Les certitudes du droit - Certainty and the Law

Sous la direction de Ejan Mackaay

On peut se procurer le présent ouvrage à:

Les Éditions Thémis
Faculté de droit, Université de Montréal
C.P. 6128, Succ. Centre-Ville
Montréal (Québec) H3C 3J7
Canada

Courriel : themis@droit.umontreal.ca
Internet : www.themis.umontreal.ca
Téléphone : (514) 343-6627
Télécopieur : (514) 343-6779

© 2004 – Éditions Thémis Inc.
Toute reproduction ou distribution interdite
Limits to Freedom of Speech : The Case of Incitement

Raphael Cohen-Almagor

Introduction

I. Incitement: Examples from the Recent Israeli Experience
   A. Stickers carrying the slogan “Rabin Should Be Killed”
   B. Rabin in black S.S. uniform
   C. Rabbi Ginsberg’s seditious pamphlet
   D. Pulsa Denurah

II. Further Thoughts

---


2 Department of Communication, University of Haifa; Fulbright-Yitzhak Rabin Scholar and Visiting Professor, UCLA School of Law (1999-2000).
On November 4th 1995 Prime Minister Yitzhak Rabin was assassinated in the main square of Tel Aviv. It was at the close of a large demonstration that had called for peace and protested against violence. Following the assassination people felt the need to ponder their own activities and statements before the assassination. Questions were raised about whether the leadership, the media and others were responsible for the atmosphere which might have been conducive to the rise of people like the assassin Yigal Amir. People who had a say in public forums utilized the media to ask themselves whether they had a share in creating a violent atmosphere that nourished murderous thoughts. Voices were raised declaring that there was “too much freedom in Israel,” too much freedom of expression, too much freedom on the part of the media.

In this essay, the issue of inciting speech is discussed, focusing attention on four examples of incitement prior to Prime Minister Rabin’s assassination that required intervention but in which insufficient measures were taken to forestall them or to punish the individuals involved. These cases occurred after the signing of the Oslo Accords in September 1993, which increased the rift between the political "left" and "right" in Israel. The first two are examples of stark political extremism. The other two are examples of incitement under religious disguise. The four cases do not exhaust all cases of incitement that took place during the heated early 1990s. Rather, they are illustrative of the incitement campaign against the Rabin government and the peace process.

I. Incitement: Examples from the Recent Israeli Experience

A. Stickers carrying the slogan “Rabin Should Be Killed”

On October 30th 1993 stickers were circulated in a small town called Or Akiva during a visit of the then Minister of Labor, Ora Namir. The stickers conveyed the following statement: “Rabin Should Be Killed.” A target was mapped, and a clear statement conveyed as to what the target’s fate should be. It was an explicit call for murder. Moreover, the social setting was such that it increased the likelihood of taking
harmful action. The stickers were distributed during a visit of a minister in Rabin’s government and there was a likelihood that one or more of the people in the public, many of whom objected to the Oslo Accords and the policies of Rabin’s government, might take measures to kill Rabin’s representative. This statement constitutes incitement that should not be protected under the Free Speech Principle.

The two who circulated the stickers, Ahuva Vaanunu and Gil Sharon, stood trial for conducting seditious actions (under section 133 of the Penal Law, 1977) and for circulating seditious publications (under sections 134a, 26 and 499 of the Penal Law, 1977). They received very lenient sentences. Judge Amiram Sharon sentenced them to three months imprisonment; six months probation and a fine of NIS 1,500 each (roughly $500). This sentence could not be regarded as a proper deterrent against those who incited the murder of Prime Minister Rabin. Instead of raising a loud voice that the courts would not tolerate explicit calls for murder, the court dismissed the issue as a mistake made by the two defendants, ignoring the context in which the stickers were circulated and the heated atmosphere that required law-and-order intervention to calm it down.

B. Rabin in black S.S. uniform

In October 1995, during a large demonstration by the Israeli political right protesting against the Oslo Accords, some extreme right wing activists associated with the Kach party4 waved photomontages of Rabin dressed in a black S.S. uniform. The Prime Minister’s face was placed over the body of the notorious Nazi leader Heinrich Himmler. The legal authorities took no steps to curtail those incitements or to prosecute those who waved the alarming pictures. In the Israeli culture it is clear what the fate of a Nazi should be. Nazis are the most vehement enemies of the Jews and, therefore, have no place within Israeli

3 Criminal file 152/94, State of Israel v. Gil Sharon and Ahuva Vaanunu, Hadera Magistrate’s Court.

4 See R. COHEN-ALMAGOR, “Vigilant Jewish Fundamentalism: From the JDL to Kach (or ‘Shalom Jews, Shalom Dogs’)”, (1992) 4 Terrorism and Political Violence, No 1, 44.
society. They should be eliminated. In this context, the difference between calling a group "Nazis," and targeting one individual by this revolting title and dressing him in a black S.S. uniform is emphasized. The legal authorities ignored this clear incitement. Only after Prime Minister Rabin’s assassination were measures taken to track down the inciters and to investigate them.

The two activists stood trial for brandishing the photomontage. Defendant 2 was also accused of writing the slogans “Rabin A Victim of Peace, Peres Is Next” and “Peres Continues the Way of Nazi Hitler.” Judge Ben-Dor noted that the two defendants had no previous criminal records and that the behavior stemmed from their ideological convictions. In his opinion, the balancing formula required withdrawal of freedom of expression when the harm to public order was severe and serious, as was the case here. The defendants’ offense severely damaged public order. The photo of Prime Minister Rabin dressed in S.S. uniform evoked outrage in every Jew. People who conceive the prime minister as a traitor, as a person whose policies might lead to the destruction of Israel as the Nazis brought about the destruction of the Jewish people, are urged by this photomontage to harm Prime Minister Rabin. Graffiti such as “Rabin A Victim of Peace, Peres Is Next” evoked similar feelings in like-minded people. Both defendants were accordingly convicted. The sentence of Defendant 1 was three months conditional imprisonment for one year and 152 hours of communal work. Defendant 2 had just opened a new business and needed to devote his time and energy to this enterprise, so the considerate judge sentenced him to three months conditional imprisonment for one year and a fine of NIS 950 (roughly $300).

With all due respect I think that these are ludicrous sentences. Incitement must be excluded from the Free Speech Principle and regarded as a criminal offense carrying severe punishment. Today there is room for legislation in Israel which would sharpen the distinction between incitement and advocacy. A meticulous analysis is required. The Penal Law must be

---

5 Criminal file 673/95, Magistrate’s Court, Jerusalem, March 17th 1996, judgment delivered by Judge Uri Ben-Dor.
amended to exclude incitement. It is important to add an article concerning calls for the death of a certain person or people, to determine a severe punishment by law for such a transgression, and to ensure its proper and serious implementation. The legal establishment has refrained from enforcing the Sedition Law because it is not specific enough. As the former government’s legal advisor, Professor Yitzhak Zamir was quoted saying that which he had repeatedly stated in private conversations, namely, that the broad definition and ambiguous wording of the Sedition Law would justify daily legal actions against newspapers which publish what could be interpreted as incitements. The fear of the slippery-slope syndrome has caused legal paralysis. It is important that the law be corrected, the sooner the better. It is further noted that article 138 of the Penal Law reads that an act, speech or publication must not be considered seditious if its purpose is to prove that the government was misled or mistaken, or to point out imperfections or flaws in the laws of the state; it also states that the means by which these observations are presented must be kosher, i.e. proper, within the law.6

I also support the amendment of the Penal Law to the effect of prohibiting the use of Nazi symbols in the Israeli political culture (possibly within the Law Prohibiting the Denial of the Holocaust, 1986). It seems that Israeli legislators did not consider passing such a law previously, assuming that Jews would refrain from using Nazi symbols for political purposes. The last months of 1995 proved them wrong. Legislation must say its words clearly: there is no room for Nazi symbols in the Israeli social arena/political culture.

It should be added that apparently not all the facts of this affair were revealed to the public. There is room to suspect that the lenient sentences were handed down also because of the involvement of the Israeli Internal Security Service (SHABAC) in the circumstances. It appears that the ideological zealots were acting under the directives of a SHABAC agent named Avishai Raviv. The SHABAC actually helped to found a terrorist organization named Eyal that was headed by Raviv. I repeat: it was not a case of installing an agent into an existing terrorist

---

6 Chapter Eight, Article One, Section 138 (1) (2) (3) of the Penal Law. Laws of the State of Israel, Special Vol.: Penal Law, 5737-1977.
organization in order to disclose its activities and warn against violent actions. Rather it was a case of founding a new terrorist organization that was extremely instrumental in generating an atmosphere of hatred and incitement that was conducive to the assassination of Prime Minister Rabin. Raviv was one of the leading figures in the radical camp that fought against any compromises for peace and for the unity of Eretz (the Land of) Israel since the mid-1980s. His activities included not only hate speech but also violent attacks on Arabs. I confess that the logic of assisting such a person, making him a leader against the foundations of Israeli democracy, and financing a terrorist organization by a prominent government agency is beyond my understanding.

A further note has to be made with regard to the role of the media in the Rabin/Himmler affair. In the Israeli culture and social context, printing photos showing Prime Minister Rabin in Nazi uniform is unethical. It is one thing to report that during a demonstration pictures of Rabin dressed in a Nazi uniform were waved, and quite another to actually print the pictures in the newspapers and by this serving the interests of the inciters. The media should not serve as a platform for spreading hatred and violence.

Indeed, Moshe Vardi, Editor of the major Israeli newspaper, Yedioth Ahronoth, applied self-censorship and refrained from printing these pictures. This is an example of applying ethical codes without the need for governmental or legal interference.

C. Rabbi Ginsberg’s seditious pamphlet

In September 1994 Rabbi Yitzhak Ginsberg published a pamphlet entitled “Baruch the Man: Five General Commandments (Mitzvot) that are Intrinsic Perspectives in the Act of Saint Rabbi Baruch Goldstein” in which he set forth Halachic (derived from Jewish law) and ideological justifications for the murder in the Cave of Machpellah (the
burial place of the Patriarchs and their wives in Hebron) in February 1994.\(^7\)

The five Mitzvot which were the impetus for Dr. Goldstein’s act, according to Rabbi Ginsberg, were revenge, removal of evil, Kiddush Ha’shem (sanctification of the Holy Name), deliverance of souls, and war. Such a pronouncement calls into question whether Judaism is compatible with humanism. It was for Attorney General Ben-Yair to examine whether this praise constituted sedition according to the Penal Law.

The Penal Law defines “sedition” *inter alia*, as:

> arousing discontent or resentment amongst inhabitants of Israel or promoting feelings of ill will and enmity among different sections of the population.\(^8\)

I am not too happy with the language of this law, which provides great latitude to restrict essential freedoms. The law should be reformulated in more restrictive terms. Nevertheless, it is argued that, on some occasions involving incitement, it is better to apply the law as it is than to convey an indulgent message to inciters that their malicious declarations may be voiced and nothing will be done to curb them.\(^9\) In my view Rabbi Ginsberg’s pamphlet arouses discontent and resentment amongst Palestinians and Israelis and it prompts feelings of ill will and enmity among different sections of the population. There was room to try him for sedition.

Moreover, Rabbi Ginsberg should have stood trial for violation of two other laws. He should have been prosecuted for

\(^{7}\) On February 25\(^{th}\) 1994, Dr. Baruch Goldstein entered the Cave of Machpellah and massacred in cold blood some twenty-nine Palestinians praying in the mosque inside the Cave.

\(^{8}\) Chapter Eight, Article One, Section 136 (3) (4) of the Penal Law. Laws of the State of Israel, Special Vol.: Penal Law, 5737-1977.

\(^{9}\) For further discussion see R. COHEN-ALMAGOR, “Combating Right-Wing Political Extremism in Israel: Critical Appraisal” (1997) *Terrorism and Political Violence* N° 4, 82.
violation of Section 4 of the Prevention of Terrorism Ordinance, 1948, which says:

\[
\text{a person publishing praise, sympathy or encouragement for acts of violence calculated to cause death or injury, and a person assisting the organization in its activities,}
\]

\[
is subject to criminal proceedings and a maximum penalty of three years' imprisonment and/or a fine of £1,000.
\]

Alternatively or additionally Rabbi Ginsberg should have been charged for “incitement to racism” under sections 144 (A-E) of the Penal Code. In August 1986, in its fighting against the Kach movement established by Meir Kahane,\textsuperscript{10} the Knesset passed a law that specifies “incitement to racism” as a criminal offense.

\[
\text{Anyone who publishes anything with the purpose of inciting to racism is liable to five years imprisonment (144B);}
\]

and:

\[
\text{anyone who has racist publications in his or her possession for distribution is liable to imprisonment for one year (144D).}
\]

The term “racism” is defined as:

\[
\text{persecution, humiliation, degradation, manifestation of enmity, hostility or violence, or causing strife toward a group of people or segments of the population – because of color or affiliation with a race or a national-ethnic origin (144A).}\textsuperscript{11}
\]

\textsuperscript{10} For discussion on Meir Kahane’s ideology see his most comprehensive book Uncomfortable Questions for Comfortable Jews, Secaucus, N.J., Lyle Stuart, 1987.

The reading of Justice Matza’s recent judgment in *Rabbi Ido Elba v. State of Israel* leads me to infer that today, after Prime Minister Rabin’s assassination, Rabbi Ginsberg might have been charged for inciting to racism.

However, at that time Rabbi Ginsberg did not stand trial neither for incitement to racism nor for sedition, nor for contravention of the Prevention of Terrorism Ordinance. Only after the abominable assassination did the authorities take action against him. On March 10th 1996 Rabbi Ginsberg was put under administrative detention, one of the most anti-democratic measures in the legal framework of Israel, for a period of two months. The grounds for his detention order were classes in which Rabbi Ginsberg told his students that there was a *Halachic* duty to take revenge against Arabs for the massacres conducted by the *Hamas* and the *Islamic Jihad* in Jerusalem, Ashkelon and Tel-Aviv.

Rabbi Ginsberg appealed to the Supreme Court against the detention decision. His main contentions were that nothing in what he said could serve as basis for the assumption that a probable connection existed between his statements and harm inflicted upon Arabs by his students. Rabbi Ginsberg maintained that his views were not one sided. In support he brought evidence showing that in one of his publications he said that:

> it is forbidden to harm a non-Jew who is not at war with us.  

The state representative argued in response that Rabbi Ginsberg exercised strong influence on his followers and that his preaching to take revenge on Arabs established grounds to suspect that the students might act upon their Rabbi’s instructions.

---


14 *Id.*, paragraph 3.
Justice Dalia Dorner accepted Rabbi Ginsberg’s appeal. She explained that there was scope for administrative detention when standard measures were deemed insufficient to secure public peace. Thus, when it was impossible to issue an indictment because the evidence, though reliable, was inadmissible and a near probability existed that forbearance from detention might bring about substantial harm to public and state security, then detention was justified. In the present matter it was not claimed that Rabbi Ginsberg might do things that would endanger public security. Rather, Rabbi Ginsberg was arrested out of fear that his pronunciations might prompt his students to harmful conduct. Moreover, Rabbi Ginsberg lectured frequently to hundreds of people and published his views on paper.

Justice Dorner mentioned that in one of his publications, Rabbi Ginsberg regarded Baruch Goldstein’s massacre at the Cave of Machpellyah as Kiddush Ha’shem. Unfortunately she refrained from voicing an opinion as to whether this writing contravened Israeli law. It was nevertheless obvious that Rabbi Ginsberg’s lectures and publications exhibited no lack of clear evidence and material, so there was no need to resort to the exceptional measure of administrative detention.

