Failing to comply: on deviating from social conventions

Paul E. Dunne
Dept. of Computer Science
University of Liverpool
Liverpool L69 7ZF, UK
ped@csc.liv.ac.uk

Abstract

The analysis of how social conventions emerge and become established is rightly viewed as a significant study of great relevance to models of legal and social systems. Such conventions, however, do not operate in a monotonic fashion, i.e. the fact that a convention is recognised and complied with at some instant is no guarantee it will continue to be so indefinitely. In total rules and protocols may evolve, with or without the consent of individual members of the society, even to the extent that some cease to be observed or effective. In this paper we examine a framework for examining such changes in behavioural conventions that uses a proposed "taxonomy of social conventions" as the basis of a qualititative model deriving from value-based argument systems.

1 Introduction

Choose any field of human activity and one will find action, expression, and obligation within it governed (at least implicitly) by some canon of rules, regulations, and acceptable practice. The nature of such systems varies greatly in the extent to which they are codified; their flexibility of interpretation; their capacity to evolve; and the scope of their applicative domain. Thus, at one extreme there are the rules prescribing legal play in games such as Chess that precisely specify, for example, how pieces may be moved and the conditions under which games finish together with their outcome. Tending towards less rigid schemata are the numerous attempts to promulgate coherent systems of "aesthetic principles", e.g. in music, literature, painting. Ultimately, and arguably richest in its liberality of provision, the vast panoply of human social conventions and norms, together with the companion corpora of codified and understood strictures, rights, and duties.

There is, of course, a significant body of research dedicated to the study of how social norms and conventions become established, ranging from essentially game-theoretic analyses of action in multiagent systems, e.g. [12, 19, 20, 27], through considerations of the rôle and properties of normative settings and their relationship with deontic logic, e.g. [5, 6, 21, 26], to general studies, such as those presented in [1, 13, 25].

Our interest in the present paper is in a rather different aspect of social convention: we consider not the issue of how such conventions emerge and are established, but rather the mechanisms and influences by which recognised conventions cease to reflect societal attitudes to the extent that these can no longer be regarded as genuinely normative.

One may advance a number of reasons why what one might term the degradation (as opposed to evolution) of social conventions is worthy of attention. Firstly, it is undoubtedly the case that such development *does* occur, e.g. "dress codes" that prescribed "formal" attire as *de rigeur* have in many arenas fallen into abeyance; societal attitudes to the use of so-called "recreational drugs" have changed significantly over the past century, especially if included within this category are tobacco and alcohol. These attitudinal shifts may be reflected in legislative changes, e.g. the 18th Amendment (1919); its repeal by the 21st Amendment (1933);[1] and statutes occasioned by changing societal *mores* and tolerances, etc. As a further point, it is clear that the susceptibility of social conventions to become disregarded is influenced by many diverse factors. For example, that an attitude is endorsed by "popular" support may, in itself, fail to provide sufficient impetus for subsequent changes in legislation to be enacted. In addition, there is the nature of the relationship between a group of whom observance of a particular convention is expected and those (outside but interacting with such groups) from whom such expectations arise, e.g. individuals receiving medical treatment or advice expect the codes that govern patient-doctor relationships to be observed irrespective of the fact that they themselves are not medical professionals. In such circumstances as those outlined above it is natural to consider to what extent a particular convention is inherently "fragile" or "unstable", that is to assess its potential to shift from being seen as "social norm" to a state where its observance is treated as socially neutral or, even, in extreme cases, socially anathema.

In sum these motivating issues, raise the principal concern addressed in the main technical matter of this paper: the formulation of an approach to describe processes of normative degradation, together with a review of its analytic properties and applicability. As minimal *desiderata* one would expect such a framework sensibly to encapsulate "normative instability".

Central to the view promoted by the model described in Section 5 is the contention that the changes in societal attitude of interest are appropriately assessed through qualitative rather than quantitative techniques. Thus, in the widely cited study of emergence of norms presented in [20], the analytic and empirical systems are based on a schema within which normative conventions are modeled through "strategy profiles" representing choices of action adopted by individual agents in a k-player game: each such profile affording a particular pay-off to the agents according to the element it indexes in the game matrix. This view leads to an intepretation of (systemic) norm as constituted by those profiles representing divers game-theoretic equilibria. Now, although it is undoubtedly true that such approaches have at their disposal a highly-developed and

[1] One should, however, note the commentary of Munger and Schaller [16] who have argued that neither amendment was driven by prevailing contemporary social attitudes.

powerful analytic apparatus via established game-theoretic tools, it is arguable to what extent such models provide "effective" insight into the evolution of norms in practice. Thus it can be argued that compliance with social conventions is determined not by explicit quantitative motivations of "reward" and "penalty" but rather by more elusive qualitative factors, e.g. "social standing".

The foundational basis we seek to exploit subsequently is provided by the *argument systems* of Dung [8], in particular in their development to the *value-based* frameworks (VAFs) introduced by Bench-Capon [2, 3] and whose properties are studied in detail in recent work of [9, 10]. While we consider the formal instantiation of VAFs as vehicles for modeling norm decay in Section 5, it is useful at this juncture to mark some salient properties that, we contend, render these especially well-suited to the concerns of this paper. In particular we note the significance of the rôle of *audience*: that a subgroup within a society should coalesce so as collectively to deviate from compliance with a given convention can be explained in terms of that subgroup reassessing those "values" (in the sense of [3]) it prioritises in determining actions and beliefs. As such a social convention can be interpreted as indicative of a "shared" sense of values governing behaviour so that a sufficiently "radical value shift" accounts for non-compliance. We note that this view admits an interpretation of "set of social conventions" as a non-monotonic structure: actions become/cease to be normative as competing audiences shift in and out of "fashion".

Of course, as we observed in the opening paragraphs of this paper, the stability of a set of social conventions is dependent on the nature of the system and society to which these pertain.

2 Overview and Related Work

Our starting point for what we understand by "social convention", derives from the classical view of Lewis [13], so that we interpret a convention as a regularity in behaviour that has evolved within a society of at least two agents. Castelfranchi and Tummolini [6] consider such conventions as mechanisms leading to solutions to coordination problems: thus, in the terminology of [6], conventions emerge in the "positive expectation" that particular actions in recognised situations will achieve coordinated behaviour. A significant aspect of the argument and analysis of [6] is in their separation of "forecasts", – i.e. agent X predicts that executing an action α will lead to agent Y executing action β – from "positive expectations", i.e. in order to facilitate coordination, X *expects* that after α, Y *will* perform β. Thus, in recurring situations the expectation "if X does α then Y will do β" assumes the strength of a *prescriptive obligation* for Y once X has acted.

Our focus is not primarily concerned with the semantics of conventions as a device for realising coordination or cooperative planning among autonomous agents. Thus, our interest is more in developing descriptive models relating to the breakdown of social conventions. In this regard, nearer to the concerns of this paper, is the work of Bicchieri [4] in which she discusses *why* conformance should arise: her view of "social norm" seeing this as dependent on matters of individual choice, preferences, and beliefs. As has been observed in [6], in so far as the concept of "expectation" arises in [4] it is as a means of explaining conformance rather than a stage towards the development of a prescriptive rule.

The two principal components of our approach are introduced over the next sections. In Section 3 we elaborate ideas concerning relative strengths of conventions by means of an informal taxonomy of these: the motivation supporting this is to separate conventions that are of the nature of *de facto* prescriptions from those which are rather more vulnerable to being ignored. We present illustrations of these ideas in some specific social contexts.

