we find no case law for this entry because the norm violation is either waived by an officer of the government (such as a police officer) or falls into another category.

Practical validity. Re A [England & Wales, 2001] concerned conjoined twins, one of whom was much weaker than the other and relied on her sister's heart to survive. There was a surgery that could save one of the twins, thereby leaving the other guaranteed to die. If the surgery was not performed, neither child would survive. The parents, devout Catholics, did not want the surgery to be performed. The Court of Appeal of England and Wales held that it was lawful to act without parental approval because it was necessary to operate in order for one of the children to survive. We classify this case as practical validity because it is based on trading off the harm and benefit to the two children and the decision to limit the parents' ability to decide when the life of a child is at stake.

Promote Client Interest

Objective validity. Consider protecting a client's confidential information. In general, as the US Supreme Court noted in Jaffee v. Redmond [SCOTUS, 1996], notes taken during an individual's therapy session are protected, and the psychiatrist or therapist cannot be legally compelled to give that information to the Court. In this case, the notes may have implicated Redmond, a police officer who had shot and killed someone. A jury awarded damages (without having seen the notes but with an indication by the judge that the notes should have been revealed). That decision was thrown out because it was based on the suggestion that the notes should have been provided and that the therapist was wrong to have withheld them. This case established the importance of a psychiatrist or therapist's duty of confidentiality to their client. We classify this case as illustrating objective validity because the facts of the therapist-client relationship, based on which the therapist violated the apparent legal requirement, were not in dispute.

Subjective validity. We did not find clear cases concerning autonomy exercised by someone based on what they imagine would serve their client's interest. This lack of subjective cases for deviation from norms suggests an expectation that professionals are able to accurately identify their client's interest, so if the deviation does not promote the client's interest, there are doubly at risk (of both violating a norm and failing to preserve their client's interest).

Practical validity. This case is a counterpoint to Jaffee, discussed above, in which there was no imminent danger to anyone. The privilege established in Jaffee has its limits. In Tarasoff v. Regents of University of California [California, 1976], the plaintiffs were the parents of the deceased Tatiana Tarasoff, who had been killed by her classmate, Poddar. Two months before her murder, Poddar confided his intent to kill Tarasoff to his psychiatrist. The Supreme Court of California held that the plaintiffs could state a claim against the psychiatrist because though the psychiatrist warned some campus police officers, they did not warn Tarasoff or her parents, or do anything to confine Poddar to prevent him from murdering her. The psychiatrist was negligent in failing to warn. We classify this case as practical validity because it hinges on the context: the client being in imminent danger of causing harm.

6.2 Promote Personal Values

Objective validity. There is no clear argument for the objective validity of values, which are psychological constructs. Hence, we do not expect to find any entry here.

Subjective validity. The Masterpiece Cakeshop case [SCOTUS, 2018] discussed above demonstrates the limits of where personal values can be applied. Masterpiece's owner, Phillips, was motivated by his values, which included a lack of support for gay marriage. However, his attempt at justifying the violation on subjective grounds was trumped by its

lack of practical validity because discrimination against people based on sexual orientation is illegal. We classify this case as subjective because the motivation was not based on empirical facts but on an opinion about sin.

Practical validity. We place conscientious objectors, individuals who refuse to participate in military service, in this category. The law entertains their autonomy with respect to the kind of military service they would undertake, even though they might not take up a function that involves fighting and injuring or killing others.

6.3 Promote Public Interest

Objective validity. Laws that limit individual freedom (such as prohibiting smoking onboard commercial aircraft or in airports) are in the public interest. That is, a government agency has the authority to go beyond what is normally allowed constitutionally to be able to limit the freedoms of individuals. We classify these laws as objective because whereas the few may suffer (e.g., due to their freedoms being limited), the public (as a whole) benefits.

Subjective validity. There are no good cases of subjective validity of deviation from norms in the public interest, but dress-code laws such as in France may be argued as being in the public interest. However, these are firmly subjective since there is no evidence to suggest the public benefits from such laws. Indeed, they may be practically invalid since they promote discrimination against religious minorities.