I agree with Justice Dorner that there was no room to place Rabbi Ginsberg under administrative detention. I have strong reservations with regard to the employment of this measure in democratic societies. In another article I wrote that the procedure of administrative detention is manifestly unjust for it lacks proper hearing and due process of law. It is contrary to the democratic spirit and to liberal reason that proscribes arbitrary arrests. This procedure is commonplace in authoritarian regimes. It is the kind of instrument despots use to suppress opposition. They see no obligation to insist on rules of evidence and to disclose information to individuals under arrest. In contrast, democracies require that all legal procedures be exhausted before putting individuals behind bars. In a court of law, the prosecution has to prove that criminal offenses have been committed which justify penalties. Defendants have the right to be represented by lawyers, to summon witnesses and to cross-examine them. The administrative detention procedure eschews this and, therefore, is contrary to the notion of doing justice. Thus, my contention is the following: let the prosecution prosecute, the defendants defend themselves and the court of
justice meet out justice in accordance with material evidence. And if there is not sufficient evidence to prosecute, or if the prosecution is unable to produce relevant material, the defendants should retain their freedom. No procedure should exist to override the administration of justice.\textsuperscript{15}

While agreeing that detention should not be considered just another preventive measure to be selected from the arsenal of preventive measures and that it should not serve as a tranquilizer or as a substitute for criminal proceedings, nevertheless there was room to file criminal charges against Rabbi Ginsberg for his inciting statements. It is reiterated that he should have stood trial for inciting to racism, for provoking acts of terror and for sedition.

One more observation regarding the Penal Law is in place. After Prime Minister Rabin’s assassination, the Minister of Justice David Libai and Attorney General Ben-Yair recommended that the Penal Law dealing with seditious conduct be refined and defined more clearly. Professor Libai asked Professor Mordechai Kremnitzer of the Hebrew University Law Faculty to prepare a draft proposal for a specific incitement law.\textsuperscript{16} However, at some later point this initiative was abandoned. I am sorry to say that for political reasons the Ministry of Justice never implemented its own recommendation.

\textbf{D. \textit{Pulsa Denurah}}

The last example of incitement concerns a religious curse called \textit{Pulsa Denurah}. In October 1995, on the eve of the most sacred day in the Jewish calendar, \textit{Yom Kippur}, a person named Avigdor Askin, together with some other people,

\begin{itemize}
\item \textsuperscript{16} I thank Mota Kremnitzer for providing me with the text of the draft proposal.
\end{itemize}
distributed this curse which was composed by three Cabbalists (Mekubalim) against Prime Minister Rabin. Mr. Askin was photographed during the recitation of the Pulsa Denurah prayer outside the Prime Minister’s official residence in Jerusalem. The prayer called on Rabin to cease his wrongful deeds in this world; it was recited in the presence of media reporters who were invited to the scene to publicize the ceremony and to deliver an inciting message to the public. The message was that Prime Minister Rabin could not escape the death curse that was placed upon him because of his evil policies. In effect, Rabin’s blood was allowed. This was a provocative measure calling for his death.

The legal authorities took no action against Mr. Askin. Only after the assassination, when Askin appeared on television and declared that “our prayer was fulfilled in full” did the authorities begin to look for him. Liberals may dismiss the entire story as ridiculous, saying Pulsa Shmulsa. But Liberals are not prone to believe in such curses. They will not be moved to help God in executing such wishes. This prayer constituted an incitement that fell on eager ears and helped to generate an atmosphere that was conducive to triggering Yigal Amir and encouraged him to carry out his heinous act.

In early March 1996 Palestinian terrorists launched a series of vicious attacks which caused the death of tens of civilians. Following those massacres, on March 6th, Mr. Askin approached the media and announced that it was his intention to perform the Pulsa Denurah ceremony once again, this time against Prime Minister Shimon Peres. After all, the curse proved very effective the first time, so why not give it a second shot. Once again, the media served as a good mobilizer of his intentions. On March 7th the two popular daily newspapers, Yedioth Ahronoth and Maariv, published Askin’s contentions. In doing so they provided a platform for incitement.

Askin stood trial for (1) performing the Pulsa Denurah ceremony and for noting the connection between the ceremony and the death of Prime Minister Rabin the day after he was assassinated and also (2) for declaring that he intended to perform a similar ceremony calling for the death of Mr. Rabin’s successor, Shimon Peres. The Court found Mr. Askin guilty on both accounts for violating Section 4 of the Prevention of Terrorism Ordinance, 1948. His verdict was imprisonment for a
period of four months and an additional one year conditional sentence for a period of three years. Askin appealed against the decision and as of November 1998, the appeal was still pending and Askin was still free, continuing to spread hatred and incitement against those conceived by him as “Israel’s enemies.” Yedioth Ahronoth informed that a survey conducted among Russian immigrants reveals that Mr. Askin was regarded by this large sector as the fourth most prominent political personality in Israel in 1997. There are speculations that Askin had also some connections with the SHABAC.

II. Further Thoughts

Prime Minister Rabin’s assassination forced us to think harder than before about the limits of liberty and tolerance in our democracy. Israel is a young democracy. It is in process of development and undoubtedly it will face further challenges and tests. I hope these tests will not be of the nature and scope of the tragic murder of the 4th of November 1995. On the whole, Israeli democracy coped quite well with the challenge imposed by Prime Minister Rabin’s assassination. Immediately after the assassination people feared that we might loose our brakes and that illiberal measures would be introduced that might hinder the Israeli nation-building tradition as a democratic state. I am happy to say that those fears were too pessimistic. Nevertheless, we must acknowledge that the assassination opened up new frontiers of political radicalism, and that ample safeguards should be installed to protect our vulnerable reality. We live in an era of political violence and extremism and we need to find answers to the radical forces that seem to go from strength to strength and to overcome them. We need to declare that incitement is well outside the boundaries of tolerance. We need to adopt legal measures to exclude it from the protection of the Free Speech Principle and not hesitate to prosecute people who call for murderous attacks on others. Unfortunately, nowadays we hear constant threats against high-ranking officials whose conduct runs counter to certain beliefs of extremists. It seems

---

17 Criminal file 827/96. State of Israel v. Avigdor Askin, Jerusalem Magistrate’s Court.

that, within the radical spheres, all boundaries are broken. The assassination of Prime Minister Rabin legitimized a new mode of conduct. The terms "political assassination" and "liquidation lists" became part of this place. I do not recall hearing those expressions prior to Prime Minister Rabin’s assassination with the frequency and intensity that we hear them now. Even in the heyday of the quasifascist Meir Kahane, during the mid-1980s, people were much more careful in expressing themselves. It is our duty to curb this frightful tendency. We must take stringent action now if we do not want to face yet another assassination. The legislature, the police, the courts and the media should take measures to exclude certain modes of speech from our society. Furthermore, we need to fight down all forms of terrorism, whether directed against Jews or against Arabs. Terrorism and democracy cannot live together. One must make way and advance at the expense of the other. It is our common interest to work for the victory of democracy. We also need to build bridges and promote understanding between different factions of the population, especially between the secular and the religious factions. Terminology such as “we are enlightened liberals and they constitute the forces of darkness” which is often utilized by Israeli civil libertarians will not help the forces of democracy. There are enlightened individuals within the religious circles just as there are intolerant individuals within the secular circles. Israel, as a religious and democratic state, needs to work out ways to bring about the good of both traditions, and to enrich the citizens’ understanding of both great forces that made Israel the state it now is.
Le juge, arbitre de la certitude du droit

François Rigaux*

Introduction........................................................................................................ 19

I. Vie, mort et résurrection de la jurisprudence conceptuelle
   (Begriffsjurisprudenz)................................................................. 20
   A. Freirechtslehre c. Begriffsjurisprudenz ........... 20
   B. Le courant réaliste américain......................... 24

II. L’arbitrage du juge................................................................. 28
   A. Le conflit de normes ................................................. 28
   B. Les choix successifs de la jurisprudence civile........................................................................ 30
   C. Argument d’analogie et évaluation des précédents .................................................................. 32
   D. La division des juges.................................................. 35

III. Doute méthodique et incertitude ...................... 37
   A. La notion d’acte clair ................................................. 38
   B. La doctrine du «constitutional doubt»............. 41
   C. Les voies du contrôle de légalité sur l’interprétation des actes juridiques privés.................. 45
   D. Univocité ou ambiguïté des textes soumis au juge .......................................................... 46

* Professeur émérite de l’Université catholique de Louvain.
Conclusion : jugez sur ce que je fais plutôt que sur ce que j’en dis

A. L’administration des questions préjudicielles d’interprétation par la Cour de justice des Communautés européennes

B. Canons d’interprétation et recours aux travaux préparatoires
Selon l’épithète dont il est accompagné, le mot «certitude» a deux significations. Après l’avoir défini comme «la qualité de ce qui est certain», Littré donne les deux usages suivants : «Certitude morale, mathématique». Il faut opérer une disjonction entre ces deux variétés de certitude. Alors qu’une certitude mathématique se place à l’abri de tout doute, il se glisse dans la certitude morale plus qu’une nuance d’hésitation. Impliquant l’adhésion du sujet, elle est empreinte de subjectivisme; ce qui est certain pour Pierre ne l’est pas pour Paul, pareil relativisme caractérisant le modèle de certitude que le juge confère au droit. Mais notons d’abord l’érosion quasi générale du vocabulaire. Une expression voisine, «sans doute», a perdu en français son sens originel, il faut y suppléer par «hors de doute» ou «assurément» si l’on veut évincer l’idée de simple probabilité qui est devenue le sens le plus usuel aujourd’hui de «sans doute».

À l’incertitude du droit correspond plutôt, selon le langage des Réalistes américains, la notion d’indétermination (indeterminacy). En mathématiques, on tient pour indéterminé un problème qui a un nombre indéterminé de solutions et ici on se trouve très proche de l’incertitude ou de l’indétermination du droit. À la vérité, l’expression est elle-même ambiguë et, si l’on ose ce jeu de mots, incertaine. Pour qui, en faveur de qui, dans quel contexte le droit se pare-t-il de la vertu de certitude? La certitude est-elle un idéal à poursuivre? Par qui, par quels


2 Selon C. L. KUTZ, loc. cit., note 1, 1001, il s’agit plutôt d’underdeterminacy que d’indeterminacy : the legal system permits the logical derivation of more than one conclusion from any given set of premises.
moyens? S’agit-il d’une question de sémantique, c’est-à-dire d’interprétation, ou y va-t-il de l’application de la règle à une situation litigieuse? Mais voici qu’apparaît la figure emblématique du juge. Le droit est-il un ensemble de règles ou de normes ou faut-il se laisser guider par des principes, des directives (guidelines), des « standards»?3 Le mot « principe » met sur la voie d’un autre glissement sémantique : l’expression « en principe » signifie que la règle ainsi modalisée tolère des exceptions, c’est-à-dire qu’un principe est un peu moins fort, moins certain, moins déterminé qu’une règle.

Pour tenter de répondre à quelques-unes de ces questions, d’ailleurs liées entre elles, il est proposé de développer les points suivants. D’abord, le contraste entre l’âge d’or de la certitude et l’âge de fer, qui est le nôtre, où l’on s’efforce laborieusement de restituer au droit la certitude qu’il aurait perdue. Le second point a pour objet le rôle du juge dans cette tentative, ce qui conduit aussitôt à la relation que les juges entretiennent avec la loi et avec les précédents judiciaires. Mais de tous les côtés c’est le conflit qui domine : multiplicité des lois dont les aires respectives ne sont qu’imparfaitement délimitées, diversité des précédents, irréductibilité du cas litigieux à aucun de ceux auxquels il se laisse comparer, division du siège dans les juridictions collégiales. Dans un troisième et dernier point, il faudra s’efforcer de démêler les fils d’un écheveau passablement embrouillé : comment dépasser l’incertitude des méthodes?

I. Vie, mort et résurrection de la jurisprudence conceptuelle (Begriffsjurisprudenz)

A. Freirechtslehre c. Begriffsjurisprudenz

Ce sont les hérésies qui mettent en relief les traits fondamentaux de la religion dont elles ne s’écartent que partiellement. De même, le positivisme dogmatique – aux États-

Unis on dira le formalisme – de la jurisprudence du XIXe siècle n’a été perçu que par la voix de ceux qui l’ont critiqué. C’est à peu près à la même époque qu’en Allemagne et en Autriche l’École de la libre recherche scientifique (Freirechtslehre) et aux États-Unis les premiers Réalistes dénoncent l’aveuglement de ceux qui pratiquaient une «jurisprudence conceptuelle» (Begriffsjurisprudenz). Mais, par définition, un aveugle ne saurait voir et les juristes dogmatiques n’auraient pu se percevoir tels que d’autres les ont vus. Mais que signifiait «jurisprudence conceptuelle»? Deux choses, semble-t-il. D’abord, que le droit est fait d’un ensemble de propositions hiérarchisées et logiquement liées entre elles. Ensuite, que toute situation particulière doit pouvoir être appréhendée (subsumiert, subsumée) par une des normes appartenant à l’ordonnancement juridique dont la pureté cristalline épargne toute incertitude quant à la nature et à l’étendue des droits et devoirs respectifs des parties. Énoncé en ces termes, le modèle frise la caricature, mais on ne saurait assez le répéter, ce sont toujours les autres qui parlent.

Durant l’âge d’or de la certitude du droit, celle-ci est faite de logique et, l’on peut ajouter, d’une logique binaire dominée par le principe de non-contradiction et du tiers-exclu. Le juge choisit entre le blanc et le noir, le oui et le non, sur le modèle du récit de la Genèse où Dieu a séparé les eaux de la terre ferme, divisé le temps entre le jour et la nuit, créé l’homme et la femme. On en trouve une expression nostalgique dans une opinion convergente du juge Clarence Thomas4. Le schéma binaire de l’outil informatique est, lui aussi, une nouvelle émergence du modèle ancestral.

Si contestable que puisse paraître l’utilisation de catégories trop appuyées, les conquêtes foudroyantes de la

---

4 À propos d’un problème qui ne méritait sans doute pas une telle envoûtement (il s’agissait de la répartition des compétences entre les juridictions fédérales et celle des États à propos d’un accident survenu dans une voie navigable), le juge Clarence Thomas écrit qu’il existe une règle déterminant la compétence:

«almost as clear as the 9th and 10th verses of Genesis : — And God said, let the waters under heaven be gathered together unto one place, and let the dry land appear : and it was so. And God called the dry land Earth; and the gathering together of the waters called he Seas : and God saw that it was good. — The Holy Bible, Genesis 1:9-10 (King James version) : Grubart v. Great Lakes Dredge and Dock Cy, 130 L Ed 2d 1024, 1045 (1995).»
sciences au XIXe siècle et le positivisme juridique vont de pair. À l’âge d’or de la certitude a succédé l’ère du soupçon ou, si l’on préfère, de l’incertitude. Le droit n’y a pas échappé et il a, l’un des premiers, abordé l’époque postmoderne, qui désigne, précisément, la fin des illusions. De ce côté-ci de l’Atlantique, c’est un article du Juge Oliver Wendell Holmes, Jr., qui a amorcé un renversement de tendance. La définition qu’il y donne du droit est relativiste et elle fait la place belle aux juges :

*The object of our study, then, is prediction, the prediction of the incidence of the public force through the instrumentality of the courts.*

Il y revient plus loin :

*The prophecies of what the courts will do in fact and nothing more pretentious, are what I mean by the law.*

L’Europe continentale n’est pas en reste, une pléiade de jeunes juristes menant le combat contre les Pandectistes et la jurisprudence conceptuelle. Les plus notables sont Eugen Ehrlich, qui enseigne à cette époque à Cernowitz (aujourd’hui Tchernovtsy) en Bucovine, province orientale de l’Empire austro-hongrois, et Hermann Kantorowicz. Les titres de leurs écrits sont significatifs. Du premier paraît, en 1903, *Freie Rechtsfindung und freie Rechtswissenschaft*. Le second publie quelques années plus tard, sous le pseudonyme de Gnaeus

---


6 *Id*. 461. Douze ans plus tard, rédigeant l’opinion de la Cour suprême dans *American Banana Cy v. United Fruit Cy*, il paraphrase un passage de l’article cité à la note précédente : *Law is a statement of the circumstances in which the public force will be brought to bear upon men through the courts. But the word commonly is conferred to such prophecies as threats ...* : 213 US 347, 356 et 357 (1909).