The second component, described over Section 4 and Section 5, involves a development of the Value-based Argument Framework (VAF) introduced in [2, 3], so that the taxonomy discussed in Section 3 together with changes in conventional stength are intrepreted via changes consequent upon shifts in audience preferences. In Section 6 we examine some analytic properties of the resulting system. We summarise and discuss developments in the conclusion.

3 A Taxonomy of Conventions

Notwithstanding the central issue that an action happening to define conventional behaviour for a society at some instant is no guarantee that this action will *always* be seen as such, it is nevertheless the case that in many domains there are defining conventions and expectations of such strength that to go against these is to act in a manner incompatible with participation in the society concerned. Thus, to take a very basic example, consider the board game Chess: should one or even *both* players decide that Black should make two moves in response to each move made by White (instead of the strict alternation of one move at a time by each player) then the board game being played is *not* Chess. In other words, embedded within the understanding of what constitutes "playing Chess" there are immutable social conventions binding players to an acceptance of those procedural "rules" defining the game, i.e. while there may be agreement or attempts to apply different procedures, the nature of this level of social convention is such that one cannot change the fundamental conventions *and* continue to be "playing Chess".

Conventions of the type outlined are, relative to the internal mechanics of the social system, effectively *irrefragable*. On the other hand, if we again focus on Chess play as an illustration, one also has far less stringent conventions governing what might be termed play etiquette, e.g. "a player should concede defeat (resign) when it is clear that their opponent has a winning material advantage – say King and Rook versus King – without demanding that checkmate be delivered." Now one cannot claim that a player who ignores this convention by insisting that a game be played to a conclusion is "not playing Chess". In total within a domain although there may be a number of differing conventions it is not always the case that these carry equal strength.[2]

Our interest is principally in those conventions "most likely" to cease as exemplars of normative standards, and thus before we can begin to set up possible models of this

[2]Notice this suggests some "Draw Rules" cannot be considered as irrefragable: there is a difference between "a game is drawn if only Kings remain" (irrefragable) and "a game is drawn if fifty moves are played without a piece being taken". The latter could be changed but the resulting game is still Chess.

process, we require some mechanism for "classifying" social conventions within the range of extremes illustrated by the two examples above. Our aim in this section is to present such a "taxonomy of social conventions", defining its components and, finally, discussing its interpretations with respect to three widely different general contexts.

3.1 Overview

As a whole, given some social system we regard conventions within it as falling into one of the classes delineated in Table 1.

Name	Interpretation/Nature	Compliance
Irrefragable Conventions	Deviation is incompatible with participation in the society.	Universal
Thoughtless Conventions	Those so deeply entrenched as to be observed without *conscious decision*. Non-compliance is *possible* but is *always* a *deliberate act* that while not taking the agent outside the society, is seen as "eccentric" or as challenging the social order.	Near Universal
Standard Conventions	Those that are adhered to as a conscious choice. Non-compliance is "tolerated" but may attract sanctions.	"Significant" Majority
Fragile Conventions	Those respected by a subgroup, however, equally ignored by others. The extent to which non-compliance is penalised depends on how "cohesive" the factions ignoring it are and their influence on others.	"Marginal" Majority

Table 1: A Taxonomy of Social Conventions

We do not claim that the schedule presented in Table 1 is definitive, in the sense that its division as exactly four classes suffices. Nor is it asserted that, for any given system, its conventions at any one time are readily classified. The rationale supporting the categories presented is that these provide a starting point for the analysis in subsequent sections of this paper.

It is, of course, the case that (with the possible exception of the class of "fragile conventions") the other categories have been considered under various guises elsewhere. Thus what we have termed "irrefragable conventions" are of the same general nature as the "Constitutive Rules" in the analyses of Searle [18]. Similarly our terminology "thoughtless convention" originates from the study of these by Epstein [11].

The breakdown described in Table 1 will be of interest to the extent that it presents a basis for examining the factors that lead to social conventions becoming weaker, in particular those influences that relegate "standard" (or even "thoughtless") conventions to a lower status or push conventions outside a system altogether.

Some comments regarding our interpretation of changing status via the taxonomy presented are in order. First the class of irrefragable conventions in a society is assumed to be *monotone*: that is, a convention *may become* irrefragable, however it cannot *cease* to be so. It follows, therefore, that irrefragable conventions have a special significance in the sense that one may view these as the definitive arbiters of what constitute actions consistent and compatible with operating within the society. One might argue that the nomenclature "convention" for this class is, perhaps, inappropriate (a view, for example, implicit in Searle's term "constitutive *rule*"). While recognising this assertion as legitimate and of some interest, a greater elaboration of such issues lies outside the immediate scope of this work. Nevertheless we note one distinction that partly justifies our usage "convention": taking a "rule" to be a prescriptive requirement, the "conventional" nature with respect to an "irrefregable convention" lies in the *expectation* that agents comply with it (*sub poena* ostracism from the societal domain should they fail to do so). So for the Chess example: one rule requires strict alternation of player moves; the relevant irrefragable convention is that players are expected to observe this if they are "playing Chess".

As a further point, we note that our interpretative terminology is deliberately chosen to use rather vague (but, it is hoped) intuitively understood notions: we are not seeking, in the first instance, to ossify the structure of Table 1 within any formal quantitative or logical basis, e.g. an intuitive sense of what constitutes a "significant" majority is assumed. Finally, given the intention of describing the mechanics of normative degradation, in terms of the taxonomy proposed one would expect – having eliminated shifts from the irrefragable state – relegation from "thoughtless" to a "lower" category to be rare and that the bulk of those once-held conventions that cease to apply arise among those which at some point become "fragile". As such the schema presented in Section 5 is intended to highlight movement between standard and fragile states.

3.2 Some Example Domains

We continue by considering how this taxonomy relates to selected topics regarding the status of conventions in three general contexts, specifically:

A. Conventions in the presentation of Mathematical concepts.

B. The evolution of conventions in Western "Art" Music.[3]

C. Irrefragable conventions in Law – Principle and Practice.

[3] This term, while not ideal, is used in preference to both "Classical Music" (which technically refers to a specific period) and the, rather affected, alternative "Serious Music".

A. Mathematical presentation

One distinctive characteristic concerning "social convention" in this field is the nature of the relation between practitioners (internal to society) and observers (audiences external to the society, i.e. non-mathematicians): whether the example conventions we outline below are complied with or not, is largely a matter of indifference to external observers, e.g. the considerations of the "patient-doctor" scheme mentioned in Section 1 do not arise. From Table 1 we can identify, for example

I. **Irrefragable Conventions**
"Proofs must be *sound*, corresponding to a logical derivation of consequences from a set of accepted initial premises.", e.g. by correct application of logical inferences to axioms.

T. **Thoughtless Conventions**
"Assertions and claims whose 'validity' is 'non-obvious' require justification by a formal proof".

S. **Standard Conventions**
"Proofs are arranged so that dependencies in lengthy arguments are made clear *before* the proof itself.", e.g. a theorem which relies on a sequence of lemmata, identifies these in advance, rather than subsequently extracting "unsubstantiated" claims used in the theorem proof and proving them.

F. **Fragile Conventions**
"Use of a result (originally) obtained by X, should be accompanied by accurate citation of where X published it."

As a simple illustration of these ideas and specifically the sanctions for non-compliance, consider a mathematical paper that is submitted for publication but violates some of the conventions described. One would expect an article breaching (I) to be rejected without any detailed comment provided, particularly if the "proofs" within it showed no regard for the principle of "logical derivation", e.g. use of "non-mathematical" approaches[4] or elementary errors in argument. In case of (T) being ignored, depending on the significance of the omitted proof one might expect anything from outright rejection of the article (with some explanatory remarks being given) to an insistence that proofs be given before further review. Similarly, for (S), although not necessarily constituting reasons for immediate rejection, it would be normal for the author to be asked to revise the organisation of material before resubmitting the article.