Practical validity. The Andrews case [Canada, 1989] illustrates that deviation from norms by an organization is limited when the deviation might harm a discriminated group. We classify Andrews under practical validity because its legitimacy hinges upon the context of discrimination of an outgroup. Protesting a norm (the next reason) and public interest overlap. For example, the Civil Rights movement in the US in the 1950s and 1960s [Young, 2001] used protests to target the desegregation of public schools and the overruling of Jim Crow voter suppression laws. The public interest is hurt by injustice, and the practices and policies resulting from protests against discriminatory and other norms advanced the public interest by getting rid of some discriminatory norms.

6.4 Protest a Norm

Objective validity. These cases are hard to find ahead of time but can be determined after the fact. That is, if the norms actually change, we can see that the protesters were vindicated. However, the validity of a protest can be overruled by a violation of public interest and values, e.g., if the protest itself causes harm to life or property. In 1977, the US Supreme Court ruled in favor of a Nazi group's free speech right to violate the norms of the town of Skokie, Illinois by marching downtown [SCOTUS, 1977]. (The Nazis withdrew their plans to march after winning a long legal battle.)

Subjective validity. May be determined based on communications of intent.

Practical validity. Thoreau's idea of civil disobedience is exemplified by the boxer Muhammad Ali. In 1966, Ali refused to be drafted into the United States military because he did not want to assist in the marginalization of the Vietnamese. Ali was arrested for refusing the draft in 1967. After the Fifth Circuit Court of Appeals upheld his conviction, it was overturned by the Supreme Court because of the failure of the draft board to provide a reason for denying Ali a conscientious objector exemption [SCOTUS, 1971]. Ali passes Young's activism test. He was acting in the interest of the Vietnamese public, who were suffering at the hands of both the Soviet Union and the United States as these superpowers fought to advance their geopolitical agendas. Ali was not being discriminatory in his actions, and he was speaking for the powerless, in this case, the Vietnamese, who were unable to lobby the US government to end the war. Ali's actions affected society at large because of his celebrity status and his being an exceptional speaker: He drew attention to the injustice he saw and started a national discussion on war and the role of the US in world politics.

7 Applying the Framework

A benefit of using Habermas's conception of validity claims as the central organizing principle is that the giving and taking of accounts in a dialectical sense comes naturally with it.

Argumentation offers a flexible way to express stakeholder needs and values and capture reasoning about norm deviation [Walton *et al.*, 2008]. An *argument scheme* captures patterns of reasoning that an agent may follow under specified circumstances. An argument scheme comprises a major premise, a minor premise, and a conclusion. This reasoning from the premises to the conclusion is not deductive but defeasible: informal logic does not guarantee true conclusions from true premises but it is the kind of reasoning people apply in practice. Shams *et al.* [2020] describe and apply some argument schemes for reasoning about goals, plans, and norms.

Importantly, argument schemes are centered on *critical questions* [Walton *et al.*, 2008]. Since the reasoning in the scheme is not deductive, it is the critical questions (and answers to them) that carry the essential knowledge that makes an argument viable. That is, the critical questions bring forth the situations under which the argument fails—e.g., where the relevant context does not hold or the premises fail or a counterargument would prevail over this argument.

Argument schemes can thus capture criteria for legitimate norm deviation. The following scheme addresses client interest. Similar schemes address the remaining reasons, thus producing one scheme for each row of Table 1.

Major premise:	X ought to protect X's client's interest, Y
Minor premise:	One way for X to protect Y (X's client's interest) is to deviate from norm N
Conclusion:	X should deviate from norm N

The associated critical questions include one for each of the validity claims for a putative norm deviation: Is it objectively valid? Is it subjectively valid? Is it practically valid?

7.1 Illustration

Let us consider a mortgage-loan scenario loosely based on Tubella *et al.*'s [2019] setting. A banker (a bank employee) decides on a loan application. Relevant norms include granting or denying an application based on (1) the ratio of the loan

amount to the appraised value of the property; (2) the ratio of the monthly payment to the borrower's income; (3) uniformity in the treatment of borrowers based on their income and zip code. Norm deviations could go in either direction: granting or denying loans in violation of an applicable norm.

The personal interest of an employee (e.g., a year-end bonus) as a reason for deviation is not practically valid. A banker could serve their client's interest by relaxing or tightening the thresholds to maximize potential profit and lower risk. Doing so could be valid in all three respects.