Flavius, Der Kampf um die Rechtswissenschaft. Face aux contradictions et aux lacunes qu’ils relèvent dans le système juridique, ils prônent sous le nom d’« Ecole du droit libre » (Freirechtslehre) une méthode qui libère le juge du carcan – sans doute plus imaginaire que réel – du droit écrit, du droit posé, dans lesquels le juge trouverait une solution prédéterminée, selon l’image que propose beaucoup plus tard Josef Esser :

L’interprétation judiciaire reçoit la forme mythique de l’éveil d’un sens sommeillant dans le texte à la manière de la Belle au bois dormant, sens qui serait donné depuis toujours.9

La force obligatoire du droit écrit est, toutefois, plus apparente que réelle. Les juges n’avaient pas attendu ni Holmes ni Ehrlich ni Kantorowicz pour se libérer, quand il était nécessaire, de la pesanteur législative, mais selon une méthode qui caractérise la fonction même de juger, ils le faisaient sans le


dire, ou plus exactement ils faisaient le contraire de ce qu’ils déclaraient faire\textsuperscript{10}.

B. **Le courant réaliste américain**

Aux États-Unis, le juge Holmes ouvrit la voie à ce qui devint une forme d’école qualifiée de réaliste bien qu’elle eût constitué un groupe plutôt disparate de personnalités indépendantes\textsuperscript{11}. Il est symptomatique que la common law et la doctrine des précédents avaient suscité une frustration au moins égale à celle qui s’était manifestée dans les pays de droit écrit et, même, de codification. Non moins significative apparaît l’époque à laquelle la *Freirechtslehre* se manifeste en Allemagne, elle coïncide avec l’entrée en vigueur, le 1\textsuperscript{er} janvier 1900, du BGB, oeuvre combien plus dogmatique et sûre d’elles-même que la codification napoléonienne à laquelle avaient contribué des survivants de l’Ancien Régime\textsuperscript{12}, praticiens peu imbus de théorie.

\textsuperscript{10} Kantorowicz le remarque lui-même à propos de la Cour de cassation de France (*der nach seiner regelmässig gepredigten, aber ebenso regelmässig unbefolgt gelassenen Auffassung*): «Aus der Vorgeschichte der *Freirechtslehre*», Mannheim, Bennsteiner, 1925, reproduit dans *Rechtswissenschaft und Soziologie*, Karlsruhe, Verlag C.F. Müller, 1962, p. 63.


\textsuperscript{12} Le BGB a été, dès son entrée en vigueur, une des cibles préférées de la *Freirechtslehre*. Voir par exemple : EHRlich, *Grundlegung der Soziologie des Rechts*, op. cit., note 7, pp. 410 et 411; *Die juristische Logik*, op. cit., note 7, pp. 165-179. La critique est double. La codification entérine le droit du passé au lieu d’ouvrir les voies de l’avenir et même du présent et il induit le praticien à une méthode d’interprétation conceptuelle (*Begriffsjurisprudenz*). Voir aussi : Ernst FUCHS, *Was will die Freirechtsschule?* Im Greifenverlag
Le legs des Réalistes américains a été pieusement recueilli par le mouvement des Critical Legal Studies. L’indétermination – ou l’incertitude – du droit est l’un des fondements de leur doctrine, mais ils s’écartent sans doutes des Réalistes par la moindre confiance qu’ils font au juge pour dissiper pareille incertitude. Ici apparaît aussi l’ambiguïté de la notion d’indétermination. S’il s’agit, en effet, d’appliquer une règle de droit – ou, le cas échéant, un standard – à une situation particulière, seul le juge est à même de le faire, et la question se ramène alors au choix de la méthode appropriée. Selon la Begriffsjurisprudenz, l’opération se réduirait à une subsomption purement logique. Mais cela suppose aussi que le sens de la


norme soit lui-même certain et ainsi le problème de la certitude du droit s’identifie avec une question d’interprétation.

Deux arguments peuvent être avancés à l’appui des méthodes nouvelles. Le premier a pour objet le caractère inéluctable de l’interprétation contextuelle. Aucun concept n’a de sens que référé à un contexte déterminé. En d’autres termes, la signification de la loi, du standard, du précédent ne va se laisser déceler que dans le contexte qui en fixe les contours. Comme tout autre phénomène de langage, le droit n’est ni absolument certain ni radicalement incertain, il n’acquiert une certitude (relative) que dans le contexte où il s’insère. Or, cette opération est l’œuvre du juge, ce qui conduit au second argument; il n’y a pas d’interprétation affranchie de l’application et les circonstances dans lesquelles une situation particulière s’offre au juge aident celui-ci à choisir et à interpréter le droit qui y est applicable.

Les données du cas d’espèce révèlent ainsi les conflits de normes que le juge est seul en mesure de trancher. Les règles protectrices des droits fondamentaux, telles qu’elles sont inscrites dans une constitution écrite ou dans la Convention de sauvegarde des droits de l’homme et des libertés fondamentales, sont énoncées en termes généraux et elles couvrent des domaines qui se recouvrent partiellement. Stanley Fish a écrit, il y a quelques années : «There’s no such thing as free speech and it’s a good thing, too...»15. Le juge de la constitutionnalité n’appréhende la liberté d’expression que par les exceptions qui y sont apportées et dont il doit décider si elles sont tolérables ou non. Les conflits entre la liberté d’expression et la protection de l’honneur et de la vie privée sont bien connus, mais on peut aussi évoquer le conflit entre cette liberté et la liberté d’association16 ou le conflit entre la liberté d’expression et le


droit de propriété. Le propriétaire d’un lieu ouvert au public, telles la société qui gère le bâtiment d’accès à un aéroport\textsuperscript{17}, celle qui a construit une ville\textsuperscript{18}, gère un espace commercial dont les voies d’accès ne se laissent pas distinguer d’une rue ouverte aux piétons\textsuperscript{19}, peut-il interdire ou réglementer la diffusion de dépliants publicitaires ou les sollicitations adressées au public\textsuperscript{20}? Alors que dans \textit{Krishna}, la liberté d’expression était renforcée par la liberté religieuse, dans une autre affaire jugée la même année, la Cour suprême des États-Unis a décidé que l’interdiction de diffuser une information sur le parking d’une entreprise attenant au bâtiment principal pouvait s’étendre aux informations données aux travailleurs par une organisation syndicale\textsuperscript{21}.

Aucun législateur ne saurait, à l’avance, déterminer les réponses à de telles questions. Seul le juge est à même de le faire et, après qu’il s’est prononcé sur un cas d’espèce déterminé, toute incertitude n’est pas levée à l’égard d’hypothèses futures différentes.


Si éloignée qu’elle puisse paraître d’un système juridique codifié ou législatif, la culture du précédent ne suscite pas des problèmes radicalement différents. Au même titre que les lois, les précédents sont multiples, il faut choisir le plus pertinent ou concilier ceux qui envoient des messages contradictoires. Il faut, dans les deux cas, en interpréter le contenu. La différence la plus notable a pour objet la liberté de certains juges d’écarter (overrule) un précédent qui a cessé de paraître satisfaisant, encore que les tenants les plus radicaux de la Freirechtslehre n’hésitent pas à encourager le juge à prendre des libertés avec la loi elle-même.

II. L’arbitrage du juge

A. Le conflit de normes

La théorie du droit s’est, de longue date, intéressée au problème des lacunes. Il s’agit, toutefois, d’un faux problème. Il n’existe que durant une phase éphémère de l’élaboration du dispositif juridique : au terme de celle-ci, il appartient au juge de se prononcer. À cela s’ajoute que la lacune n’est souvent qu’apparente, avancée par un plaideur astucieux ou issue d’une connaissance insuffisante des sources. À la lacune, présentée comme un conflit négatif, paraît s’opposer le conflit positif : la situation est appréhendée par de multiples sources de droit. La dogmatique traditionnelle a construit une série de règles ou de « canons » : principe de hiérarchie (la norme constitutionnelle prime une loi ordinaire, le droit fédéral brise le droit de l’État, Bundesrecht bricht Landesrecht), Specialia generalibus derogant, Lex posterior ... Il s’agit, cependant de sécurités illusoires, traversées par d’autres canons d’interprétation. On ne peut prêsumer que le législateur ordinaire ait voulu transgresser un précepte constitutionnel ou une norme internationale. Aux deux normes dont le conflit n’est qu’apparemment régi par le principe de hiérarchie, le juge va donner une interprétation accommodante, mais l’assouplissement ne vaut pas moins pour la norme de rang supérieur que pour celle qui y est constitutionnellement subordonnée.

La jurisprudence de la Cour suprême des États-Unis à propos du premier amendement est la plus significative à cet égard. En dépit des termes catégoriques de la première phrase
du premier amendement (Congress shall make no law [...] abridging the freedom of speech, or of the press...), la Cour suprême n’a pas hésité à assortir la règle d’une quantité d’exceptions22 et de modaliser celles-ci selon la nature du discours protégé23 ou le moyen de communication utilisé24. De la même manière, à propos de la liberté de l’art et de la science qui, selon les termes de l’article 5 de la Loi fondamentale allemande est, à la différence de la liberté d’expression, soustraite à toute restriction ou immixtion législatives, le Tribunal constitutionnel fédéral a décidé que la liberté de l’art est garantie «sans réserve» (vorbehaltlos) mais «non sans


23 La jurisprudence a tracé une échelle graduée allant de la discussion de questions d’intérêt général (qui contribue davantage à la formation de l’opinion publique), à la défense d’intérêts économiques ou professionnels, à la publicité commerciale, aux publications de divertissement, à celles qui sont indécentes sans franchir la barrière impermissible de l’obsénité. Mais qui établira la différence entre la «simple» indécence et l’obsénité ?

limites» *(nicht schrankenlos)*25. L’art d’accommoder la constitution fut porté à son suprême degré durant la Seconde Guerre mondiale quand les Américains de lignée *(ancestry)* japonaise furent déplacés de la côte du Pacifique où ils résidaient régulièrement et transférés dans des camps de concentration *(relocation camps)*. La Cour suprême a justifié de telles mesures par les nécessités de la guerre bien que, selon l’opinion dissidente du juge Murphy, elles fussent critiquables pour deux motifs au moins : l’utilisation de «profils», tous les Américains de lignée japonaise étant présumés suspects, et la discrimination raciale dont ils furent victimes, les Américains d’origine allemande ou d’origine italienne ayant, durant la même période, été soumis à un contrôle sélectif limité à ceux dont la loyauté paraissait douteuse26.

**B. Les choix successifs de la jurisprudence civile**

Non moins que la jurisprudence constitutionnelle, l’application des lois civiles donne matière à des choix successifs des cours et tribunaux. Quand la victime d’un accident mortel imputable à un tiers vivait en concubinage, la jurisprudence française a d’abord hésité à mettre à charge de l’auteur responsable l’obligation de réparer le préjudice causé à la survivante. Il fallait admettre que l’application de l’article 1382 du Code civil n’était pas limitée à la lésion d’un droit mais qu’elle s’étendait à l’atteinte portée à un simple intérêt, à condition que celui-ci fût juridiquement protégé. Or l’objection dirigée contre l’action de la concubine était le caractère immoral et, partant, illicite, des rapports sexuels en dehors du mariage. En France, l’objection fut définitivement levée en 1970 par un arrêt de la Chambre mixte de la Cour de cassation27. Encore ne s’agissait-il que d’un concubinage simple. À quoi pouvait prétendre la femme qui, étant engagée dans les liens d’un mariage, demandait réparation pour le décès d’un autre homme avec lequel elle s’était mise en ménage? La concubine adulte

---


aurait-elle les mêmes droits que la concubine simple? La transgression permanente et délibérée des devoirs de fidélité et de cohabitation ne rendrait-elle pas illégitime l’intérêt dont la demanderesse se prévalait? La motivation des deux premiers arrêts de la chambre criminelle qui accueillent l’action de la concubine adulte est prudente et paraît dépendre des particularités du cas d’espèce. En 1972, un arrêt de rejet concède que le caractère illicite du concubinage dans lequel est engagée une handicapée physique abandonnée par son mari depuis 36 ans «n’est pas certain»28. L’année suivante la même chambre casse un arrêt de la cour d’appel de Bordeaux qui a refusé le droit d’action à une concubine séparée de son mari en 1937, au moment de la guerre civile espagnole, et qui n’en a plus reçu de nouvelles depuis : l’existence du conjoint étant devenue «incertaine», la «base légale» de la décision disparaît29. On notera comment dans les deux cas le regard d’incertitude jeté par le juge sur le fait conduit à une nouvelle certitude du droit. Ce n’est qu’en 1975 que la Chambre criminelle renonce à pareil subterfuge en adoptant une motivation juridique péremptoire. L’arrêt d’appel qui refuse le droit d’action parce que l’adultère est réprimé par le Code pénal est cassé : il s’agit, en effet, d’un délit dont la poursuite est subordonnée à la plainte de l’époux offensé et dont l’auteur de l’homicide involontaire est d’autant moins fondé à se prévaloir qu’il s’agit «d’un état de fait touchant à la vie privée de la partie adverse»30. Mais on notera aussi la connivence entre le juge et le législateur : moins d’un mois plus tard, la loi n° 75-617 du 11 juillet 1975 portant réforme du divorce supprime l’adultère comme cause spécifique de divorce et abolit la répression pénale qui y était attachée.

On a parfois rapproché la concubine adulte de la victime en situation irrégulière. Le soldat qui a fait le mur de la caserne, le prisonnier en cavale sont blessés par un chauffeur en état d’ivresse qui les a renversés sur un trottoir. Celui qui conduit une voiture sans permis, mais n’a commis aucune autre faute, celui qui au temps des Sunday laws de certains États

américains, chassait, ou, en dehors de circonstances exceptionnelles, voyageait le dimanche, sont victimes d’un accident qui ne serait pas survenu s’ils n’avaient pas eux-mêmes transgressé une obligation légale ou réglementaire. Toutefois, ce que demande la victime en situation irrégulière est d’être traitée comme ceux qui respectent la loi, sans que l’auteur du fait dommageable soit autorisé à se prévaloir d’un fait qui, tout en jouant un rôle déterminant dans un modèle mécaniste du rapport de causalité, laisse intacte la responsabilité morale de l’auteur du fait dommageable. À la différence de la victime en situation irrégulière, la concubine adultère se prévaut du fait même que d’aucuns tiennent pour illégitime. Loin de gommer ce fait elle entend obtenir qu’il figure parmi les éléments, qu’il soit même l’élément central de la relation causale prévue par l’article 1382 du Code civil.