It is worth remarking on the fragile convention identified, particularly since it could be argued that "proper citation" should be viewed as at least of the strength of standard convention, and, indeed, may even be observed thoughtlessly. To counter this objection, the precise wording of this convention must be noted: it does not relegate the convention of "citation" *in general* to "fragility" but is a comment about how widely known a *specific* contribution might become. As a concrete example of this, consider the following. A seminal publication in the mathematical discipline of computational

[4] e.g. from the present author's experience, "proof" that P=NP based on an analysis of the book of *Genesis*.

complexity theory is Cook's 1971 paper [7] in which he gives an analytic argument that deciding satisfiability of arbitrary propositional formulae was "unlikely" to admit an efficient algorithm, i.e. in the (now) established language of this field, [7] proved that "SAT is NP–complete". The vast majority of subsequent "NP–completeness" proofs that appeared in the mid-late 1970s include a citation to [7]: in effect referring to [7] in NP–completeness work became a "standard convention". From, the mid-1990s, however, such citations become increasingly rare in articles using "SAT is NP–complete" as a premise: instead are either references to widely used textbooks (thus the citation is indirect via the bibliography of the textbook); or no supporting citation whatsoever is supplied (indicative of a view that the methods of [7] are so firmly established in the perception of the audience that explicit citation of their source is felt to be redundant). In sum "cite [7] in NP–completeness proofs" has become a "fragile convention" from being a standard one. Ultimately one might see it no longer enjoying the force of convention at all: e.g. other than out of affectation, who among present-day combinatorial mathematicians justifies the approximation $n! \approx (n/e)^n\sqrt{2\pi n}$ by citing [22, Propn. 28]?

B. Audience/Artist Conventions in Music

In our review of (A) above, the classification of "conventions in mathematical presentation" is, we would claim, uncontroversial: certainly if these were to be ignored it is hard to imagine the collective society of non-mathematicians feeling concern. Our principal motivation for considering the development of social conventions within the topic now considered, is that this field illustrates it is not always possible easily to classify conventions within the divisions outlined in Table 1. In particular, the view of practitioners as to which actions (if any) define "conventions" may be noticeably at odds with the view of audiences.

The period since the development of a formal written notation, (*circa* 14th century), saw a process of quite radical changes in the perception of what constitutes "legitimate musical expression": thus the conventions governing plainsong differ from the intricate harmonic rules and procedures acknowledged by Bach and his contemporaries; these in turn have succumbed to a point that some commentators have described as a "crisis" in expressive form.[5]

We have noted that should a mathematical "proof" breach any (or, indeed, all) of the conventions given, then non-mathematicians are likely to be indifferent. Such equanimity, however, has often failed to be the typical reaction of audiences to works that are radically at variance with expectations. So, at one extreme, in Vienna in March 1913 the first performance of Berg's *Altenberglieder* as part of a concert promoting the new atonal form pioneered by Schönberg, sparked an audience riot of such violence that the remaining programme was abandoned with police intervention being necessary to restore order; barely two weeks later, the premiere of Stravinsky's *Le Sacre du Printemps* in Paris met with a similar fate, the intensity of the reaction in the latter case being exacerbated by the composer's "failure" to produce a work comparable to that idiom with which he had already achieved great success in Paris, e.g. *L'Oiseau de*

[5] e.g. "Few would dispute that a crisis exists. Some musicians who less than ten years ago expressed enthusiasm for the festivals of modern music now recoil perplexed." [17].

feu in 1909; *Petrushka* in 1911. Less dramatically, it is has been known for audiences to express disapproval of (what is perceived to be) an artist's breach of conventional expressive standards by heckling, booing, or walking out of the performance.

One reason why extreme reactions occur is given by the following: an audience holds (certainly "thoughtlessly", possibly even with the strength of "irrefragability") very precisely defined ideas of "what is music" and "what is noise". With novel techniques, however, work presented as "music" (within the system followed by its composer) is heard as "noise" (by audiences accustomed to traditional forms). In consequence audiences may feel "insulted" (seeing the new forms as a joke at their expense); and/or confused since they have no basis on which to "appreciate" or understand the logic of a piece. As a result it is unsurprising that such perceptions are manifested in the form of hostility towards and denigration of the works presented. It should be noted, however, that such hostility may not be permanent: as audiences become familiar with new idioms, works once regarded as "noise" are accepted, e.g. Stravinsky's *Le Sacre du Printemps* is now a standard concert piece.[6]

The significant issue illustrated by this example, is that consistent views of conventions and their strength may not be shared by those located within the system and those external to it, e.g. the former (composers) may feel free to disregard the convention "valid musical expression must recognise and relate to accepted objective standards"; the latter (audiences) may deem this to be"irrefragable".

C. The Concept of Irrefragable Convention in Law

In our final example overview we consider one legal code – the U.S. Constitution – and to what extent it may be interpreted *qua* system of irrefragable provisions. Many commentators have remarked upon the largely incremental and monotonic nature of this code, Lutz [14] going so far as to describe it as "the most stable in the world". This stability is partly a reflection of its original framers' intentions, who in recognising that a finite text could not account for every nuance and eventuality that might arise, prescribe a rigorous process by which amendments could be incorporated. Although its drafters may not have thought of these processes in terms of the notion of "irrefragable convention", it seems reasonable to argue that some degree of empathy with such is evident.

In principle, this view places the 27 (to date) ratified amendments as defining a set of "irrefragable provisions": thus, the approved amendments as a necessary clarification of lacunae within the original constitution, and, as *necessary*, therefore irrevocable. To write as such, however, is to write not of the reality but of the ideals strived for: 27 amendments have been ratified, however, only 26 of these remain in force, the sole purpose of the 21st being repeal of the 18th (Prohibition).

The example of the U.S. Constitution provides an illustration of the final issue of systemic classification of conventions we wish to review: the *principles* that a codified constitution sets out to defend *are* irrefragable, but the evolution of the *expressed* statement of these is the responsibility of *fallible agencies*: politicians, legislators, etc. The

[6]Such change may take considerable time: Berg's *Altenberglieder*, Schönberg's *Erwartung*, are even now, rarely performed.

dynamics of the processes underlying the 18th/21st Amendment ratifications exemplify this clash between the inviolable nature of an ideal and the fact that maintaining a codified realisation of these is at the mercy of fallible mechanisms.

In order to interpret the U.S. Constitution as a system of "irrefragable conventions" (that is, in principle if not in practice) one could view the Repeal Amendment as a discrepancy, with the advantage that such acknowledges the principle of fixed incremental change. This, however, has one drawback: it sanctions the validity of "deviant" amendments whereby the legislature (albeit in good faith) ratifies amendments in the interest of minority pressure groups, rather than from the stance of acknowledging a constitutional principle. The 18th Amendment has been noted in [23] as "the only constitutional attempt to incorporate a sumptuary provision into the fundamental law of the land", and stronger than merely "sumptuary" is the only ratified amendment that asserts the supremacy of the state over the freedom of action by private individuals where such action does not impinge upon the rights of others. Thus, an alternative interpretation is given by arguing that the 18th Amendment was an "unconstitutional" foray into areas in which the original drafters did not intend the state to interfere.[7]

3.3 Summary

Addressing the question of those mechanisms by which once established conventions cease to be observed, implicitly recognises that not all conventions within a system carry the same force, and thus the 'weakest' are those 'most likely' to be dispensed with. In this section our aim has been to outline a basic general classification of social conventions according to their 'strength'. Within the three example contexts we have considered a number of issues specifically related to this taxonomy.