Suppose a property is foreclosed by the bank (taken over when the borrower fails to make loan payments). The borrower may resist eviction to protest not the general norm of eviction from property one does not own, but the norms of predatory lending, resulting in usurious terms for their loan.

The reason for deviation matters. Imagine a volunteer helping low-literacy (and poor) borrowers complete loan applications. The volunteer is subject to a norm of reporting their client's income and expenses truthfully. They may overstate the client's income to—in their view (i.e., subjectively)—help them buy an expensive house. But locking a poor client into a large loan payment may not objectively be in the client's interest. Suppose instead our volunteer overstates their client's income to compensate for the bank's bias against minorities. Thus, the same deviation (overstating the income) may be valid in all three respects in the second case but not the first.

Norm deviation can trigger dilemmas between the reasons. Suppose the banker (validly) exercises their personal values to relax loan thresholds for members of their ethnicity. But doing so would go against the public interest because of unequal treatment based on ethnicity. That is, the validity of a deviation relies on chaining arguments [Freeman, 2021, p. 120]. We need additional critical questions addressing the tradeoffs between various reasons in the case at hand.

7.2 Uses of the Framework

Since our framework maps to argument schemes for practical reasoning, we observe that it supports not just deliberation about whether to comply or deviate but also the negotiation of norms [Aydoğan *et al.*, 2021]. An agent can put forth an argument, leading another to counter with facts that negate its premises or bring out critical questions that undermine its applicability. Justified deviation applies naturally in competitive settings, e.g., where one agent may justify why it violates the terms of a contract or an autonomous vehicle may justify why it would not concede to another according to prevalent norms. Another possible use is in an engineering methodology for responsible agents, akin to those for goal-based [Bresciani *et al.*, 2004] and value-based [Liscio *et al.*, 2022] design.

In general, the critical questions would depend on specific domains to accommodate the relevant aspects of the context, e.g., for privacy [Kökciyan and Yolum, 2022], personal assistance [Kola *et al.*, 2020], or health care [Shah *et al.*, 2021]. Whereas modeling context is nontrivial in general, in specific domains, it can be tractable. An effective approach would be, like Kökciyan and Yolum [2022], to ask for human intervention for difficult cases. Another approach would be to adapt Telang *et al.*'s [2021] formal model centered on *practical rules* with an oracle for answering the critical questions. Such oracles may be built using machine learning from data obtained through experience with user interactions. For example, knowing that a norm deviation serves one's client's interests but does not violate the public interest may be an adequate justification for deviating from that norm.

8 Discussion

This work differs from previous studies in its (1) conceptual approach to justified norm deviation by incorporating Habermas's notion of validity into a legal perspective and (2) validation based on using case law as an empirical basis instead of introspection or speculation.

Our five reasons for deviation can be readily ordered in terms of the ease of realizing them in AI agents. Interests are the easiest conceptually and may be reflected in utilities, for example. In cases where the distinction between personal and client interest is essential, we might associate personal interest with the agent's creator or operator or even itself (positing it as sentient) and its client's interest with a user. Values are more challenging than interests because the underlying constructs are not readily made computational. Public interest involves developing social intelligence in our agents. Current work, as described in Section 2, addresses elements of what is needed for the above reasons. Protest appears beyond today's multiagent systems approaches because it requires a deeper past and future historical analysis of a social system.

Social and historical context is an essential factor in all aspects of responsibility. Suppose an agent were to allocate loans or suggest sentences for juvenile defendants, which would require not blindly following norms but purposefully deviating from some norms in contexts that make the deviation valid. The framework of this paper thus shows where the context is relevant in decisions about norms by both (1) the agent who acts and (2) the others, AI agents or humans, who assess the validity of the first agent's public actions.

The research community, e.g., [Falcone *et al.*, 2013; Nardin *et al.*, 2016], identifies norm compliance as enhancing trustworthiness and trust. However, this is not so in real life. As the examples above illustrate, justified norm deviations indicate the benevolence and integrity components of trust and trustworthiness [Singh and Singh, 2023] better than blind norm compliance. Specifically, the relevant justifications state how the agent promoted a stakeholder's interests or values. This holds even for performative behaviors such as protests, where the stakeholder may be an individual, an organization, or the public.