C. Argument d’analogie et évaluation des précédents

L’évolution de la jurisprudence française sur le droit d’action de la concubine illustre à la fois le caractère relatif de la notion de cas difficile (hard case) et l’évaluation des précédents. Tant qu’aucun arrêt de la Cour de cassation n’a statué sur une telle action introduite par une concubine simple, la demande peut être rangée parmi les cas difficiles. La situation change après un tel arrêt mais la catégorie des cas difficiles accueille alors la demande de la concubine adultère. Cette situation peut être réglée par une solution d’espèce limitée aux circonstances particulières d’un adulte dont le caractère illicite ou la réalité ne sont pas certains, ou, ce qui implique un pas de plus, par une décision définitivement favorable à toutes les concubines adultères. Le développement de l’assurance de la responsabilité civile a, certes, favorisé pareille extension des obligations mises à charge de l’auteur d’un fait dommageable. Si celui-ci devait s’acquitter sur ses propres biens, on pourrait concevoir un conflit entre la faute, supposons-le légère, d’un homme marié et devant subvenir à l’entretien d’une famille nombreuse et l’intérêt de la victime de cette faute à laquelle il devrait garantir une pension de survie.

Quant à la doctrine du précédent, elle n’épargne pas au juge le devoir de lever lui-même les incertitudes du droit. Il s’agit d’une variété de l’argument d’analogie, mais il est bien
connu que celui-ci implique une comparaison dont l’issue n’est pas certaine : entre l’hypothèse explicitement réglée par la loi et le cas déjà jugé ou celui qui l’est actuellement, les similitudes l’emportent-elles sur les différences? La concubine adultère peut-elle être assimilée à une concubine simple? Et la question qui se pose aujourd’hui avec une acuité particulière en France est la suivante : la vie commune de personnes de même sexe équivaut-elle au concubinage protégé par la loi ou de nature à s’inscrire dans la suite des précédents? Un arrêt récent de la troisième chambre civile de la Cour de cassation de France a donné à cette question une réponse négative pour l’interprétation de l’article 14 de la loi du 6 juillet 1989, qui dispose que «lors du décès du locataire, le contrat de location est transféré (...) au concubin notoire (...) qui vivait avec lui depuis au moins un an à la date du décès».

Pour rejeter le pourvoi, la Cour de cassation se borne à énoncer :

Qu’ayant retenu, à bon droit, que le concubinage ne pouvait résulter que d’une relation stable et continue ayant l’apparence du mariage, entre un homme et une femme, la cour d’appel n’a ni violé l’article 26 du Pacte international relatif aux droits civils et politiques, ni l’article 8, alinéa 1er, de la Convention européenne de sauvegarde des droits de l’homme et des libertés fondamentales.31

Pour l’interprétation d’une disposition particulière de la loi, la Cour de cassation a donc écarté l’argument d’analogie et il est possible qu’elle ferait de même s’il elle était invitée à étendre sa jurisprudence sur le droit d’action de la concubine à un partenaire de même sexe. Toutefois, certains juges du fond ont appliqué l’article 1382 du Code civil au partenaire du même sexe de la victime d’un accident32.


L’arrêt du 17 décembre 1997 invite aussi à s’interroger sur la fonction du juge pour définir les termes non définis par le législateur lui-même. Les rédacteurs de la loi du 6 juillet 1989 auraient pu dire ce qu’ils entendaient par «concubin notoire» et préciser si cette qualité était limitée à une relation «ayant l’apparence du mariage, entre un homme et une femme». En conférant, en ces termes, une interprétation limitative à la volonté du législateur, la Cour de cassation paraît prendre appui sur une conception surdéterminée du droit positif, qu’elle se réfère à la «morale civile» ou à l’idée de nature.

La tendance actuellement dominante à la Cour de cassation ainsi que dans une large part de la doctrine française et qu’il est permis de qualifier d’homophobe n’est pas indemne du soupçon d’incohérence. En quoi la concubine adulte mérite-t-elle un sort meilleur que les partenaires d’un couple de deux personnes de même sexe auxquelles la loi qui leur interdit le mariage n’offre d’autre modèle de vie commune que le concubinage? Et l’on rappellera à cet égard le grief d’incohérence dirigé contre la Cour suprême des États-Unis selon laquelle la doctrine de la privacy garantit le droit de la femme à obtenir, sous certaines conditions, une interruption médicale de grossesse alors qu’elle ne protège pas contre les visites domiciliaires et la menace d’une sanction pénale les cohabitants de même sexe.

33 Les travaux préparatoires révèlent des tentatives à cette fin mais elles ne purent aboutir. Voir sur ce point les conclusions précitées de l’avocat général Weber.


Le conflit de normes de même rang a déjà été mis en parallèle avec un conflit négatif souvent qualifié de lacune. Mais le conflit positif dissimule lui aussi une lacune. La technique consistant à tenir en équilibre les deux intérêts en concours (balancing of interests test, Interessenabwägung) est un faux-semblant. Pour arbitrer le cas, le juge doit nécessairement faire prévaloir sur l’autre l’intérêt qui pèse le plus lourd. Le plus souvent, toutefois, il élimine le conflit de normes par énoncé d’une troisième norme, proprement nouvelle, produit de l’ingéniosité judiciaire. Roe v. Wade en procure l’illustration la plus convaincante. Pour concilier la liberté de la femme enceinte, rattachée un peu arbitrairement à la doctrine de la privacy, et la somme d’intérêts divers que l’État peut faire valoir, la Cour suprême a enfermé la liberté de la femme dans des limites précises : l’interruption médicale de grossesse ne peut intervenir que dans les trois premiers mois de la conception et elle doit être pratiquée en milieu hospitalier par un médecin. Ainsi, le juge de la constitutionnalité n’est pas seulement « un législateur négatif », selon la formule de Kelsen, il se voit contraint d’élaborer une règle inédite.

D. La division des juges

Non seulement le conflit de lois, le conflit de normes exigent un arbitrage permanent des juges mais la logique du conflit se poursuit jusqu’au sein de la fonction judiciaire. Même dans un pays comme la France où l’absence d’opinions séparées ou d’opinions dissidentes donne aux décisions judiciaires un vernis d’unanimité, il est possible d’observer de telles divergences. L’action de la concubine en fournit un exemple puisqu’elle a opposé la Chambre criminelle à la Chambre civile de la Cour de cassation, l’arrêt de la Chambre mixte de 1970 ayant fait triompher la position plus libérale de la première. Il n’est, dès lors, pas exclu que si la Chambre criminelle venait à être saisie de l’action civile du partenaire de même sexe de la victime d’un homicide, elle serait plus accueillante à l’argument d’analogie que ne l’a été une Chambre civile en 1997. Dans l’affaire tranchée par ce dernier arrêt, l’avocat général Weber avait conclu en sens contraire, en invitant la Cour à « interpréter

les textes en fonction de l’évolution de la société à laquelle ils s’appliquent. Pour encourager la Cour à ce faire, il avait relevé que la formation saisie était la Chambre spécialisée des baux d’habitation dont la décision n’aurait pas la même force de précédent qu’un arrêt de l’Assemblée plénière. La doctrine de l’acte clair n’a pas non plus fait l’unanimité (infra, note 40).

À la Cour européenne des droits de l’homme, où la diversité des motivations peut se faire jour, il existe des décisions dont le dispositif n’est obtenu que par une majorité boîteuse. Un arrêt du 25 mai 1993 était relatif à l’accusation, dirigée contre la Grèce, de violation des libertés de religion (art. 9) et d’expression (art. 10) garanties par la Convention. Un Témoin de Jéhovah avait subi une condamnation pénale en vertu d’une loi sur la répression du prosélytisme manifestement discriminatoire puisqu’elle n’avait jamais été appliquée qu’à la diffusion de croyances religieuses autres que la religion orthodoxe dominante. Composée de neuf juges, la Cour s’est divisée en trois groupes de trois. Pour le premier la loi est, comme telle et par l’application qui y est donnée, discriminatoire. Trois autres juges estiment au contraire que ni la loi ni la condamnation faisant grief au requérant n’étaient contraires à la Convention. Les trois juges qui énoncent l’opinion de la cour suivent une voie médiane; bien que la loi ne soit pas discriminatoire et qu’elle poursuive une fin légitime, l’application qui en a été faite en l’espèce ne satisfait pas au principe de proportionnalité. Les trois membres du premier groupe se joignent au verdict de condamnation et cette majorité boîteuse n’adresse pas un message clair ni aux États ni aux personnes relevant de leur compétence.

À la Cour suprême des États-Unis les opinions composites (plurality opinions) sont devenues de plus en plus nombreuses. En voici un exemple récent. La loi fédérale sur les naturalisations (8 USC § 1409 (a)(4)) soumet à des conditions plus rigoureuses l’octroi de la nationalité américaine à l’enfant né à l’étranger des relations entre un Américain et une étrangère

---


qu’à celui dont la mère est américaine. Pareille discrimination entre le père et la mère (mais dont seul l’enfant est victime) est-elle contraire à l’*Equal Protection Clause* du cinquième amendement? Le grief d’inconstitutionnalité a été rejeté par six juges sur neuf39, mais pour des motifs incompatibles. Le juge Stevens annonce la décision de la Cour et, selon son opinion, à laquelle ne se rallie que le Chief Justice (Rehnquist), la disposition législative critiquée résiste au contrôle le plus strict (*heightened scrutiny*). D’après la juge O’Connor et le juge Kennedy elle ne satisfait pas à ce test, mais bien à celui de la *rational scrutiny*. Enfin, les juges Scalia et Thomas estiment que les conditions auxquelles le Congrès subordonne l’acquisition de la nationalité par les personnes nées à l’étranger sont soustraites à tout contrôle juridictionnel. Les trois juges émettant des opinions dissidentes ne parviennent même pas à y donner une formulation unique, la juge Ginsburg ayant émis une opinion dissidente à laquelle se joignent les juges Souter et Breyer et qui est complétée par l’opinion du juge Breyer, à laquelle se joignent les juges Souter et Ginsburg.

Pour s’en tenir à l’opinion composite qui conduit au rejet de la requête, force est de constater qu’elle s’inscrit en dehors de toute ligne de précédent. L’opinion la plus radicale (celle du juge Scalia) refuse de se prononcer sur la constitutionnalité de la loi. Les deux autres groupes d’opinions s’opposent sur une question qui n’a pas cessé de diviser la jurisprudence américaine : alors que le test le plus rigoureux du contrôle de constitutionnalité s’applique incontestablement aux discriminations raciales et ethniques, il demeure controversé si le même test doit être étendu aux discriminations sexuelles (*gender discriminations*). Bien loin d’y mettre fin l’arrêt nourrit la controverse.

On peut donc conclure cette deuxième partie de l’exposé en observant que les juges sont parfois en défaut d’arbitrer l’incertitude du droit et qu’en affichant leur division, ils contribuent à l’entretenir.

III. Doute méthodique et incertitude

Quand il faut appliquer la constitution, une loi ou un précédent, il y va d’un problème d’interprétation, et l’un des pièges auxquels le juge est exposé, mais auquel il est rarement capable de résister, consiste à séparer à l’aide d’une formule préétablie le bon grain de l’ivraie. Seuls les textes ambigus ont besoin d’être interprétés, la plupart des conflits sont évacués par le doute qui est jeté sur leur réalité. On aura reconnu ici la doctrine française de l’acte clair et la théorie américaine du constitutional doubt. En Allemagne aussi la jurisprudence fait le partage entre un texte ou une notion univoque (eindeutig) et celle qui est ambiguë (ou plurivoque, zweideutig). Voici quelques illustrations de la méthodologie du doute.

A. La notion d’acte clair

La théorie de l’acte clair s’est greffée sur une scission entre l’application et l’interprétation. Selon la tradition française il est des normes que le juge doit appliquer sans pouvoir les interprétérer. Tel est le cas, notamment, pour les traités internationaux dont l’interprétation relève en principe du gouvernement auquel il appartient de déterminer l’étendue de ses obligations dans les relations qu’il conduit au nom de la République avec les autres États. Une interprétation erronée ou malencontreuse du juge national pourrait engager la responsabilité internationale de l’État. C’est par la voie d’un référé au ministre des Affaires étrangères que les cours et tribunaux ainsi que les juridictions administratives sollicitent une interprétation qui, au moins dans l’ordre interne, doit être tenue pour authentique. La prolifération des traités internationaux devant être appliqués par les tribunaux a conduit à restreindre la portée de la règle. Cela se fait de diverses manières qu’il serait d’autant plus superflu de développer ici qu’elles illustrent une observation précédemment faite, à savoir la divergence entre les divers organes de l’État40. Même dans les cas où le juge serait en principe tenu de solliciter

40 Voir notamment Henri BATIFFOL et Paul LAGARDE, Traité de droit international privé, 8e éd., t. 1er, Paris, LGDJ, 1993, n° 3 : «Cette question complexe qui oppose depuis longtemps la Cour de cassation et le Conseil d’État, et même les différentes chambres de la Cour de cassation entre elles, est actuellement en cours d’évolution».
l’interprétation du Gouvernement, il peut s’en abstenir si l’interprétation du texte n’est pas douteuse. Le texte doctrinal dont l’exégèse a soutenu la notion contemporaine d’acte clair (bien qu’il ne contienne pas cette expression) est l’œuvre du vice-président du Conseil d’État, Laferrière :

*Cela revient à dire que, pour qu’il y ait question préjudicielle, il faut qu’il y ait une question, et qu’elle préjuge en tout ou en partie le jugement du fond. Il faut qu’il y ait une question : c’est-à-dire une difficulté réelle, soulevée par les parties ou spontanément reconnue par le juge, et de nature à faire naître un doute dans un esprit éclairé.*41

Et de préciser sa pensée à la page suivante :

*Dans le doute il faut surseoir. S’il hésite, il doit surseoir, car lorsqu’on est embarrasé de dire si une question est ou non douteuse, tout porte à croire qu’elle l’est réellement.*

Le problème du référend en interprétation va rejaillir avec l’entrée en vigueur de l’article 177 du traité CE, qui, en certaines circonstances, impose aux juridictions des États membres, en principe compétentes pour appliquer le droit communautaire, d’adresser à la Cour de justice des Communautés européennes une question préjudicielle d’interprétation. Dans ses conclusions précédant un arrêt du 27 mars 1963, l’avocat général Lagrange paraphrase la doctrine de Laferrière, ce qui n’est guère étonnant puisqu’il était lui-même issu du Conseil d’État. Mais il ajoute au texte de 1896 en évoquant explicitement «la théorie de l’acte clair» :

*si le texte est parfaitement clair, il n’y a plus lieu à interprétation, mais à application, ce qui ressortit à la compétence du juge chargé précisément d’appliquer la loi. C’est ce qu’on appelle parfois, d’une expression d’ailleurs peu exacte et souvent mal comprise, la théorie de l’acte clair : à vrai dire, il s’agit simplement de la*  

---

lignes de démarcation entre deux compétences. Bien entendu, comme toujours en pareil cas, il peut y avoir des cas douteux, des cas limites; dans ce doute, évidemment, le juge devrait prononcer le renvoi.42