The schema offered by our discussion of conventions in "Mathematical Presentation" could be seen as a benchmark of how convention shifts in 'ideal' systems 'should' operate: with no movement between the strongest categories and, for the example of fragile convention offered, changes that do occur being highly specific in nature rather than shifts in general principle..

Much of the stability within this area, however, stems from the fact that the society in which its conventions apply is (to all intents) independent from external scrutiny. As seen with the second example, when such external factors are relevant then participants and observers may hold very different views of conventions and strengths thereof operating within a system.

Finally our example from one particular legal code highlights the distinction, important as regards identifying irrefragable conventions, between the concepts of such as resulting from a system of idealised principles and their implementation via some written codification of such principles formulated by fallible agents.

[7] Although Tribe [24] has pointed out the peculiar nature of Section 2 of the Repeal Amendment has the unusual effect of placing transportation of alcohol into 'dry' states on a par with forcible enslavement: as one of "two ways, and two ways only, in which an ordinary private citizen, acting under her own steam and under color of no law, can violate the United States Constitution".

4 Value-Based Argument (VAF)

We now present a brief summary of the principal elements and concepts arising in Value-Based Argument Frameworks. These were introduced by Bench-Capon [2, 3] as a development of Dung's argument systems [8] which define an argumentation framework as:

Definition 1 *An* argumentation framework *is a pair $\langle \mathcal{X}, \mathcal{A} \rangle$ where $\mathcal{X}$ is a set of* arguments *and $\mathcal{A}$ is a binary relation on $\mathcal{X}$, i.e. $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{X}$. For two arguments x and y, the meaning of $\langle x, y \rangle \in \mathcal{A}$ is that x represents an attack on y. An argument $x \in \mathcal{X}$ is* acceptable *with respect to a set of arguments S if for evey $y \in \mathcal{X}$ that attacks x there is some $z \in S$ that attacks y. A set S of arguments is* conflict-free *if for every pair of arguments x and y in S neither attack the other. A conflict-free set of arguments S is* admissible *if for each x in S, x is acceptable to S. A set of arguments S is a* preferred extension *if it is a maximal (with respect to set inclusion) admissible subset of $\mathcal{X}$.*

There are a number of technical complications that arise in the "pure" schema offered by Dung's framework. Addressing some of these is one of the motivations underlying the extension to value-based frameworks. Formally, these are defined as.

Definition 2 *A* value-based argumentation framework *(*VAF*) is a 4-tuple: $\langle \mathcal{X}, \mathcal{A}, \mathcal{V}, \eta \rangle$, where $\langle \mathcal{X}, \mathcal{A} \rangle$ is an argument framework, $\mathcal{V}$ is a non-empty set of* values*, and η is a function which maps from elements of $\mathcal{X}$ to elements of $\mathcal{V}$. We say that an argument x* relates *to value v if accepting x promotes or defends v: the value in question is given by $\eta(x)$.*

An audience specific VAF *(*AVAF*) is a 5-tuple:*

$$\text{VAF}(\alpha) = \langle \mathcal{X}, \mathcal{A}, \mathcal{V}, \eta, \succ_\alpha \rangle$$

where $\mathcal{X}$, $\mathcal{A}$, $\mathcal{V}$ and η are as for a VAF*, $\succ_\alpha$ is a* value-preference relation *(transitive, irreflexive and asymmetric) over $\mathcal{V} \times \mathcal{V}$ reflecting the value preferences of an* audience *α. We write v_1 is preferred to v_2 as $v_1 \succ_\alpha v_2$. The* AVAF *relates to the* VAF *in that $\mathcal{X}$, $\mathcal{A}$, $\mathcal{V}$ and η are identical and $\succ_\alpha$ is the set of preferences derivable from α in the* VAF.

An argument x defeats *an argument y for audience α if and only if both $\langle x, y \rangle \in \mathcal{A}$ and $\neg(\eta(y) \succ_\alpha \eta(x))$. An argument x is* acceptable to audience α *with respect to a set of arguments S if: for each $y \in \mathcal{X}$ if y defeats x with respect to α then there is some $z \in S$ that defeats y with respect to α.*

A set S is conflict-free for audience α *if*

$$\forall x, y \in S \;\; \langle x, y \rangle \in \mathcal{A} \;\Rightarrow\; \eta(y) \succ_\alpha \eta(x)$$

A conflict-free for audience α set S is admissible for an audience α *if every $x \in S$ is acceptable to audience α with respect to S.*

A set of arguments S is a preferred extension for audience α *if it is a maximal (with respect to set inclusion) admissible set for audience α.*

Given a VAF*, $\langle \mathcal{X}, \mathcal{A}, \mathcal{V}, \eta \rangle$, an argument $x \in \mathcal{X}$ is* objectively acceptable *if and only if for all α, x is in the preferred extension for audience α.*

An argument $x \in \mathcal{X}$ is subjectively acceptable *if and only if for some α, x is in the preferred extension for audience α.*

An argument which is neither objectively nor subjectively acceptable (such as one attacked by an objectively acceptable argument with the same value) is said to be indefensible.

Two computational properties of interest for VAFs are given in

Theorem 1

i. For any given audience – $\langle \alpha, \succ_\alpha \rangle$ – there is a unique, non-empty preferred extension, and an efficient algorithm to determine it. (Bench-Capon [2])

ii. Given S there is an efficient algorithm to discover the class of audiences for which S is the preferred extension. (Dunne and Bench-Capon [9])

For a VAF, $H = \langle \mathcal{X}, \mathcal{A}, \mathcal{V}, \eta \rangle$ and audience α with preference relation $\succ_\alpha$, we denote this unique preferred extension by $PE(H, \alpha)$.

5 Conventions and Arguments

Value-based argument frameworks provide a model to account for how the ordering of abstract values among different agents relates to those arguments that the agents consider "acceptable" and are able to defend: as an agent's value priorities change, so too might those arguments that it regards as acceptable.

Recalling that our interest is in describing processes by which the strength of conventions alters, our model is based on the view that when an agent ceases to comply with a particular convention then this indicates the agent no longer "accepts" the convention as relevant to its actions. As to *why* such changes might occur, among possible reasons are:

> The reasons that led to the agent respecting the convention in the past are no longer relevant.
>
> The penalties for failing to comply are now seen as insufficiently discouraging.
>
> The advantages of non-compliance are felt to outweigh those of conforming.

Our central contention and that which underpins the use of VAFs, is that these are consequences of changes in an agent's value preferences: such changes allowing the agent either consistently to ignore the convention – i.e. although it is still "acceptable" in the sense of being consistent with the values held, the agent is not compelled to observe it – or, in more extreme cases, rendering the convention unacceptable since the agent cannot defend its observance with respect to the value ordering held.

5.1 Issues and Social Views

The informal description just given thereby leads to a model wherein abstract concepts of "social convention" correspond to the abstract notion of "argument" and "value" within a VAF.

In total to model the potential change in a convention's status between "standard", "fragile" and the level at which it ceases to be observed we have:

a. A representation of the arguments and factors determining whether or not an arbitrary agent is bound by the convention.

b. A representation of those factors that influence a given agent's interpretation of the system delineated by (a).

c. A process for classifying how given instantions of (b) across a society of agents determine the convention's state.