One potential shortcoming is that our underlying theories and case law take a Western democratic perspective. Cultural effects can be important: e.g., in collectivist cultures, personal interest would offer weak(er) grounds. However, the cultural dependence could be turned into an advantage in that an agent could determine whether or not to deviate from a norm based on which cultural perspective is applicable.

This paper poses new theoretical challenges. One, how can we model compositionality in regards to norm deviation by an organization vis à vis its members, analogous to group ability [Singh, 1991]? Two, how can we capture accountability in STS [Chopra and Singh, 2016] in light of norm deviation?

Acknowledgments

We thank the anonymous reviewers for their helpful comments. MPS thanks the NSF (grant IIS-2116751) for support.

References

[Aydoğan *et al.*, 2021] Reyhan Aydoğan, Özgür Kafalı, Furkan Arslan, Catholijn M. Jonker, and Munindar P. Singh. NOVA: Value-based negotiation of norms. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(4):45:1–45:29, August 2021.

[BBC, 2022] BBC. Tesla adds chill and assertive self-driving modes, January 2022. https://www.bbc.com/news/technology-59939536. Accessed 2023-01-16.

[Bohman and Rehg, 2017] James Bohman and William Rehg. Jürgen Habermas. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. 2017.

[Bresciani *et al.*, 2004] Paolo Bresciani, Anna Perini, Paolo Giorgini, Fausto Giunchiglia, and John Mylopoulos. Tropos: An agent-oriented software development methodology. *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 8(3):203–236, May 2004.

[California, 1976] California. Tarasoff v. Regents of Univ. of California, 551 P.2d 334, 1976. Supreme Court of Calif.

[Canada, 1989] Canada. Andrews v. Law Society of British Columbia 1 S.C.R. 143, 1989. Supreme Court of Canada.

[Chopra and Singh, 2016] Amit K. Chopra and Munindar P. Singh. From Social Machines to Social Protocols. In *Proc. WWW Conference*, pages 903–914, Montréal, April 2018.

[Chopra and Singh, 2018] Amit K. Chopra and Munindar P. Singh. Sociotechnical systems and ethics in the large. In *Proc. AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, pages 48–53, February 2018.

[Cranefield *et al.*, 2017] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. No pizza for you: Value-based plan selection in BDI agents. In *Proc. 26th IJCAI*, pages 178–184, Melbourne, 2017.

[Dastani *et al.*, 2018] Mehdi Dastani, Paolo Torroni, and Neil Yorke-Smith. Monitoring norms. *Knowledge Engineering Review*, 33:e25, December 2018.

[Dell'Anna *et al.*, 2020] Davide Dell'Anna, Mehdi Dastani, and Fabiano Dalpiaz. Runtime revision of sanctions in normative multi-agent systems. *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 34(2):43:1–43:54, June 2020.

[Dubljević *et al.*, 2022] Veljko Dubljević, Sean Douglas, Jovan Milojevich, Nirav Ajmeri, William A. Bauer, George F. List, and Munindar P. Singh. Moral and social ramifications of autonomous vehicles. *Behaviour & Information Technology*, 42, April 2022.

[Dworkin, 1988] Ronald Dworkin. *Law's Empire*. Harvard University Press, Cambridge, Massachusetts, 1988.

[England & Wales, 2001] England & Wales. Re A (conjoined twins) [2001] 2 WLR 480, September 2001. Court of Appeal of England and Wales.

[Falcone *et al.*, 2013] Rino Falcone, Cristiano Castelfranchi, Henrique Lopes Cardoso, Andrew J. I. Jones, and Eugénio Oliveira. Norms and trust. In Sascha Ossowski, editor, *Agreement Technologies*, pages 387–408. Springer, 2013.

[Finnis, 2003] John M. Finnis. Law and what I truly should decide. *The American Journal of Jurisprudence*, 48(1):107–129, June 2003.

[Flanagan, 1954] John C. Flanagan. The critical incident technique. *Psychological Bulletin*, 51(4):327–358, July 1954.

[Freeman, 2021] James B. Freeman. Walton on ethical argumentation. *Journal of Applied Logics: The IfCoLog Journal of Logics and their Applications*, 8(1):124–144, February 2021.