Bien qu’aucun des deux arrêts de 1963 et de 1964 n’ait dû entériner l’enseignement de son avocat général, des décisions ultérieures paraissent le refléter en ce compris la réticence exprimée à l’égard de la notion d’acte clair. Elle exclut notamment le devoir de poser la question préjudicielle dans l’hypothèse suivante:

Enfin, l’application du droit communautaire peut s’imposer avec une évidence telle qu’elle ne laisse aucun doute raisonnable sur la manière de résoudre la question posée. Avant de conclure à l’existence d’une telle situation, la juridiction nationale doit être convaincue que la même évidence s’imposerait également aux juridictions des autres États membres et à la Cour de justice. Ce n’est que si les conditions sont remplies que la juridiction nationale pourra s’abstenir de soumettre la question à la Cour et la résoudre sous sa propre responsabilité.43

En 1964, un an après la position prise par l’avocat général Lagrange devant la Cour de justice des Communautés européennes, la doctrine du Conseil d’État fait retour à son lieu d’origine, précisément à propos de l’application de l’article 177 du traité CE. C’est sans faire référence aux conclusions précitées, mais seulement à Laferrière que Madame Questiaux, commissaire du gouvernement, conclut dans les termes suivants:


L’application d’actes clairs au procès ne soulève pas de question préjudicielle.44

Les conclusions de Madame Questiaux et la doctrine de Laferrière sont encore évoquées à propos de la procédure en cassation :

Cette référence, faite par Laferrière, à un « esprit éclairé » et impartial, qui est celui du magistrat, est indispensable.45

Le moindre reproche qu’on puisse adresser à la théorie de l’acte clair est qu’elle pèche par imprécision. Le texte de Laferrière qui paraît véritablement la matrice de cette théorie et auquel se réfère la précédente citation exige l’opération d’un « esprit éclairé, et impartial ». Ainsi, la clarté n’est pas une vertu émanant du texte lui-même, elle nécessite le truchement d’un « esprit » éclairé. La circularité du propos ne saurait échapper : seule une lumière d’origine subjective établit la clarté du texte. Les commentateurs qui font l’ économie du concept d’acte clair recourent à des raisonnements non moins circulaires : pour que la question préjudicielle doive être posée « il faut qu’il y ait une question » (Laferrière) et si la question de la question est elle-même douteuse, il y a matière à question (Laferrière et Lagrange). Ce doute est qualifié par la Cour de justice (« aucun doute raisonnable »). À la question de la question paraît faire écho le doute sur le doute. S’il est un cas dans lequel l’ indétermination du droit se réfugie sous un langage abscons et lâche, c’est bien celui-là, et la conclusion qui en est tirée est que c’est à « l’esprit éclairé et impartial » du juge qu’il appartient de faire la lumière.

B. La doctrine du « constitutional doubt »

À la Cour suprême des États-Unis, la doctrine du constitutional doubt contribue à la police du contentieux de constitutionnalité. De même que le doute profite à l’accusé – In

44 Conclusions précédant Conseil d’État, 19 juin 1964, Rev. dr. public, 1964, 1019, 1029.

45 Jacques BORÉ, « Un centenaire : le contrôle par la Cour de cassation de la dénaturation des actes », (1972) 70 Rev. tr. dr. civil 245, 265.
Pour décisions, multiplicité des motifs avancés à l’appui de la solution. Rigoureuse successivement décisions de excluant il qui mérite doute. Pareille: juges cette avant. rang dubio 42

Elle que doctrine non est que question/nobreakspace d’inconstitutionnalité est longue, que la teneur de la partie est évidente. Il faut considérer rigoureusement entre deux questions d’interprétation et la multiplicité des motifs avancés à l’appui de la solution.

Bien qu’on se soit limité à un échantillon restreint de décisions, force est de constater la variété des formulations. Pour que la loi résiste au doute quant à sa constitutionnalité il suffit que l’interprétation avancée à l’appui de celle-ci soit "fairly possible". Pareille interprétation peut être retenue encore qu’elle paraisse contraire au "sens littéral". Même si

46 Elle a ses racines dans Murray v. The Charming Betsy, 6 US (2Cranch) 64, 143 (1804).

47 United States v. Jin Fuey Moy, 36 S Ct 658, 659 (1916); Panama R. Co. v. Johnson, 264 US 375, 390 (1924); Crowell v. Benson, 285 US 22, 62 (1932); Almandarez-Torres v. United States, 118 S Ct 1219 (1998), J. Scalia dissenting, 1243. Mais la même opinion précise le sens des mots employés: «I emphasize that fairly possible is all that needs to be established. The doctrine of constitutional doubt does not require that the problem — avoiding construction be the preferable one — the one the Court would adopt in any event. Such a standard would deprive the doctrine of all function».

48 United States ex rel. Attorney General of the United States v. Delaware and Hudson Co., 29 S Ct 527, 536 (1906): «despite the literal sense of some of the prohibitions, they should all be construed so as to accomplish the result intended». 
«le langage de la loi est plutôt complexe», il faut considérer «la question dans son contexte», ce qui oriente dans une seule direction49. Pour que la Cour puisse écarte le doute de constitutionnalité il faut que la disposition législative puisse raisonnablement recevoir deux interprétations50. Des interprétations alternatives il ne faut écarter que celle qui serait «manifestement contraire à l’intention du Congrès»51. Enfin tant le doute que le problème ou la question sont qualifiés en des termes variés : «serious doubt»52, «grave doubt» ou «doubt gravely»53, «serious constitutional problems»54; «significant constitutional question» et «serious likelihood that the statute will be held unconstitutional»55.

La seconde observation tend à distinguer plus rigoureusement deux questions d’interprétation qui peuvent, l’une comme l’autre, être douteuses. Existe-t-il un doute quant à la constitutionnalité de la loi et l’interprétation de la loi est-elle douteuse? Le juge Scalia qui distingue en ces termes les deux questions56 n’appréhende toutefois pas le cœur du problème : une chose est d’interpréter la loi, une autre chose d’interpréter la constitution. Comme il a déjà été indiqué ci-dessus, ce n’est pas seulement la loi qui est ouverte à plusieurs interprétations, il dépend aussi de l’interprétation de la constitution – quel que

49 Almandarez-Torres v. United States, précité, note 47, 1228.
55 Almandarez-Torres v. United States, précité, note 47, 1228.
56 Almandarez-Torres v. United States, précité, note 47, 1234.
soit, par ailleurs, le sens de la loi – que celle-ci satisfasse ou non à la norme supérieure telle qu’elle est interprétée. Or, la Cour paraît raisonner comme si le doute n’atteignait jamais la constitution elle-même.

Les motifs pour lesquels la Cour suprême des États-Unis refuse de tenir pour inconstitutionnelle une loi qui peut recevoir une interprétation compatible avec la constitution sont variés. Une première idée est l’économie de moyens : pourquoi résoudre une question de constitutionnalité qu’il est possible d’éviter? L’opinion dissidente de la juge O’Connor dans Rust v. Sullivan en donne un exemple : au lieu de se prononcer sur la constitutionnalité d’une loi du Congrès la Cour aurait pu se borner à décider que la réglementation prise en vertu de cette loi n’en donnait pas une interprétation raisonnable.

Un deuxième argument en faveur de la doctrine du constitutional doubt est que la déclaration d’inconstitutionnalité d’une loi du Congrès est «le plus grave et le plus délicat» parmi les devoirs que la Cour est appelée à remplir. Mais cette affirmation doit être tempérée (qualified) par la proposition selon laquelle «le souci d’éviter une difficulté ne doit pas être poussé au point de se transformer en faux-fuyant inavoué».

La motivation à laquelle plusieurs décisions paraissent sensibles est la déférence à l’égard du législateur et elle se conjugue alors avec une forme de réserve du judiciaire (judicial restraint).

Le Congrès n’est pas moins que cette Cour tenu, en vertu du serment que ses membres ont prêté, à respecter la Constitution. Les tribunaux ne peuvent dès lors


avancer à la légère que le Congrès a pu avoir l’intention de transgresser des libertés constitutionnelles ou d’usurper un pouvoir qui ne lui appartenait pas.\footnote{Edward J. DeBartolo Corp. v. Florida Gulf Coast Building and Construction Trades Council, précité, note 51, 575. Plus récemment : Almandarez-Torres v. United States, précité, note 47, 1228, citant Rust v. Sullivan, précité, note 57 : «This canon is followed out of respect for Congress, which we assume legislates in light of constitutional limitations».}

La fragilité de la doctrine du\textit{ constitutional doubt} est attestée par le grand nombre de décisions suivies d’opinions dissidentes\footnote{Kuhn v. Fairmont Coal Cy, précité, note 57; Crowell v. Benson, précité, note 47; Erie Railroad Co. v. Tompkins, précité, note 57; Rust v. Sullivan, précité, note 57; Almandarez -Torres v. United States, précité, note 47, arrêt le plus récent et rendu par cinq voix contre quatre.}. Le plus notable à cet égard est \textit{Blodgett v. Holden}, arrêt per curiam, dans lequel les huit membres de la Cour se sont divisés en deux groupes de nombre égal, l’opinion généralement citée étant celle du juge Holmes suivie par les juges Brandeis, Sanford et Stone.

C. Les voies du contrôle de légalité sur l’interprétation des actes juridiques privés

Bien que l’interprétation des actes juridiques privés ne soit pas une opération intellectuellement distincte de l’interprétation d’une loi, d’un règlement ou d’un acte de l’autorité publique et que certains actes privés – les contrats d’adhésion par exemple – soient beaucoup plus souvent appliqués que la plupart des lois, la Cour de cassation de Belgique s’abstient en principe de contrôler l’interprétation de tels actes par le juge du fond\footnote{Voir François RIGAUX, \textit{La nature du contrôle de la Cour de cassation}, Bruxelles, Bruylant, 1966, pp. 284-302; «Le juge, ministre du sens», dans \textit{Justice et Argumentation}, \textit{Essais à la mémoire de Chaim Perelman}, Bruxelles, Éditions de l’ULB, 1986, pp. 79-95; \textit{La loi des juges}, Paris, Odile Jacob, 1997, pp. 192-198.}. Il n’y est procédé que si ce juge est accusé d’avoir excédé son pouvoir d’interpréter les actes juridiques privés, bien que ceux-ci ne soient que de simples «faits» au regard du contrôle de légalité. La distinction entre le
fait et le droit\textsuperscript{63} qui, en principe, ferait obstacle à un tel contrôle est tempéré par une autre distinction, celle qui sépare les actes privés dont les termes sont «clairs et précis» de ceux qui sont ambigus. Quand le juge du fond s’est écarté du sens «objectif» ou «naturel» de telles dispositions contractuelles il a, selon la Cour de cassation de France, dénaturé ces actes ou, d’après le langage de la Cour de cassation de Belgique, violé la foi qui leur est due\textsuperscript{64}.

La Cour refuse de contrôler l’interprétation des actes juridiques privés sauf si le juge du fond les a dénaturés : «Il n’est pas permis aux juges, lorsque les termes des conventions sont clairs et précis, de dénaturer les obligations qui en résultent et de modifier la stipulation qu’elles contiennent»\textsuperscript{65}. Ou encore : «La Cour de cassation n’exerce son droit de contrôle que lorsque sous prétexte d’interprétation, le juge détourne de leur sens les termes clairs et formels d’une convention»\textsuperscript{66}.

L’adoption de ces terminologies voile (assez pauvrement) la réalité : la doctrine de la dénaturation comme celle de la foi due aux actes imposent à la juridiction de cassation de comparer l’une à l’autre deux interprétations de l’acte juridique privé, celle qui a été faite par la décision attaquée à celle que révèle à la Cour elle-même le «sens naturel» des termes. Tant la distinction entre les actes clairs et ceux qui sont ambigus que la distinction du fait et du droit ne sont que des chevilles bouchant les interstices d’un raisonnement qui s’épuise dans ses propres dichotomies. De même que la Cour de cassation répute question de fait les points qu’elle se refuse à examiner, elle affirme l’ambiguïté des écrits privés sur l’interprétation desquels elle étend son contrôle. Ce faisant, elle donne à ses propres méthodes une certitude illusoire. Est-ce moins arbitraire que les nombreuses décisions

\textsuperscript{63} Voir F. RIGAUX, La nature du contrôle de la Cour de cassation, op. cit., note 62, pp. 75-140.

\textsuperscript{64} Sur la différence d’approche entre les deux cours de cassation, voir F. RIGAUX, La nature du contrôle de la Cour de cessation, op. cit., note 62, pp. 285-288. Sur la doctrine française de la dénaturation, voir l’article de J. BORÉ, loc. cit., note 45.

\textsuperscript{65} Civ., 15 avril 1872, D., 1872, I, 176.

\textsuperscript{66} Cass. (Belg.) (1er Ch.), 9 février 1933, 2de espèce, Pas., 1933, I, 103, 125.
de rejet de la requête en certiorari que la Cour suprême des États-Unis motive par la constatation que la demande ne soulève pas une question «substantielle» de droit constitutionnel ou de droit fédéral?

D. Univocité ou ambiguïté des textes soumis au juge

Un exemple emprunté à la jurisprudence constitutionnelle allemande illustre la nature illusoire de la recherche du sens usuel, du sens naturel ou du sens littéral d’un texte isolé de l’application qui en est faite en articulation avec des normes diverses. Des militants pacifistes sont poursuivis pour avoir apposé sur leur véhicule un autocollant reproduisant une phrase de Kurt Tuscholsky : «Soldaten sind Mörder». Il leur est reproché de diffamer la Bundeswehr, mais l’accusation ne vise pas, même collectivement, les membres de celle-ci, elle ne lui impute pas les crimes commis par la Wehrmacht durant la Seconde Guerre mondiale, elle affirme plutôt ce que l’auteur du texte tient – à tort ou à raison, peu importe – pour une vérité universelle, à savoir que tout soldat équipé d’armes aptes à donner la mort et exercé à s’en servir est un meurtrier potentiel. Il s’agit donc d’une opinion plutôt que de l’énoncé d’un fait, et, en régime démocratique, les opinions, dont il est impossible de démontrer la vérité ou la fausseté, bénéficient d’une protection constitutionnelle renforcée. Dès lors, l’interprétation du texte litigieux est elle-même subordonnée à l’interprétation de la loi en vertu de laquelle son auteur ou son utilisateur pourrait être réprimé. Mais la loi ne peut être appliquée, et interprétée, qu’en fonction du sens attribué à la norme – constitutionnelle ou internationale – qui garantit la liberté d’expression. Si l’on y ajoute les autres normes constitutionnelles ou internationales qui protègent le droit à l’honneur au rang de bien de la personnalité ou qui sauvegardent la présomption d’innocence de l’accusé en matière pénale, on aura à peu près fait le tour d’un problème qui ne saurait se réduire à la lecture ou à l’exégèse de chacun des textes séparés des autres. Aucun de ceux-ci n’a de sens littéral et quand le juge constitutionnel allemand se fonde sur le caractère non univoque (nicht eindeutig) du texte de l’autocollant pour l’interpréter à la lumière de tous les autres textes avec lesquels il entretient une relation contextuelle, il se

---

donne une raison commode d’échlorer le fantôme du sens littéral. L’interprétation de la phrase «Soldaten sind Mörder» est inséparable de l’interprétation des quatre normes en présence et la signification d’aucune de celles-ci ne se laisse disjoindre de celle qui est donnée aux trois autres. Ce n’est pas à dire que ce qui est décoré du titre de sens littéral soit privé de toute signification : il peut se concevoir «littéralement» qu’un groupe de personnes désignées comme des soldats soient publiquement accusées de meurtre et c’est sous le bénéfice de cette signification que le juge pénal a cru pouvoir prononcer une condamnation du chef de diffamation. Mais il s’agit d’une interprétation parmi toutes les autres et pour le choix entre celles-ci le juge du fond doit notamment tenir compte des règles de poursuite et d’incrimination en matière pénale : parmi plusieurs significations il doit choisir celle qui est la plus favorable à l’acquittement de l’accusé. En outre, le contexte oriente vers la même interprétation : la phrase est empruntée à un écrivain qui n’a pu viser une situation factuelle déterminée, et le choix du sens anagogique se trouve renforcé par l’utilisation qui en est faite par un mouvement pacifiste.