In abstract terms these lead to the concepts of *Issue* and *social view* to describe (a) and (b). That is,

Definition 3 *An* Issue *is a pair* $\mathcal{I} = \langle H, p\rangle$ *whose first component is a* VAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{V}, \eta\rangle$ *and whose second component, p, is an argument in* $\mathcal{X}$*, called the* proposed convention*.*

An r*-party* social view *is defined by a set of* $r \geq 2$ *pairs*

$$\mathcal{S} \ = \ \{ \ s_1, s_2, \ldots, s_r\} \ \ ; \quad s_i = \langle \alpha_i, \pi_i\rangle$$

with α_i *an ordering of* $\mathcal{V}$*, i.e. an* audience*, and* $\pi_i \subseteq \mathcal{X}$ *a set of arguments defining the* prejudices *or* beliefs *of an agent* a_i*.*

So in the scheme of Definition 3, *Issues* encapsulate not only the reasoned arguments that govern "acceptance" of the proposed convention ($p \in \mathcal{X}$) but also the values endorsed by these arguments. We regard *Issues* as "static" in the sense that having formulated the components relating to a particular convention under debate, these do not subsequently change. In contrast, the *social view* may develop so that over time the ordering of values in the component α_i of s_i is affected. Finally, before considering (c), we briefly discuss the rôle of π_i. For a given agent, two factors may influence its attitude towards p. The first being its ranking of the values within the system (as described by α); the second that the agent may wish to continue complying with some existing conventions or will accept p only if this also permits it to accept other arguments in addition. Thus the subset π_i of $\mathcal{X}$ is intended to represent such existing beliefs and conventions as are expressed within the VAF defined by the $Issue$. We note that while it would be natural to expect $\langle \alpha, \pi\rangle$ to be "consistent" in the sense that π is a subset of the preferred extension $PE(H, \alpha)$ we do not require or assume this to be the case.

The following definition proposes an approach to classifying the status of a convention p within an $Issue$ $\mathcal{I} = \langle H, p\rangle$ from the perspective of an r-party social view $\mathcal{S}$.

Definition 4 *Let* $\mathcal{I} = \langle H, p\rangle$ *be an Issue and* $\mathcal{S}$ *an* r*-party social view. The (proposed) convention* p *is* accepted by $s_i \in \mathcal{S}$ *if* $p \in \pi_i$ *or* $p \in PE(H, \alpha_i)$*. Furthermore,*

p is thoughtless *if: for each $s_i \in \mathcal{S}$, $p \in \pi_i$.*
p is standard *if:* $|\{\ s_i\ :\ p \in \pi_i\ \}| \geq \lceil 2r/3 \rceil$ or both $|\{\ s_i\ :\ p \in \pi_i\ \}|\ \geq\ \lceil\ r/2\ \rceil$ and

$$\frac{|\{\ s_i\ :\ p \notin \pi_i\ \ and\ \ p \in PE(H, \alpha_i)\}|}{|\{\ s_i\ :\ p \notin \pi_i\}|}\ \geq\ \frac{1}{6}$$

p is fragile *if: p is accepted by at least $\lceil\ r/2\ \rceil$ members of $\mathcal{S}$ and p is* not *a standard convention.*

We use the notation $\mathbf{T}(H, p, \mathcal{S})$, $\mathbf{S}(H, p, \mathcal{S})$, $\mathbf{F}(H, p, \mathcal{S})$ to indicate p being thoughtless, standard, or fragile.

Notice that we give translations from the classes of Table 1 only for those cases where non-compliance is compatible with continuing membership, i.e. we do not present an analogue for irrefragable conventions. We have two reasons for this: firstly, the framework is intended to capture ideas of conventions becoming weaker and it has been assumed that such behaviour is not relevant for irrefragable conventions; secondly, whether a convention is irrefragable is not a property determined by the value ordering (α) or belief set (π) of an individual agent. It is noted, however, that in principle one could define a convention, p, (in the scheme of Definition 3) as irrefragable if $p \in PE(H, \alpha)$ for *every* possible value ordering, i.e. if p were *objectively acceptable*: one reason for *not* doing so is given in Theorem 3(a).

We conclude by commenting on the choice of definitions for the other categories used in Definition 4 and how these relate to the informal descriptions of Table 1.

In all three cases, p in order to qualify as a convention must be accepted by a majority of of the members of $\mathcal{S}$. What distinguishes the different categories is the extent to which p already features in the belief set, π_i, of each member of $\mathcal{S}$. For the concept of *thoughtless* convention it is required that p belong to every such set: in effect the interpretation is that when $p \in \pi_i$, s_i does not have to "think" about conformance, since this is given *a priori* and from s_i's view requires no further formal justification, e.g. that p is consistent with the value ordering α_i in the sense that $p \in PE(H, \alpha_i)$. It is noted that the condition "$p \in \pi_i$ for all s_i" is more restrictive than that suggested by the compliance category of Table 1: this, however, is adopted merely to simplify the presentation.

The formulation of "standard convention" is more involved: we have as a precondition that the majority of agents accept p thoughtlessly, i.e. have $p \in \pi$, and in order to ensure some degree of "significant majority" it may further be required that sufficient of the remainder can be persuaded complying with p is consistent with the value ordering they endorse. An immediate consequence of these two conditions is that a standard convention in this sense must be accepted by at least two-thirds of the agents in $\mathcal{S}$.

Finally we regard acceptance of p by a majority of $\mathcal{S}$ but by one of insufficient strength to meet the conditions for p to be a "standard convention", to define p as a "fragile convention". The reasoning supporting this view is as follows. Consider the ways in which p could be accepted by a majority yet *fail* to be a standard convention. Either we have $|\{s_i\ :\ p \in \pi_i\}| \geq \lceil r/2 \rceil$ but without enough of the remaining agents holding a consistent basis upon which to accept p, i.e. $p \in PE(H, \alpha)$; or, we have $|\{s_i\ :\ p \in \pi_i\}| < \lceil r/2 \rceil$ and sufficiently many of the remainder can be persuaded that p may be observed without contravening the agent's value ordering. Recalling that, unlike $\langle H, p \rangle$, the system $\mathcal{S}$ is not static, there is the possibility of individual agents revising their value ordering (α) and/or their belief sets (π).

Now, suppose p is a fragile convention by reason of $|\{s_i \; : \; p \in \pi_i\}| \geq \lceil r/2 \rceil$. In this case a majority of agents are predisposed to comply with p (in effect "thoughtlessly") and thus in order for p to strengthen in status to a standard convention requires "only" that a sufficient number of non-accepting agents either change or *can be persuaded* to change their value ordering so as to admit p; alternatively, that agents for whom $p \in PE(H, \alpha_i)$ can be persuaded to commit p to their belief set. Thus this case can be viewed as a fragile convention which, nevertheless, has some "reasonable" possiblility of becoming a standard one.

Alternatively suppose p is a fragile convention with $|\{s_i \; : \; p \in \pi_i\}| < \lceil r/2 \rceil$. The status of p being accepted by a majority is dependent on the value orderings of agents that have $p \in PE(H, \alpha)$: should sufficient of these change to leave $p \notin PE(H, \alpha)$, then the result may be that p is no longer accepted by a majority. Thus this second case can be viewed as a fragile convention which may potentially cease to define normative behaviour once a minority of agents observe it.