[Habermas, 1984] Jürgen Habermas. *The Theory of Communicative Action, volumes 1 and 2*. Polity Press, Cambridge, UK, 1984.

[Hohfeld, 1919] Wesley Newcomb Hohfeld. *Fundamental Legal Conceptions as Applied in Judicial Reasoning and other Legal Essays*. Yale University Press, New Haven, Connecticut, 1919. A 1919 printing of articles from 1913.

[Kafalı *et al.*, 2020] Özgür Kafalı, Nirav Ajmeri, and Munindar P. Singh. DESEN: Specification of sociotechnical systems via patterns of regulation and control. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 29(1):7:1–7:50, February 2020.

[Kökciyan and Yolum, 2022] Nadin Kökciyan and Pinar Yolum. Taking situation-based privacy decisions: Privacy assistants working with humans. In *Proc. 31st IJCAI*, pages 703–709, Vienna, July 2022.

[Kola *et al.*, 2020] Ilir Kola, Myrthe L. Tielman, Catholijn M. Jonker, and M. Birna van Riemsdijk. Predicting the priority of social situations for personal assistant agents. In *Proc. PRIMA*, LNCS 12568, pages 231–247, Nagoya, Japan, November 2020. Springer.

[Liscio *et al.*, 2022] Enrico Liscio, Michiel van der Meer, Luciano Cavalcante Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. What values should an agent align with? An empirical comparison of general and context-specific values. *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 36(1):23, 2022.

[Montes and Sierra, 2021] Nieves Montes and Carles Sierra. Value-guided synthesis of parametric normative systems. In *Proc. 20th AAMAS*, pages 907–915, Online, May 2021.

[Murukannaiah *et al.*, 2020] Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn M. Jonker, and Munindar P. Singh. New foundations of ethical multiagent systems. In *Proc. 19th AAMAS*, pages 1706–1710, Auckland, May 2020. IFAAMAS. Blue Sky Ideas Track.

[Nardin *et al.*, 2016] Luis G. Nardin, Tina Balke-Visser, Nirav Ajmeri, Anup K. Kalia, Jaime S. Sichman, and Munindar P. Singh. Classifying sanctions and designing a conceptual sanctioning process model for socio-technical systems. *Knowledge Engineering Review (KER)*, 31(2):142–166, March 2016.

[Rokeach, 1973] Milton Rokeach. *The Nature of Human Values*. Free Press, New York, 1973.

[dos Santos *et al.*, 2017] Jéssica Soares dos Santos, Jean de Oliveira Zahn, Eduardo Augusto Silvestre, Viviane Torres da Silva, and Wamberto Weber Vasconcelos. Detection and resolution of normative conflicts in multi-agent systems: A literature survey. *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 31(6):1236–1282, November 2017.

[Savarimuthu *et al.*, 2008] Bastin Tony Roy Savarimuthu, Maryam Purvis, Martin K. Purvis, and Stephen Cranefield. Social norm emergence in virtual agent societies. In *Declarative Agent Languages and Technologies VI: Revised Selected and Invited Papers*, LNCS 5397, pages 18–28, Estoril, Portugal, May 2008. Springer.

[Schafheitle *et al.*, 2020] Simon Schafheitle, Antoinette Weibel, Nadine Meidert, and Dirk Leuffen. The road to trust. A vignette study on the determinants of citizens' trust in the European Commission. *Journal of Common Market Studies (JCMS)*, 58(2):256–275, March 2020.

[Schwartz, 2012] Shalom H. Schwartz. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1):3–20, December 2012.

[SCOTUS, 1971] SCOTUS. Clay v. United States: 403 U.S. 698, June 1971. Supreme Court of the United States.

[SCOTUS, 1977] SCOTUS. Nat'l Socialist Party of Am. v. Vill. of Skokie, 432 U.S. 43, June 1977. Supreme Court of the United States.

[SCOTUS, 1996] SCOTUS. Jaffee v. Redmond, 518 U.S. 1, June 1996. Supreme Court of the United States.

[SCOTUS, 2018] SCOTUS. Masterpiece Cakeshop, Ltd. v. Colorado Civil Rights Comm'n 138 S. Ct. 1719, June 2018. Supreme Court of the United States.