**Conclusion : jugez sur ce que je fais plutôt que sur ce que j’en dis**

On pourrait s’étonner que la doctrine américaine du *constitutional doubt* ne soumette au «doute» que l’interprétation de la règle inférieure et non celle de la norme de rang supérieur apte à déterminer la constitutionnalité de la première. L’une des explications est que l’organe habilité à formuler chacune des interprétations n’est pas identique dans les deux cas. La juridiction investie en dernier ressort du contrôle de constitutionnalité est l’interprète autorisé des normes constitutionnelles mais il se peut que ses différents membres ne s’accordent pas sur la signification d’un texte ou que la force d’un précédent soit récusée par une nouvelle majorité. La

---

68 Selon la formule de Marbury v. Madison, 5 US (1 Cranch) 137 (1803), réitérée dans Cooper v. Aaron, 358 US 1, 18 (1958) : «It is emphatically the province and duty of the judicial department to say what the law is. [...] It follows that the interpretation of the Fourteenth Amendment enunciated by this Court in the Brown Case is the supreme law of the Land, and Article 6 of the Constitution makes it of binding effect on the States...»
certitude ainsi atteinte est affectée d’une double relativité : dans le temps, d’abord, puisqu’elle est exposée à être renversée, quant à la matière ensuite, car sans faire l’aveu d’un revirement de jurisprudence une décision ultérieure peut limiter la portée de celle qui la précède\textsuperscript{69}. En revanche, pour l’interprétation de la norme inférieure la juridiction constitutionnelle ne bénéficie pas d’une égale latitude. La doctrine du \textit{constitutional doubt} implique le respect de la volonté du législateur ordinaire, éclairée par l’exploitation des travaux préparatoires, même si celle-ci est souvent controversée et permet des conclusions opposées\textsuperscript{70}. Quand la norme dont l’inconstitutionnalité est alléguée appartient au droit d’un État fédéré, la Cour suprême de cet État est en principe compétente pour y donner une interprétation qui lie les juridictions fédérales\textsuperscript{71}.

À la différence de la Cour suprême des États-Unis qui peut appliquer et interpréter le droit fédéral, les juridictions constitutionnelles spécialisées n’ont certainement pas le pouvoir d’appliquer les lois ordinaires ni sans doute même celui de les interpréter. Un arrêt récent de la Cour d’arbitrage atteste bien la réserve observée à cet égard par la juridiction constitutionnelle belge. Sur une demande de question préjudicielle relative à la nature de la mission du directeur des contributions, la Cour d’arbitrage établit une alternative entre deux interprétations de la disposition légale pertinente (l’article 366 du Code des impôts sur le revenu 1992) : si celle-ci a conféré au directeur des contributions une mission juridictionnelle, et, à défaut d’une procédure de récusation, l’article 366 précité est contraire au principe d’égalité des articles 10 et 11 de la Constitution. S’il

\textsuperscript{69} Selon un arrêt récent de la Cour suprême des États-Unis : \textit{«Stare decisis is not an inexorable command»} : State Oil Co v. Khan, 118 S Ct 275 (1997).

\textsuperscript{70} Voir par exemple Ratzlaf \textit{v. United States}, 126 L Ed 2d 615, 626 (1993) : \textit{«There are, we recognize, contrary indications in the statute’s legislative history.»} Ou encore National Organization for Women \textit{v. Scheidler}, 127 L Ed 2d 99 (1994) : \textit{«Both parties rely on legislative history to support their positions»}.

s’agit au contraire d’une mission administrative, le vice d’inconstitutionnalité doit être écarté. La Cour d’arbitrage ne choisit pas entre les deux interprétations et elle n’a pas le pouvoir de le faire puisqu’il ne lui appartient pas d’appliquer la loi fiscale. À l’arrière-plan de la demande de question préjudicielle il faut noter que la Cour de cassation, compétente pour contrôler la légalité des décisions rendues en matière fiscale, avait donné à la mission du directeur des contributions la qualification juridictionnelle à présent tenue pour inconstitutionnelle par la Cour d’arbitrage. Il ne restera, dès lors, à l’avenir, à la Cour de cassation que de modifier sa jurisprudence, sans que la cour d’arbitrage ait le pouvoir de la censurer si elle s’abstient de le faire. Mais le juge du fond qui a posé la question est en principe tenu de donner à la disposition légale l’interprétation en harmonie avec la Constitution.

A. L’administration des questions préjudicielles d’interprétation par la Cour de justice des Communautés européennes

Le caractère artificiel de la plupart des distinctions dichotomiques qui ne donnent qu’une apparence de certitude au raisonnement judiciaire est encore attesté par l’évolution de la jurisprudence de la Cour de justice des Communautés européennes à propos de la distinction entre l’application et l’interprétation d’une norme de droit communautaire.

Dans ses premiers arrêts, la Cour de justice insiste avec force sur cette distinction et sur le partage de compétence qui en est la conséquence. Un arrêt du 30 janvier 1974 fait encore une application rigoureuse de la distinction en décidant à propos de la question préjudicielle qui lui est posée :

---


73 De la même manière, le Tribunal constitutionnel fédéral allemand laisse à la juridiction de cassation en matière civile (Bundesgerichtshof) la compétence de contrôler l’interprétation du droit civil : BVerfG 14 février 1973, Prinzessin Soraya, BverfE, 34, 269, 279 et 280. Voir F. RIGAUX, La loi des juges, op. cit., note 62, pp. 139-142.


75 Voir notamment l’arrêt du 27 mars 1963, précité, note 42, 76.
qu’il s’agit non d’une question d’interprétation mais d’application réservée à la juridiction nationale.\textsuperscript{76}

Toutefois, bien qu’elle répète, à titre de principe, qu’il appartient exclusivement à la juridiction nationale de décider si la disposition ou la décision de droit interne est ou non conforme au droit communautaire tel qu’il a été interprété dans la réponse à la question préjudicielle, la Cour ne laisse pas de faire implicitement pareille «application» de ce droit en décidant, par exemple, «qu’une taxe présentant les caractéristiques mentionnées ci-dessus» ne doit pas être «remboursée à l’exportateur du produit en cause»\textsuperscript{77}.

La barrière de la fiction est délibérément franchie quand la Cour écrit, par exemple que :

\textit{Le droit communautaire s’oppose à l’application d’une règle de procédure nationale qui, dans des conditions telles que celles de la procédure en l’espèce au principal, interdit au juge national, saisi dans le cadre de sa compétence, d’apprécier d’office la compatibilité d’un acte de droit interne avec une disposition communautaire, lorsque cette dernière n’a pas été invoquée dans un certain délai par le justiciable.}\textsuperscript{78}

De la même manière que la Cour d’arbitrage a, sans le dire expressément, censuré l’interprétation donnée par la Cour de cassation à une règle de droit fiscal et indirectement contraint la juridiction suprême en matière civile à aligner sa jurisprudence sur le verdict d’inconstitutionnalité, après que la Cour de justice des Communautés européennes a énoncé l’incompatibilité avec son interprétation du droit communautaire d’une règle de droit interne présentée comme


imaginaire mais qui est la sœur jumelle de la norme nationale dont l’application avait été l’occasion de la demande de question préjudicielle, le juge national qui reçoit cette réponse y trouve quasi dictée la solution qu’il doit donner au conflit de normes qui avait été soulevé devant lui.

B. Canons d’interprétation et recours aux travaux préparatoires

Certains canons d’interprétation faisant appel au contenu substantiel de la norme constitutionnelle ou internationale renforcent ou tempèrent, selon le cas, la présomption de constitutionnalité. La vertueuse hostilité que la Cour suprême des États-Unis affecte aujourd’hui à l’égard des classifications liées à la race ou à l’appartenance ethnique justifie d’élever le niveau du contrôle de constitutionnalité (*heightened scrutiny*). Le statut des libertés fondamentales appelle aussi un contrôle particulièrement vigilant : c’est pourquoi on ne saurait imputer au Congrès l’intention de « jeter au bûcher toutes nos traditions » inscrites dans le quatrième amendement\(^7\). Le Tribunal constitutionnel fédéral allemand a mis en œuvre une directive d’interprétation analogue pour annuler un arrêt de l’Oberverwaltungsgericht qui avait restreint par son interprétation la garantie constitutionnelle des libertés de réunion et de manifestation\(^8\). Le principe constitutionnel de légalité des incriminations et des peines requiert aussi un canon interprétatif qui restreigne le champ d’application d’une norme pénale selon une vieille directive qui remonte à l’ancien droit, *Odiosa sunt restringenda* ou *lenity principle*\(^9\). Toutefois la mise en œuvre de ce principe ramène à la doctrine du sens clair car il ne s’appliquerait que pour évicier les termes ambigus de


la norme pénale. Là où la volonté du législateur est «claire», s’arrête le pouvoir d’interprétation du juge :

If the intent of Congress is clear, that is the end of the matter; for the court, as well as the agency, must give effect to the unambiguously expressed intent of Congress.\(^8\)

C’est encore au couple clarté-ambiguïté que l’on revient pour la mise en œuvre d’une autre méthode d’interprétation, l’utilisation des travaux préparatoires (legislative history, Entstehungsgeschichte) de la loi ou du traité.

Outre que l’interprétation des travaux préparatoires est elle-même parfois conflictuelle\(^8\) et n’est alors qu’un «faible roseau» (a rather thin reed) donnant appui à une interprétation de la loi, les juges adoptent souvent un double langage à l’égard des travaux préparatoires. Tantôt ils déclarent n’y avoir recours que pour contribuer à l’interprétation d’un texte obscur ou ambigu\(^8\) et l’écartent quand le texte est clair\(^8\). Toutefois, quand la majorité de la Cour suprême des États-Unis s’appuie fortement sur la clarté d’un langage non équivoque (plain language), assurance discréditée par des voix dissidentes autorisées\(^8\), on peut douter de la valeur d’une méthode d’interprétation accrochée à la notion de langage univoque (plain language). Tantôt, et ceci vient renforcer les doutes, les juges affirment que le texte est «suffisamment clair en lui-même» et rend superflu tout recours aux travaux préparatoires, pour ensuite y rechercher une confirmation de l’interprétation.

---


déduite du seul texte. Pareille expression de l’insécurité des juges est partagée par la Cour internationale de Justice87 et par la Cour suprême des États-Unis88. Mais d’éminents juges critiquent pareil affaiblissement de la doctrine du sens clair par un recours inutile (et, par conséquent, nuisible) aux travaux préparatoires. C’est ce que fit, dès 1932, le juge Anzilotti dans son opinion dissidente sous un Avis consultatif de la Cour permanente de Justice internationale :

Si vraiment l'article 3, d'après le sens naturel des termes était parfaitement clair, il ne serait guère admissible de chercher une interprétation autre que celle qui répond au sens desdits termes.89

De même, le juge Schwebel, actuellement président de la Cour internationale de Justice, a joint une opinion dissidente sous l’arrêt précité du 15 février 199590. On peut encore citer l’opinion convergente du juge Scalia sous INS v. Cardozo-Fonseca. Il y déplore l’examen circonstancié des travaux préparatoires de la disposition litigieuse :

Although it is true that the Court in recent times has expressed approval of this doctrine, that is to my mind an ill-advised derogation from the venerable principle that if the language of a statute is clear, that language must be given effect — at least in the absence of a patent absurdity.91

---

Entre l’opinion du juge Scalia et celle des présidents Anzilotti et Schwebel, il existe une différence notable : alors que le premier paraît faire confiance à la possibilité de fonder la solution sur la doctrine du sens clair, les deux juges internationaux sont plus circonspects à cet égard. Anzilotti paraît même sceptique :

*Je ne vois pas comment il est possible de dire qu’un article d’une convention est clair avant d’avoir déterminé l’objet et le but de la convention, car c’est seulement dans cette convention que l’article assume sa véritable signification.*

Le même scepticisme est partagé par la doctrine.

Que les principales théories jurisprudentielles analysées dans le présent exposé paraissent se contredire s’explique par l’éclectisme et, osons le dire, l’opportunisme des motivations. Mais le plus significatif reste le caractère illusoire et, même, simplificateur de ces théories. Les juges ne cessent de faire ce dont ils déclarent devoir s’abstenir et les concepts qu’ils utilisent comme les canons d’interprétation auxquels ils se réfèrent ne donnent même pas une justification formelle à leurs décisions.

Cela ne les empêche pas de conférer au droit une certitude qu’il n’aurait pas sans eux, mais une certitude éphémère qui ne contribue guère à la prévisibilité des solutions.

---

92 *CPJI*, Série A/B, n° 50, p. 383.
Copyright and Parody: Touring the Certainties of Intellectual Property and Restitution

Wendy J. Gordon

Introduction

I. Goals

A. Market failure and market success

B. Market failure in the absence of intellectual property

---

1 Copyright 2000 by Wendy J. Gordon. Professor of Law and Paul J. Liacos Scholar in Law, Boston University School of Law; Visiting Senior Research Fellow, St. John’s College, Oxford; and Visiting Fellow, Programme in Comparative Media Law and Policy, Centre for Socio-Legal Studies, Oxford (1999-2000). Thanks for useful comments are owed to Ejan Mackaay, Bob Howell and Michael Meurer, and to participants in the faculty workshop at the University of San Diego School of Law.

C.Success of the intellectual property market ......................................................... 73

II. Comparing “clear rules” with the uncertainty of case-by-case responses .......... 74
   A. Definitions: formal property rules as compared with substantive reasonableness standards ................................................................. 75
   B. Copyright’s unusual mid-range status ................................. 83
   C. Comparing Copyright and Restitution ......................... 86

III. Systemic intellectual property rules to minimize market imperfections .............. 93

IV. Using standards to minimize market imperfections: turning to the fair use doctrine and the question of parody ............ 96
   A. Suppression Triangle ....................................................... 103
   B. Pecuniary losses .............................................................. 105
   C. Managerial discretion ...................................................... 106
   D. Endowment or wealth effects: pricelessness ............................................. 106

Conclusion ................................................................................. 112
One of the supposed certainties of the common law is that persons need not pay for benefits they receive except when they have agreed in advance to make payment. The rule takes many forms. One of the most familiar is the doctrine that absent a contractual obligation, a person benefited by a volunteer ordinarily need not pay for what he has received. This rule supposedly both encourages economic efficiency and respects autonomy.

To illustrate the baseline rule: While I am out of town, my neighbor drains his swamp and in the process also dries up the mosquito haven in my backyard. I am benefited. Nevertheless, common law will probably require me neither to shoulder part of the drainage costs, nor to hand over to my neighbor any portion of the increase in land value which his actions have given me. For me to retain the benefit, and even to profit from it willfully, is not “unjust enrichment.” Had my neighbor desired to have me share the costs or profit with him, he should have approached me in advance and sought my consent, by contract.