5.2 Summary

The abstract system just described presents an approach to describing levels of compliance with a convention p using a VAF structure $H = \langle \mathcal{X}, \mathcal{A}, \mathcal{V}, \eta \rangle$ (with $p \in \mathcal{X}$) to delineate the underlying arguments for and against p being accepted. In order to relate this static form to the question of how compliance with p evolves relative to the classification of social conventions in Table 1, the society of agents concerned are described in terms of individual value preferences (α) and beliefs ($\pi \subseteq \mathcal{X}$). The mechanisms described in Definition 4 give a rather basic translation onto the classes "thoughtless", "standard", and "fragile" convention and it is, of course, the case that thresholds other than those given could be considered, e.g. relaxing the condition $p \in \pi_i$ for all s_i with respect to thoughtless acceptance; modifying the simple majority aspects in the formulation of standard convention. The point of Definition 4, however, is not to promote the literal thresholds used as definitive: the values indicated are intended merely as "reasonable" choices. This key idea proposed within this definition is that the characterisation of a convention's strength is dependent on the relationship between the proportion of agents accepting p without argument, i.e. with $p \in \pi$ and the proportion of the remainder accepting p on the basis of their value preferences, i.e. for which $p \in PE(H, \alpha)$: both belief sets and value orderings, in contrast to H, being susceptible to change.

6 Analytic Properties

Given an Issue $\langle H, p \rangle$ and r-party social view $\mathcal{S}$ there are a number of natural questions that arise. So, for example,

a. Which, if any, of the following is true of p?

$\mathbf{T}(H, p, \mathcal{S})$	i.e. is p a thoughtless convention?
$\mathbf{S}(H, p, \mathcal{S})$	i.e. is p a standard convention?
$\mathbf{F}(H, p, \mathcal{S})$	i.e. is p a fragile convention?

b. Is it the case that the belief set of s_i is consistent with its value ordering in the context of H, i.e. is $\pi_i \subseteq PE(H, \alpha_i)$?

c. Is the belief set of s_i a consistent set with respect to *some* value ordering, i.e. is there a value ordering α, for which π_i is *admissible* with respect to α?

d. Is the belief set of s_i a *maximally* consistent set with respect to *some* value ordering, i.e. is there a value ordering α, for which $\pi_i = PE(H, \alpha)$?

As a straightforward consequence of Theorem 1, it is easy to see

Theorem 2

a. Each of the properties $\{\mathbf{T}(H, p, \mathcal{S}), \mathbf{S}(H, p, \mathcal{S}), \mathbf{F}(H, p, \mathcal{S})\}$ *is decidable by a polynomial-time algorithm.*

b. Checking if $\pi_i \subseteq PE(H, \alpha_i)$ *can be done in polynomial-time.*

c. Checking if there is a value ordering, α*, for which* π_i *is admissible with respect to* α *and constructing such an ordering if one exists, can be accomplished by a polynomial-time algorithm.*

d. Checking if there is a value ordering, α*, for which* $\pi_i = PE(H, \alpha)$*, and constructing such an ordering if one exists, can be accomplished by a polynomial-time algorithm.*

Proof: (a) and (b) are immediate from Theorem 1(i) and Definition 4; both (c) and (d) follow from the algorithm of Theorem 1(ii) presented in [9], with a minor modification to this required in order to deal with (c). ▫

While Theorem 2 demonstrates the tractable nature of some basic questions within the model of Definitions 3 and 4, in principle, other such questions are less readily answered. One such result provides a reason for not formulating "p is an irrefragable convention" via "$\forall\ \alpha\ \ p \in PE(H, \alpha)$".

Theorem 3

a. If $\mathbf{I}(H, p, \mathcal{S})$ *– "p is an irrefragable convention" – is defined to hold whenever* $p \in PE(H, \alpha)$ *for every value ordering* α*, then deciding* $\mathbf{I}(H, p, \mathcal{S})$ *is* CO-NP*–complete.*

b. Deciding if there is at least one value ordering, α*, for which* $\pi_i \subseteq PE(H, \alpha)$ *is* NP*–complete.*

Proof: Part (a) is simply [10, Theorem 2], since the decision problem is equivalent to deciding if an argument is objectively accepted. Similarly (b) follows by considering the special case of π_i containing exactly one argument, so that the problem is equivalent to deciding if an argument is *subjectively* accepted: a problem shown to be NP–complete in [10, Theorem 1]. ▫

As was noted in [10], however, the constructions by which the classifications of Theorem 3 are derived, could be seen as rather unrealistic in the context of such VAFs as might occur in realistic applications since both results employ VAFs in which $|\mathcal{V}|$ is $O(|\mathcal{X}|)$: in typical cases it is more likely that the number of distinct values is constant, in which event both decision problems in Theorem 3 may be dealt with using polynomial time algorithms.

The questions considered in Theorems 2 and 3 are all concerned with somewhat static aspects of a system $\langle\langle H, p\rangle, \mathcal{S}\rangle$. Of rather more immediate interest, however, are questions regarding properties of fragile conventions as these evolve or degrade: given that, within the schema presented, deciding if a convention *is* fragile, i.e. whether $\mathbf{F}(H, p, \mathcal{S})$ holds, is efficiently testable, we concentrate in the remainder of this section on systems for which the proposed convention has this property.

In these questions the aspects of interests concern dynamic properties of the system $\langle\langle H, p\rangle, \mathcal{S}\rangle$. Thus suppose that p is fragile. As we observed there are two possibilities, one for which p has a tendency towards satisfying $\mathbf{S}(H, p, \mathcal{S})$, the other where p may cease to be observed by a majority of agents. Now, noting that the relevant applicable option is easily determined, in the former we have the following: given that $1/2 \leq |\{s_i : p \in \pi_i\}|/r < 2/3$ if the agents holding p within their belief sets wish to promote p to the status of standard convention then they require some case of sufficient weight to change the view of those agents for whom $p \notin \pi_i$. One may interpret this idea of "case" as, in effect, a meta-argument of the form that a "rational" consequence of an agent's (current) state (α_i, π_i) is that one (or both) of these components must change. Such cases may be expressed via a template

$$\Phi(s_i, \langle H, p\rangle) \ \triangleright \ (s_i := \gamma_1(s_i))| \ldots |(s_i := \gamma_t(s_i))$$

i.e. "it is a rational consequence ($\triangleright$) of a particular property holding of the current state with respect to the Issue ($\Phi(s_i, \langle H, p\rangle)$) that this state be replaced by one of a number of defined alternatives ($s_i \ := \ \gamma_j(s_i)$)".

For example agents might try to raise a fragile convention to the level of standard convention by putting the following meta-argument ("consistency of belief set" – CBS) to agents for whom $p \notin \pi_i$

$$p \in PE(H, \alpha_i) \ \ \triangleright \ \ (\pi_i := \pi_i \cup \{p\}) \ | \ (\alpha_i \ := \ \beta_i)$$

That is, given that p is accepted under the value ordering endorsed by s_i it can be added to the belief set π_i, the alternative being that the value ordering be revised. Notice that those agents who accept CBS and act on it increase the proportion of agents in the system for whom $p \in \pi_i$, thereby moving p closer to the status of standard convention.

In order for CBS to effect a change it is necessary that the proportion of agents accepting p – either through $p \in \pi_i$ or $p \in PE(H, \alpha_i)$ – be at least $2/3$: otherwise, even if every agent with $p \in PE(H, \alpha_i)$ modifies its belief set in accordance with CBS, p may still remain fragile. If the number of agents accepting p is less than $\lceil 2r/3 \rceil$, then a different meta-argument could be applied: one which is predicated on an agent's belief set being incompatible with its value ordering thus leading to the proposal that the latter be changed. The meta-argument ("value-revision" – VR) is

$$\begin{array}{l} \pi_i \not\subseteq PE(H, \alpha_i) \text{ and } \pi_i \cup \{p\} \subseteq PE(H, \beta) \\ \rhd \\ (\alpha_i \ := \ \beta) \ | \ (\pi_i \ := \ PE(H, \alpha_i) \end{array}$$

In other words, given that π_i is not consistent with the preferred extension identified by the value ordering α_i, but (together with p) *is* consistent with that of another value ordering (β), the rational consequence that should be enacted is for s_i to adopt β.