[Serramia *et al.*, 2018] Marc Serramia, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, Manel Rodríguez, Michael Wooldridge, Javier Morales, and Carlos Ansótegui. Moral values in norm decision making. In *Proc. 17th AAMAS*, pages 1294–1302, Stockholm, 2018.

[Shah *et al.*, 2021] Vedarsh Shah, Zedong Peng, Ganesh Malla, and Nan Niu. Towards norm classification: An initial analysis of HIPAA breaches. In *Proc. 29th IEEE International Requirements Engineering Conference Workshops*, pages 415–420, Notre Dame, Indiana, 2021.

[Shams *et al.*, 2020] Zohreh Shams, Marina De Vos, Nir Oren, and Julian A. Padget. Argumentation-based reasoning about plans, maintenance goals, and norms. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 14(3):9:1–9:39, March 2020.

[Singh and Singh, 2023] Amika M. Singh and Munindar P. Singh. Wasabi: A conceptual model for trustworthy artificial intelligence. *IEEE Computer*, 56(2):20–28, February 2023.

[Singh, 1991] Munindar P. Singh. Group ability and structure. In Yves Demazeau and Jean-Pierre Müller, editors, *Decentralized Artificial Intelligence, Volume 2*, pages 127–145. Elsevier/North-Holland, Amsterdam, 1991.

[Singh, 2013] Munindar P. Singh. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):21:1–21:23, December 2013.

[Singh, 2022] Munindar P. Singh. Consent as a foundation for responsible autonomy. *Proc. 36th AAAI*, 36(11):12301–12306, February 2022. Blue Sky Track.

[NHTSA, 2022] NHTSA. Part 573 Safety Recall Report. Recall 22V-037. January 2022. Tesla rolling stop recall.

[Telang *et al.*, 2021] Pankaj R. Telang, Munindar P. Singh, and Neil Yorke-Smith. Maintenance of social commitments in multiagent systems. *Proc. 35th AAAI*, 35(13):11369–11377, February 2021.

[Thoreau, 1848] Henry David Thoreau. *On the Duty of Civil Disobedience*. Lerner Publishing, Minneapolis, 1848. 2020 reprinting.

[Tubella *et al.*, 2019] Andrea Aler Tubella, Andreas Theodorou, Frank Dignum, and Virginia Dignum. Governance by glass-box: Implementing transparent moral bounds for AI behaviour. In *Proc. 28th IJCAI*, pages 5787–5793, Macau, August 2019.

[Vermont, 1910] Vermont. Ploof v. Putnam, 83 Vt. 494, February 1910. Supreme Court of Vermont.

[Von Wright, 1963] Georg Henrik Von Wright. *Norm and Action: A Logical Enquiry*. International Library of Philosophy and Scientific Method. Humanities Press, New York, 1963.

[Walton *et al.*, 2008] Douglas N. Walton, Chris Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, Cambridge, UK, 2008.

[Weisbach, 1999] David A. Weisbach. Line drawing doctrine and efficiency in the tax law. *Cornell Law Review*, 84(6):1627–1681, September 1999.

[Wolff, 1998] Robert Paul Wolff. *In Defense of Anarchism*. University of California Press, Oakland, California, 1998.

[Woodgate and Ajmeri, 2022] Jessica Woodgate and Nirav Ajmeri. Macro ethics for governing equitable sociotechnical systems. In *Proc. 21st AAMAS*, pages 1824–1828, Auckland, May 2022. IFAAMAS.

[Yazdanpanah *et al.*, 2021a] Vahid Yazdanpanah, Enrico H. Gerding, Sebastian Stein, Corina Cîrstea, m. c. schraefel, Timothy J. Norman, and Nicholas R. Jennings. Different forms of responsibility in multiagent systems. *IEEE Internet Computing (IC)*, 25(6):15–22, November 2021.

[Yazdanpanah *et al.*, 2021b] Vahid Yazdanpanah, Enrico H. Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M. Jonker, and Timothy J. Norman. Responsibility research for trustworthy autonomous systems. In *Proc. 20th AAMAS*, pages 57–62, May 2021. Blue Sky Ideas Track.

[Young, 2001] Iris Marion Young. Activist challenges to deliberative democracy. *Political Theory*, 29(5):670–690, October 2001.