Yet if a delivery truck hits a bump so that a bag of valuable items tumbles out and onto my back yard, the owner could compel me to return the bag or pay for it. I will be liable even though I had not agreed in advance with the owner that I would pay for the items. This seems an exception to the basic rule that one need not pay for benefits except pursuant to contract.

---

2 Restatement of Restitution, section 2 (1937). It is sometimes said that when recovery is denied, plaintiffs tend to be called “intermeddlers,” but when they win, they are more likely to be called “volunteers.” Both words refer, however, to the same basic pattern: conferring benefits on someone who has not asked for them. This article uses the terms interchangeably.

3 The same puzzle recurs — but is less obvious — when I go into a store. I am not free to take whatever I want, even though I have never agreed to the store owner’s entitlements over his goods. Robert Hale and other legal realists were most insistent on this point. For an application of their insights to the realm of copyright, see Wendy J. GORDON, “An Inquiry Into The Merits of Copyright: The Challenges of Consistency, Consent and Encouragement Theory”, (1989) 41 Stanford Law Review 1343 at 1422-1435 [hereinafter W. J. GORDON, “Merits of Copyright”].
The reader is probably objecting that this “exception” is no surprise — it is, rather, the familiar category of property. We all know that when we “take” others’ property, whether we actively grab it or passively retain it, we usually have to either return it or pay for it. We may even be subject to further responsibilities, such as accounting to the owner for profits earned by its use. In fact, another supposed baseline rule of the common law is that to own property is to have the right to exclude “any other individual in the universe”, so that any intentional taking of another’s personal property, or any intentional crossing of a real property boundary, is prima facie actionable. Therefore, the reader may argue, there is no uncertainty — merely the coexistence of separate established categories.

Yet these two categories, property and liberty, are so familiar to us that we often overlook the extent and variability with which one limits the other. It was in part to remedy this frequent oversight that, shortly after the turn of the last century, Wesley Hohfeld developed his now-famous taxonomy of legal relations. As Hohfeld explained, “rights” and “duties” are logical correlatives, in the sense that if someone has a “right” to exclusive use of Blackacre, others must have a “duty” to stay off that land. Similarly, where someone has a “liberty” or “privilege” to act, others have (as a logical correlative) “no right” to have the government stop the action. So an expansion of property “rights” logically entails both an expansion of the “duties”, and a contraction in the “liberties”, of non-owners.

---


6 W. N. HOHFELD, loc. cit., note 5.
“Perhaps that’s so,” the reader may be conceding, “Property can limit non-owners’ liberty. 7 But that does not mean that liberty, in the sense of freedom from nonconsensual obligations, is an uncertainty. It just means that liberty principles have to be rewritten to take account of this familiar exception.” For example, the principle with which this essay begins could be rewritten to say, “Persons need not pay for benefits they receive except when they have in advance agreed to pay or when the benefits constitute property owned by others.” It could further be rewritten to take account of the nonconsensual duties imposed by the law of promissory estoppel, by tort law, by municipal law, by restitution law, and so on.

Unfortunately for the hope of certainty, the end result is a vague statement somewhat like this one: “Persons need not pay for benefits they receive except when the law says otherwise.” Further complicating matters, sometimes the law says that property or “quasi-property” arises when one reaps where another has sown. 8

The underlying goal of the instant essay is to defamiliarize the relation of property and liberty so that the reader can see it afresh. This journey is one that frequently recurs in the legal literature, 9 but what this essay adds is a new itinerary. It leads the reader away from her accustomed tangible territory where fabled Blackacre and Whiteacre abide, into an intangible realm sometimes known as Intellectual Property. The latter, being less familiar than the realm of physical property, may be more capable of being seen free of the deadening overlay of habit.

In this realm of intangibles, we can see most vividly how, why and where the law erodes the two supposed certainties mentioned above: (1) the claim that persons need not

7 Milton Friedman and others have of course argued that property can function to increase liberty as well as to limit it. But that is a different topic.

8 See International New Service v. Associated Press, 248 U.S. 215, 239 (1918) (news items ordinarily considered in the public domain were considered “quasi-property” when taken by a competitor).

9 See, e.g., sources cited infra, note 11.
pay for benefits they receive except when they have agreed to pay, and (2) the claim that an intentional taking of personal property, or an intentional crossing of a property boundary, is _prima facie_ actionable. Along the way, I hope to persuade the reader that standards as vague as the “fairness” in “fair use” or the “unjust” in “unjust enrichment” have both an underlying logic and a legitimate role to play. While my points are not novel, the illustration, by means of Restitution and Copyright, might be. Among other things, uncertainty can have particular utility in the Intellectual Property area where the law gives monopoly power in order to provide economic incentives for creation. As Professors Ayres and Klemperer argue, “the last bit of monopoly pricing provides disproportionately small profits in comparison to its social cost”, so that social benefit can be significantly enhanced by legal doctrines that deprive IP owners of the certainty needed to extract the full monopoly price. Ayres and Klemperer explicitly recommend tempering the reach of IP law by using open-textured standards.

---

10 Standards such as “reasonableness” or “fairness” delegate so much power to decision-makers in individual cases that they make prediction difficult, that is, standards produce _ab ante_ uncertainty. A prolific literature compares and contrasts the functions served by sharp-edged and certain _rules_ with vaguer _standards_. A wonderful window into that literature is offered by two short pieces that introduce its major themes and do much to clarify them: Carol ROSE, “Crystals and Mud in Property Law”, (1988) 40 Stan. L. Rev. 577 and Frederick SCHAUER, “Dimensions”, (1997) 82 Iowa L. Rev. 911.

11 See, for example, Joan WILLIAMS, “The Rhetoric of Property”, (1998) 83 Iowa L. Rev. 277, 278 (seeking to “defamiliarize” property notions.) The second of the claims I investigate — the notion that property owners have a truly exclusive dominion — has been notably subject to scholarly analysis. See, e.g., Joan WILLIAMS, _supra_; Carol M. ROSE, “Canons of Property Talk, or, Blacktone’s Anxiety”, (1988) 108 Yale L. J. 601; Frank MICHELMAN, “Ethics, Economics, and the Law of Property”, (1982) 24 Nomos 3 (J. PENNOCK & J. CHAPMAN (eds)).


13 _Id._, at 987-988.

14 _Id._

15 _Id._, at 1024-1026.
The essay that follows examines the boundary between two sets of rules. The first set arises under the law of Restitution, particularly the rule that volunteers ordinarily need not be rewarded. (Another way to state this same Restitution rule is to say that the retention of benefit voluntarily conferred is ordinarily not “unjust enrichment”.) The second set of rules are those of Intellectual Property law, which creates property in a special kind of volunteer. My argument is simply that the law of Restitution leads almost directly to the law of Intellectual Property, though the two areas are premised on diametrically opposed baseline certainties.

For simplicity’s sake, the essay primarily uses one methodology — that of economics. Of course, American and Canadian law have many dimensions. Rules and practices are most stable when they are supported by the convergence of many policies, of which economics is merely one.16 Thus the essay does address some additional policies, such as autonomy and the principle of equal respect for persons, when they are particularly apt. Nevertheless, for ease of exposition, the economic analysis will dominate. Hopefully, it will demonstrate both why the boundaries between liberty and property are fuzzy, and the nature of some of the principles that help shift the boundaries in one direction or another.

The essay then turns to examining the doctrine of property law that the intentional crossing of a property boundary is prima facie actionable. The essay uses three American doctrines from intellectual property, “fair use”, “the idea/expression dichotomy”, and “substantial similarity”, to demonstrate that in appropriate circumstances, even this apparent certainty must give way.

In its final stage, the essay turns from exposition to advocacy. I hope to persuade my Canadian readers to reconsider the certainty with which Canadian law now favors an established artist’s interests over those of a parodist and her audience.

---

16 Guido Calabresi’s book TRAGIC CHOICES gives many examples that illustrate the kind of instability, or cyclic institutional and rule changes, that result when resource limitations make convergence of principles unavailable. (I am also indebted here to Randy Barnett.)
Today Canada and the United States have adopted quite opposed approaches to parody. In the U.S., the “fair use” doctrine will often shelter a parody that embodies a substantial portion of the work that it ridicules. This was demonstrated vividly in the recent U.S. Supreme Court case considering whether a rap group, “2 Live Crew” could, without permission, lawfully record and commercially distribute a parody of the Ray Orbison hit, “Oh, Pretty Woman.” The Court remanded the copyright owner’s infringement case for further consideration, in an opinion that stressed the open-ended nature of the fair use doctrine. In Canada, by contrast, the doctrine of “fair dealing” does not provide much shelter for parodies, and this hostility is underlined by Canada’s generous statutory treatment of what

17 Lyrics of the parody included lines like, “Big hairy woman you need to shave that stuff” and “Two timin’ woman now I know the baby ain’t mine.” The lyrics of both the Orbison song and the parody appear in full at Campbell v. Acuff Rose Music Inc., 510 U.S. 569 (1994) at Appendices A and B to the majority opinion.


19 By contrast with the open-ended fair use doctrine of the U.S., presented at note 54 below, the Canadian fair dealing provisions are narrow and specific, applying only to research or private study, news reporting, and criticism or review. Not only would the copying involved in a parody likely violate an author’s ordinary rights under copyright law, but, since a parody distorts an original work, it could also violate the right of integrity. The Canadian provisions on fair dealing follow.

Copyright Act, R.S.C., (1985) ch. C-42, Section 29 (Can.) (as amended, 1997)

“EXCEPTIONS

Fair Dealing

Research or private study

29. Fair dealing for the purpose of research or private study does not infringe copyright.

Criticism or review

29.1 Fair dealing for the purpose of criticism or review does not infringe copyright if the following are mentioned:

(a) the source; and

(b) if given in the source, the name of the

(i) author, in the case of a work,

(ii) performer, in the case of a performer’s performance,

(iii) maker, in the case of a sound recording, or
it calls the artist’s “moral right of integrity”. Although the Supreme Court of Canada has not addressed any case of “parody”, the Canadian fair dealing statute lays out a set of crystalline rules into which it would be difficult to squeeze most parody cases.

The essay elucidates an economic logic that helps to explain the uncertain, open-ended, case-by-case treatment of the United States courts. In the process, it is hoped that at least

(iv) broadcaster, in the case of a communication signal.

News reporting

29.2 Fair dealing for the purpose of news reporting does not infringe copyright if the following are mentioned:
(a) the source; and
(b) if given in the source, the name of the
(i) author, in the case of a work,
(ii) performer, in the case of a performer’s performance,
(iii) maker, in the case of a sound recording, or
(iv) broadcaster, in the case of a communication signal.”

Canada’s moral rights statute is part of its copyright law, and applies to virtually any copyrighted work. Copyright Act, supra, note 19, s. 14.1. Relevant excerpts are set forth below at note 106. Note in particular Canada’s right of integrity, which forbids alterations to a work — even a work such as a song — that prejudice an author’s reputation.

Thus, in Cie Générale des Établissements Michelin-Michelin & Cie v. C.A.W.-Canada, (1996) 71 C.P.R. (3d) 348; 1996 CPR LEXIS 2377 (Federal Court, Trial Division, 1996), a Canadian court imposed liability on a union for its parodic use of a Michelin logo cartoon character as part of an organizing campaign at a Michelin plant. The court declined to apply the reasoning of the U.S. Supreme Court in Acuff-Rose and instead applied the Canadian statute strictly, 17 C.P.R. at 380-385. The court contrasted the non-exhaustive nature of the factors listed in the U.S. Fair Use provision with the bounded and rule-like Canadian approach. The court noted:

“The exceptions to acts of copyright infringement are exhaustively listed as a closed set in sub-sections 27(2) to 27(m) and 27(3) of the [Canadian] Copyright Act. They should be restrictively interpreted as exceptions.

[...] Parody does not exist as a facet of ‘criticism’ [...] for the purposes of the Copyright Act.” Id., at 381.

Nevertheless, the court did intimate, in dicta, that some “critical variation” of the cartoon character might be permitted “within the context of a newspaper or journal article” about the company whose symbol it was. Id., at 385.
some Canadians might be led to appreciate the merits of an approach which, under the “fair use doctrine”, sometimes does allow parodies to distort copyrighted works. My argument is in part motivated by the value that transgressive and appropriative works bring to a culture. Nevertheless, as mentioned, in this piece I will largely confine my analytic tools to the economic.

Intellectual property law is exciting because it is where we can see new rights being created as we watch. It reminds us that much of what concerns scholars comes to life daily in the hands of judges and legislators.

I. Goals

Western culture has long recognized the tension between the law’s need to speak clearly to cover broad classes of cases, and the desire to do justice in the individual case. Thus, for example, Aristotle wrote, “When the law speaks universally, then, and a case arises on it which is not covered by the universal statement, then it is right, where the legislator fails us and has erred by its over-simplicity, to correct the omission [...]”

Such practices are often referred to as “equitable”, and as presented by Aristotle (at least in this translation), they seem uncontroversial. However, it over-simplifies to speak of correcting an “omission.” A judge who wants to give individualized relief will often need to do more than merely fill in an omitted blank. Rather, she may be called on to disregard a law that indeed covers the case, or she may need to utilize an

---


equitable device (often a legal fiction such as “constructive trust”) that allows her to achieve a result contrary to what the written rule would have led an observer to expect.

A judge can have good reason for such actions. A law that is always applied literally can lead to results that undermine legitimacy. Sharp-edged rules can be over-inclusive or under-inclusive, and inappropriate results can be costly in both human and economic terms.

Judge-made exceptions constitute only one route to individualized treatment. Another route is for the legislature itself to frame its dictates in terms of a broad standard, such as “fairness” or “reasonableness.” Similarly, the legislature can subject hard-edged rules to an exception described by such a broad standard. When the legislature inserts broad language such as “fairness” into a statute, it is deliberately delegating some of its power to the individual tribunals that will be required to interpret the fuzzy command. Again the goal is to avoid the costs that a more certain rule can impose when its hard edges prove over-inclusive or under-inclusive.

But just as certainty can impose costs and threaten legitimacy, so can equity and the use of broad standards. For example, a regime of flexible and individualized treatment involves not only high administrative costs, but also dangers of bias, inconsistency, and insecurity. Just as excessive rigidity can undermine a legal system’s legitimacy, so can excessive flexibility.

One way of potentially resolving this tension is to regularize the grants of equitable treatment themselves. Thus, the likelihood of individualized relief can become gradually

---

more predictable. Another potential “resolution” is the so-called free market: If property rights are well defined, individuals can trade entitlements to suit individual needs.

In a perfect market, it is argued, all goods will flow to their highest-valued uses by the consensual behavior of the affected parties. Thus, when the Invisible Hand works properly, both certainty and individuation can coexist.

Of course, the perfect market of the Invisible Hand has virtually no counterpart in reality. In the real world, for example, transactions are costly to consummate. Third parties are affected by decisions in whose making they had no part. Further, knowledge is imperfect, extreme income inequalities abound, strategic behavior can block mutually-beneficial coordination, and some resources cannot be commodified without losing part or all of their value. So for these reasons (among many others), clear-cut property rules cannot always be relied upon. Even if one begins from neo-classical economic premises, law might appropriately favor the use of standards as well as sharply defined rules.