For the second possibility, that of a fragile convention being accepted by a majority but with insufficient numbers including it in their belief set, although both of the templates CBS and VR may be applicable, there are a number of reasons why they may fail to have any effect: an agent may feel "more inclined" to act on either if they are proposed from several different sources, so that part of the motivation for adopting an alternative $s_i := \gamma_j(s_i)$ arises from a recognition that to act as such is in accordance with a consensus proposed by an existing *convinced* majority, i.e. those for whom $p \in \pi_i$. Thus in this case, there is less "external" pressure for a non-compliant agent to change their state. In terms of Definition 4, for fragile conventions one might have extremes ranging from $\forall\ i\ p \notin \pi_i \ \wedge \ p \in PE(H, \alpha_i)$ to $(\forall\ i\ p \notin \pi_i) \ \wedge \ |\{s_i \ : \ p \in PE(H, \alpha_i)\}| = \lceil r/2 \rceil$: the former requiring over half of the agents consistently to change their view in order that p change in status; the latter needing just a single agent to amend its value ordering so that $p \notin PE(H, \beta)$ for p to cease being accepted by a majority. The example we present in the next section considers one setting closer to the first possibility but which results in the proposed convention ceasing to be endorsed by a majority.

7 Example Study

To conclude we present a simplified example of how the mechanisms introduced over Sections 3–5 can be used in analysing degradation in a fragile convention's status. We consider a convention governing jury deliberation and decision making, using as a particular instance the fictional setting portrayed in Lumet's film *12 Angry Men*. This, of course, considerably simplifies the actual process of jury deliberation, compresses the timescale, and exaggerates character stereotypes for dramatic reasons. We consider this not as a "realistic" example of jury deliberation but in order to illustrate how reactions may be viewed in terms of compliance with a particular convention in the model introduced above. We do not attempt to give a full analysis in terms of the values and beliefs of the twelve individuals depicted: it is certainly possible to do so, however, this would require a quite extensive discussion. Instead we concentrate on a point which can be described via a subset of the individual jurors.

7.1 Background

The setting is that of a jury deciding a verdict in the case of a teenage boy accused of murdering his father: this must be unanimous and, in the event of a guilty decision, a capital sentence will be imposed. The film concentrates on the (ultimately successful) attempt by one juror to obtain an acquittal decision, even though all the other jurors have, initially voted guilty. A key element in this process is the conflict between verdicts formed by "irrational" reasons and those formed by interpreting the evidence. As this is elaborated the discussion evolves so that the main issue is the extent to which evidence presented should be accepted unchallenged as opposed to the jury independently assessing its reliability.

7.2 Issue and Social View Analysis

The proposed convention we examine is "The verdict reached by a jury should be based *solely* on the evidence that has been presented during the trial". Denoting this P, the VAF relating to our example involves five arguments each of which supports one of two values, which we refer to as "Rationality" (R) and "Sentiment" (S). These are given in Table 2 and graphically in Figure 1.

Notation	Meaning	Value	Attacks
P	Decide on evidence as given	R	{Y,W}
W	Background determines verdict	S	{X}
X	Defence was poorly argued	S	{P}
Y	Decide on evidence as analysed	R	–
Z	Wait before final verdict	S	{P}

Table 2: VAF for Example Issue

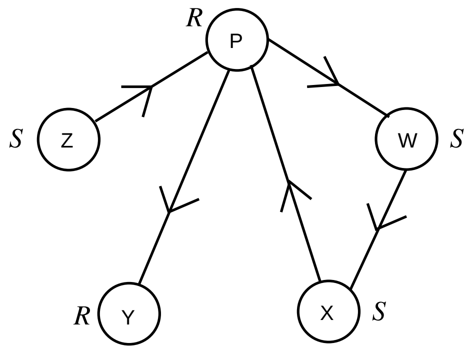

Figure 1: VAF example as directed graph H

The two value orderings result in the preferred extensions,

$$PE(H, R \succ S) = \{\text{P,X,Z}\} \;\; ; \;\; PE(S \succ R) = \{\text{W,Y,Z}\}$$

We have already discussed the proposed convention P. The remaining arguments are related to views that become apparent as the deliberation moves from the initial vote to the stage where individual jurors give a brief justification of their verdict.

Before expanding these arguments we first define the four individual agents considered and their adopted positions[8] in Table 3.

Agent	π	α
HF	{X,Z}	$S \succ R$
EB	{W}	$S \succ R$
LC	{W,P}	$S \succ R$
EM	{P}	$R \succ S$
Remaining eight	{}	$R \succ S$

Table 3: $\mathcal{S}$ – 12-party Social View for Example

Note that for the purposes of the example we view the remaining eight jurors as holding a common view.[9] In terms of Definition 4 the proposed convention is fragile: two out the twelve accept it *a priori*; of the ten remaining, eight accept P, by reason of it belonging to the preferred extension induced by $R \succ S$.

7.3 Development

The influences that establish the Issue $\langle H, p \rangle$ emerge and are resolved in six stages:

	Action	Raises
1.	Informal conversations prior to convening	P, W
2.	Open vote resulting in 11-1 majority for guilty	–
3.	HF outlines reasons for dissenting.	Z
4.	EM, LC, EB and others justify "guilty" verdict	P, W
5.	HF indicates further reasons	X, Y
6.	Secret vote (HF abstaining) gives 10-1 majority.	Y

The argument W – that the defendant's background determines the verdict – appear as a factor in both EB's and LC's views during (1), it becoming clear that both have decided their verdict having seen the accused, considering the defence arguments irrelevant.[10] Although P continues to be supported by EM, the nature of the discussion following (6) results in non-compliance with P. The significance of W is in revealing that two of the jurors – EB and LC – are not susceptible to *rational* argument: the accused is Hispanic and from a deprived background, factors which EB regards as indicative of guilt, irrespective of the trial process which he considers to be unjustifiably expensive. For very different reasons LC is also influenced by W: not on account of the racism and

[8]The initials are those of the corresponding actors – Henry Fonda, Ed Begley, Lee J. Cobb, and E.G. Marshall.

[9]This, of course, is to over simplify: one juror acts from a desire to end debate quickly.

[10]cf. LC's comment during (1), "What gets me is the way those lawyers talk and talk and talk even when it's an open-and-shut case like this one. Did you ever hear so much talk about nothing?"

class prejudice of EB, but because the age and family circumstances of the defendant provoke painful memories of unresolved issues with his own son.

The case for Z proposed by HF in (3) is, in the first instance, based entirely on the value S: given the consequence of a guilty verdict it would be indecent to return one in such a short time and HF suggests spending one hour deliberating. This proposal is accepted leading to (4) in which brief justifications of the guilty verdict are given with EB's comments indicating that his decision derives from W. The discussion progresses to (5), in which HF advances X – that the presentation of the defence case was badly handled – but this is not regarded significant: X is a consequence of using a court appointed public defendant; a more capable lawyer might well have failed to make a reasonable case against the evidence offered by the prosecution, etc. It is in trying to argue that there may be alternative interpretations of the evidence that HF then introduces Y: that the jury should try to analyse how cohesive a case the evidence presents, not simply accept it unchallenged (P). In order to strengthen this argument HF demonstrates the unsound nature of one point of the prosecution case – that the murder weapon, a knife with an ornate handle, was "unique" – by producing an identical one. This is sufficient to convince a second juror to change verdict in order that further discussion take place (6) – HF having offered to accept the majority verdict should the result of a secret vote continue to show eleven jurors returning guilty verdicts.

To conclude this example, we briefly consider the influence of EM who is of interest as the only person whose initial belief set – $\{$P$\}$ confroms with the proposed convention *and* is consistent with his value ordering. The final three jurors to reverse their initial guilty verdict are $\{$EB,LC,EM$\}$. Throughout the subsequent discussion, in which the validity of Y is acknowledged, EM's attitude favours R over S and supports P. The rational qualities, however, are limited in one significant aspect: EM's naive acceptance of the prosecution case: e.g. that witness testimony (the more so since unchallenged by the defence) is entirely accurate; that the defendant's alibi is not credible; that the murder weapon is "unique" (since the prosecutor has stated as much without contradiction). It is one noticeable feature of the dynamics of LC's thinking that he consistently echos and seeks support from EM (with value ordering $R \succ S$) rather than the equally entrenched but increasingly marginalised[11] EB (promoting $S \succ R$).

Thus while the other jurors admit "reasonable doubt" as successive points of the prosecution case are re-interpreted, EM although accepting the arguments given, compensates by placing increasing weight on the undiscredited evidence, i.e. complies with P in accordance with his belief set and the value ordering $R \succ S$. When the balance has altered to 9-3 in favour of acquittal, challenged to justify his stance, EM's reasoned response places the weight of the entire prosecution case on a single eye-witness, finally changing his verdict to not guilty, after accepting an analysis which indicates the witness wore glasses. This detail could only be established through considering evidence via the argument Y, since the witness had not worn glasses when testifying.

[11]A process that has begun during (4), when HF exposes the irrational basis of EB's decision. A key prosecution witness is from the same neighbourhood and background as the defendant: following his open statement of W, EB attempts to rationalise his verdict by referring to this witness, only to be challenged by HF's question, "You don't believe the boy. How come you believe the woman? She's one of 'them' too, isn't she?"

8 Conclusion

Our aim in this article has been to present a model of shifts in strengths of social conventions through a schema built on two structures: the informal classification of conventions discussed in Section 3; and an embedding of this into value-based argument frameworks. Some basic analytic properties of these models were discussed in Section 6. We conclude by briefly outlining some directions in which the preliminary models proposed may be developed.

We have stressed the view that changes in the status of a particular social convention are explicable in terms of underlying changes in value preferences over a section of the society, so that such qualitative ideas are as applicable as the quantitative concepts that have been used in game-theoretic approaches seeking to explain the emergence of conventions, e.g. through rewards providing incentives to comply, penalties forming sanctions for non-compliance. It would be of interest to formulate an analogue of the empirical studies presented in [20] and to assess the characteristics exhibited for varying classes of VAF and social view. Thus, considering the two meta-arguments for state change given in Section 6, one might consider criteria for agents to accept or reject these, so that the evolution of a proposed convention's status could be monitored. Similar questions concern the nature of acceptance for a proposed convention p relative to the VAF H in which it features, e.g. is there a meaningful interpretation of a proposed convention p when it is *indefensible*? More generally, it would be of interest to examine how the number of value orderings under which p is accepted relates to its stability regarding the taxonomy outlined. Finally, our example study in the previous section starts from a well-defined and "small" VAF. In modelling more complex cases, however, some suitable methodology to construct the argument setting and to allow it to evolve would be required. Addressing such issues forms the subject of continuing work.

References

[1] R. Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.

[2] T. J. M. Bench-Capon. Agreeing to differ: modelling persuasive dialogue between parties with different values. *Informal Logic*, 22(3), 2003.

[3] T. J. M. Bench-Capon. Persuasion in Practical Argument Using Value Based Argumentation Frameworks. *Journal of Logic and Computation*, 13(3):429–48, 2003.

[4] C. Bicchieri. Norms of cooperation. *Ethics*, 100:838–861, 1990.

[5] G. Boella and L. van der Torre. Δ: the social delegation cycle. In: *Proc. 7th Intnl. Workshop on Deontic Logic in Computer Science* (DEON'04) (Eds: A. Lomuscio and D. Nute), pages 29–42, Springer LNCS 3065, 2004.

[6] C. Castelfranchi and L. Tummolini. Positive and negative expectations and the deontic nature of social conventions. In: *Proc. Ninth ICAIL*, pages 119–125, ACM Press, 2003.

[7] S. A. Cook. The complexity of theorem-proving procedures. In *Proc. 3rd ACM Symposium on Theory of Computing*, pages 151–158, ACM Press, 1971

[8] P. M. Dung. On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming, and N-person games. *Artificial Intelligence*, 77:321–357, 1995.

[9] P. E. Dunne and T.J.M. Bench-Capon. Identifying Audience Preferences in Legal and Social Domains. In: *Proc. DEXA'04*, pages 518–527, Springer LNCS 3180, 2004.

[10] P. E. Dunne, T. J. M. Bench-Capon. Complexity in Value-Based Argument Systems. In: *Proc. 9th European Conference on Logics in Artificial Intelligence* (JELIA'04), pages 360–371, Springer LNCS 3229, 2004.

[11] J. M. Epstein. Learning to be thoughtless: social norms and individual computation. *Computational Economics*, 18:9–24, 2001.

[12] Y. Gal, A. Pfeffer, F. Marzo, and B. J. Grosz. Learning social preferences in games. In: *Proc. AAAI-2004*, pages 226–331. 2004.

[13] D. Lewis. *Convention: a Philosophical Study*. Harvard Univ. Press, 1969.

[14] D. S. Lutz. Toward a theory of constitutional amendment. *American Political Science Review*, 88:355–370, 1994

[15] R. McGrew and Y. Shoham. Using contracts to influence the outcome of a game. In: *Proc. AAAI-04*, pages 238–244, July 2004.

[16] M. Munger and T. Schaller. The prohibition-repeal amendments: a natural experiment in interest group influence. *Public Choice*, 90:139–163, 1997.

[17] H. Ottaway. Prospect and Perspective. In: The Symphony (Ed: R. Simpson), Pelican, 1967

[18] J. R. Searle. *Speech Acts*. Cambridge University Press, 1969

[19] Y. Shoham and M. Tennenholz. On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, 73:231–252, 1995.

[20] Y. Shoham and M. Tennenholz. On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence*, 94:139–166, 1997.

[21] T. Smith. Violation of norms. In: *Proc. Fourth ICAIL*, pages 60–65, ACM Press, 1993.

[22] J. Stirling. *Methodus Differentialis*, London, 1730

[23] W. Swindler. A dubious constitutional experiment. In: *Law, alcohol and order* (Ed: D. Kyvig), pages 53–66, Greenwood Press, 1985

[24] L. Tribe. How to violate the Constitution without really trying: Lessons from the Repeal of Prohibition to the balanced budget amendment. *Constitutional Commentary*, 12:217–221, 1995.

[25] E. Ullman-Margalit. *The Emergence of Norms*. Oxford Univ. Press, 1977.

[26] L. W. N. van der Torre and Y.-H. Tan. Distinguishing different roles in normative reasoning. In: *Proc. Sixth ICAIL*, pages 225–232, ACM Press, 1997.

[27] H. P. Young. The Evolution of Conventions. *Econometrica*, 61(1):57–84, Janaury 1993.