Although this essay touches on many topics, the central problem addressed is parody: how should a copyright court react when someone copies a copyrighted work and distorts it in a way that ridicules the original? As already mentioned, the approach is different in Canada and in the United States. In Canada, the copyright owner and potential parodist know the law contains a fairly certain rule which forbids parodies unless the copyright owner’s consent is first obtained. In the United States, by contrast, the parties know that the situation is governed by the uncertainties of the broad “fair use” doctrine. Why shouldn’t the copyright owner’s right of control be consistently enforced, leaving it to author and parodist to bargain over who values the resource use more highly? That question is addressed directly in the essay’s Section IV.

Before reaching that point, the paper explores a number of related matters. The remaining pages of Section I will briefly summarize the primary economic goals of copyright. Sections II and III compare “standards” with “rules” in the context of achieving these and related economic goals, examining doctrines within the law of both Copyright and Restitution.
Section IV addresses copying that is done for the purposes of criticism and parody. The Conclusion argues that the better approach is to utilize the broad standard of “fair use”.

A. Market failure and market success

The reader is probably familiar with the basic problem cited as the usual justification for intellectual property regimes: it is a form of market failure that prevents consumers from having available as many new intellectual products as they would be willing to pay for. Usually termed the “public goods” problem, it is briefly described below. What is less recognized is the fact that such market failure is only half of the prerequisite for justifying copyright: the other requisite condition is that there be less costly market imperfections after intellectual property is instituted than there would have been in the absence of the intellectual property regime.25

B. Market failure in the absence of intellectual property

“Public goods” are defined by having two characteristics, inexhaustibility and nonexcludability. Most intangibles have these characteristics to some extent. Intangibles tend to be inexhaustible over a large range of utilization (everyone can sing or play the same song, or build the same design of engine). They also tend to be difficult for proprietors to fence off (once encountered and remembered, anyone can reproduce the song or design).26 If production of a public good


26 That a good is “inexhaustible” does not mean that one person’s use of it will never affect others’ use. For example, at the extreme, the value of a song can become zero through saturation-level repetition, and some exhaustible physical products — such as plastic for a phonograph record or a radio set — may be necessary to afford access to the intangible. Practically speaking, then, songs may not be infinitely available to all as a valuable good.
is left entirely to the private market, lack of fencing can lead to under-production, for usually a producer needs either subsidy from government or patron, or a mode of excluding non-payors, if he or she wishes to obtain payment for what he or she has made. Without patronage or a mode of excluding free-riders, the payoff from investing in creative activities will be low, and incentives will be inadequate to induce production of as many new intangible goods as the public would be willing to pay for.

Some public goods, such as national defense, can be produced through use of a state apparatus.27 This approach has the virtue of responding to both “public goods” characteristics. State production (1) can take advantage of inexhaustibility by making the benefit available to all, and (2) resolves the problem of underproduction by requiring everyone, through taxes, to pay.

Like most nations, the United States is committed to the belief that sole reliance on state-directed production or bureaucratic subsidy is not the best way to produce inventions and art. In the realm of inventions, probably the most obvious danger of state control is the bureaucratic tendency to resist innovation. In the realm of cultural products, the most obvious dangers of state control are “lack of taste,”28 and the possibility of censorship. (A history of free enterprise, of course, also plays

---

27 The extent to which public provision of public goods is indeed necessary in various contexts is, of course, a matter of debate. For example, Ronald Coase has shown that although lighthouses are a classic public good (their light can be used by a virtually unlimited number of ships within range, and no ship can be practically excluded from their light), some lighthouses have in fact been built through non-governmental arrangements. See R. H. COASE, “The Lighthouse in Economics”, (1974) 17 J. Law & Econ. 357.

28 Or, more precisely, there is a need for taste to evolve outside the state apparatus in order for individuals to maintain some degree of genuine self-determination. See, e.g., C. Edwin BAKER, “Property and its Relation to Constitutionally Protected Liberty”, (1986) 134 U. Pa. Law Rev. 741.
a role.) Whatever the reason, there is a consensus in the United States that a diversity of private initiatives needs to be enlisted, and that state production — despite its ability to respond to both inexhaustibility and nonexcludability — should not be the primary route to follow in regard to inventions and art. As for private subsidy through foundation and the like, support from that sector is often unavailable or sparse.

Thus, the United States opted for primary reliance on use of the private market to generate incentives for the production of inventions and art. Using a market requires curing the excludability problem. Some scholars have argued that significant modes of exclusion are available independent of the law.\footnote{Scholars who are critical of copyright, or who doubt the wisdom of its expansion, typically argue that copyright may not be necessary for creators to obtain payment for their work. See, e.g., Stephen Breyer, “The Uneasy Case for Copyright: A Study of Copyright in Books, Photocopies, and Computer Programs”, (1970) 84 Harv. L. Rev. 281, 350; Tom Palmer, “Intellectual Property: A Non-Posnerian Law and Economics Approach”, (1989) 12 Hamline L. Rev. 261.  
On the U.S. computer front, various forms of technological fences (including cryptography) have been recently given a legal “assist” by Congress with the adoption of the Digital Millennium Copyright Act. Although many technological fences are permeable to hacking, the Act makes most such technical bypass unlawful. See 17 U.S.C. Section 1201 et seq.} For instance, even in a legal system without copyright (one might call such a system “copy liberty”), a writer and her authorized publisher could obtain payment through exploiting natural levers such as the advantage of being first in a market or having a reputation for providing authentic, distortion-free texts. But within the first years of the American republic, its Congress decided\footnote{The U.S. Constitution empowers Congress to enact copyright and patent “for limited times” to further “the progress of Science and the useful arts.” U.S. Const. art. I, paragraph 8, cl. 8. Congress enacted its first copyright statute in 1790.} to provide legally enforceable rights of exclusion by enacting intellectual property laws such as copyright and patent. These laws give individual creative persons the right to forbid copying\footnote{Also, American patent law prohibits even duplication of the patented invention that happens to result from completely independent efforts.} of their works.
This right valuably supplements the author’s physical control over her manuscript. It makes publishers and manufacturers willing to pay meaningful sums for the privilege of copying because the exclusive right provides some protection against unauthorized competition from outsiders.

Thus, intellectual property law responds primarily to the second “public goods” characteristic — difficulty of fencing — and does so by altering that characteristic by legal fiat. The law provides fences, which in turn assist the producers in capturing for their own pockets some of the benefits their efforts generate. The system relies on the premise that such enrichment will induce new investment in creative endeavor and that enough new investment will be created — investment that would not otherwise exist — that the value produced by this investment will outweigh the extra administrative and other costs of the intellectual property system.\(^{32}\)

---


What matters is not the absolute level of costs involved in an intellectual property system, but rather a comparison among the costs of the various potential systems. Even a system of no intellectual property rights will have significant administrative costs. Thus, assume there were a legal regime that rejected patent and copyright and recognized only individually-negotiated contracts as a limitation on the public’s ability to copy. In such a context, for example, inventors may spend a good deal of money on policing the secrecy of their inventions, composers may spend a good deal of money on obtaining contractual promises-not-to-copy from people who seek entry to concerts, and the like.

Similarly, it should not be imagined that intellectual property rights are the only way that costly decreases in public access occur. A regime without intellectual property rights will afford far from unlimited access to the public. To use the prior examples: the inventor unprotected by patent may be unwilling to trade information with rival firms, and composers unprotected by copyright may be unwilling to allow radios or television to broadcast their music to general audiences.

For further analysis of the many possible alternatives to copyright law and their costs, see W. J. GORDON, “Merits of Copyright”, loc. cit., note 3.
C. Success of the intellectual property market

An intellectual property market cannot be “perfect,” if by perfection one means a market where everyone willing to pay more than a good’s marginal cost is able to purchase the good and where the producers’ costs are also covered. No matter how low the marginal cost of producing an extra copy of an intangible might be — and it might be as low as zero\textsuperscript{33} — some people who are willing to pay for a copy at that level or above will not have access to it if intellectual property law allows producers to demand payment in excess of marginal cost.

A price above marginal cost is desirable for incentives, for it is hoped that the price will cover research and development expenses and induce other potential producers to make new intangibles. Nevertheless it is clear that a right of exclusion, although the core of intellectual property law, is only a partial response to intangibles’ public goods characteristics, for such law leaves the public unable to take full advantage of the inexhaustibility of intangibles. Instead, the failed promise of inexhaustibility merely exaggerates the deadweight loss that is a cost for all monopolies.

Thus, in addition to providing a right of exclusion, a successful intellectual property system should also be tailored to take as much advantage of inexhaustibility as possible. If something can be copied at no cost, there must be some instances in which allowing free copying will be Pareto-superior.\textsuperscript{34} As will appear, the American legal system makes

\textsuperscript{33} The marginal cost of producing an extra unit of an intangible may be zero, because of inexhaustibility, or it may be some positive sum corresponding to the cost of the intangible’s physical embodiment (such as the cost of the plastic that goes into a phonograph record).

\textsuperscript{34} It might be argued that in cases of true Pareto-superiority, the law would not need to provide a safe-harbor for free copying: if the owners of the patents and copyrights were truly unharmed by the copying (as the notion of pareto-superiority assumes), they would allow the copying to proceed without hindrance.

However, humans are both envious and insecure. The copyright owner might refuse permission not because he is suffering tangible harm, but because he is irritated that other people are getting a free ride or because he has irrational fears about future harm. A society may well choose to consider itself entitled to disregard envy and insecurity as legally relevant harms.
some effort to recapture the lost promise of inexhaustibility. Some of these efforts take the form of “certain” and hard-edged rules, such as the statutory provisions that place copyrights in the public domain after a specified number of years, and some take the form of “uncertain” or “fuzzy” doctrines such as fair use.

II. Comparing “clear rules” with the uncertainty of case-by-case responses

For an intellectual property regime to have even a chance of producing more allocative gain than a “copy-liberty” regime, the intellectual property regime must produce resource packages that are tradable. It must also minimize its deadweight costs and other imperfections. In the following, the article will explore some of the devices that American copyright

In addition, attention has to be paid to defining “harm”. Some cases are easy. For example, assume that the copyist or adaptor is taking customers who would otherwise buy or license from the author, customers who constitute part of the very market at which the author was aiming when she set out to create the work. It would be fairly uncontroversial to call such copying or adaptation “harmful”. Conversely, if the copyist or adaptor of the copyrighted work is serving a market which the original author is incapable of reaching, either directly or by licensing, then the author is not “harmed” by being denied a right over this market. See W. J. GORDON, “Fair Use as Market Failure: A Structural and Economic Analysis of the Betamax Case and Its Predecessors,” 82 Columbia L. Rev. 1600 (arguing that in cases of complete market failure, where administrative costs or other problems would prevent a copyright owner from being able to serve a given group of customers, it is likely to be appropriate to allow “fair use” to those customers when they copy without permission.)

Between these extremes many debatable cases exist for which there will be no obvious answer to the question of what constitutes “harm”. In such cases the law would have to normatively decide what kind of baseline entitlement should be secured to the author. Erosion of a relevant baseline constitutes harm.

35 After a set number of years, all works of authorship go into the public domain. 17 U.S.C. section 302-305 (duration of copyrights). Duration is discussed infra, at text accompanying note 90 and following.

36 The doctrine of “fair use” is discussed further at infra, section IV.

Copyright and Parody

Law employs to bring “success” to the intellectual property market — which means keeping its imperfections to a minimum.

Some of these devices are system-wide responses. Adjudication under sharply defined rules is sometimes accomplished by using a “formal” approach, often associated with the notion of “property owner as sovereign.” The strict liability aspects of copyright have this character, as will be noted below. Copyright law also has another kind of response: employment of open-textured standards which require case-by-case substantive inquiry. Under the latter kind of devices, a court typically makes a substantive judgment as to the desirability of commodifying the resource or behavior at issue, and decides whether (if the resource is ordinarily suitable for buying and selling) its use must be paid for by the particular defendant in the particular context.

A. Definitions: formal property rules as compared with substantive reasonableness standards

In American common law, violation of most property or personal rights will be termed a “tort.” Yet torts themselves tend to fall into two broad categories: intentional torts like battery or trespass, which are ordinarily actionable without proof that the defendant’s specific behavior was socially undesirable, and unintentional torts, which are ordinarily actionable only if the plaintiff shows that the defendant’s

---

38 The other two types of rights that can be sued upon in American civil (non-criminal) courts are those arising out of contract law and restitution.


In restitution a person brings suit on the ground that the defendant has been unjustly enriched and that this enrichment came either at the plaintiff’s expense or by violating some right of his. There is an obvious need for case-by-case adjudication in order to decide what enrichments are “unjust.” Restitution also functions as a remedy following on the violation of other rights (e.g., restitution may require a trespasser to return the profit he made by trespassing on a plaintiff’s land). Restitution is discussed further in this article at Section II, C, beginning at page 86 below.
behavior was negligent or otherwise unreasonable. As a matter of categorization, the first type of tort can be viewed as following a “property” or “formal” model and the second as following a “nonformal tort,” “reasonableness” or “substantive” model.39

In the “property” or “formal” model, the courts defer to the property owner as if she were a mini-sovereign, making no inquiry into whether the owner’s decision to exclude a defendant was proper or improper, or whether the defendant’s use of the owner’s resource was harmful or productive. A classic example under American law is trespass to land. Someone who enters land reasonably but mistakenly thinking he has the right to do so will be liable as a trespasser, as will someone who entered the land out of a pressing (but not life-saving) need for a shortcut.40 Following out the analogy to sovereignty, the primary relevant question in these cases is essentially jurisdictional, inquiring into whether the defendant crossed a boundary over which the owner possessed an

39 I am indebted here to the work of William Powers. Comparing trespass with negligence, for example, Professor Powers notes:

“Ownership embodies a formal methodology, since [...] questions concerning appropriate use are answered wholly by asking whether a proposed use has been sanctioned by the owner. A decision by the landowner [...] concludes legal debate under the ownership model. On the other hand, a duty of reasonable use embodies a nonformal methodology because it makes direct, ad hoc reference to efficiency [or other measures of social desirability]. Under this model, a decision concerning the landowner [...] would depend on a comparison of relative costs and benefits in the specific case.”


40 Admittedly, even in intentional torts American courts may take cognizance of excuses (such as incapacity) and justifications (such as necessity or self-defense). This does not undermine the distinction between intentional and unintentional torts, however. Not only is the burden of proving such intentional-tort defenses typically on the defendant, but these defenses also permit a court far less latitude than does the broad balancing of costs and benefits which a court engages in under a reasonableness inquiry. Thus, a person taking a shortcut through another’s land can take advantage of the necessity defense only if an imminent danger made the shortcut imperative.

41 See W. C. POWERS Jr., loc. cit., note 39.
exclusive right and did so without obtaining the owner’s consent. If so, the defendant has broken the relevant rule and is liable.

By contrast, in the “nonformal tort” or “reasonableness” model, a court does not assume as a prima facie matter that deference is owed to the decisions of the property owner. Instead, the court makes its own substantive inquiry into the desirability of the defendant’s boundary-crossing. Further, the burden of proof will likely be placed on the plaintiff to satisfy the court that the defendant’s behavior was wrongful, and typically wrongfulness will be defined by reference to an open-textured standard. A classic example in the United States is negligence law: in unintentional auto accidents, unless a defendant is found to have lacked “due care,” she will not be required to pay for the damage she caused.

American copyright law follows an uneasy middle course between the more certain “formal” model and the less certain “reasonableness” model. On the one hand, virtually any unauthorized substantial copying of a protected subject matter is subject to a prima facie prohibition. For example, even “unconscious copying” gives rise to liability. Similarly, if a clever plagiarist convinces a magazine publisher that a short story is original, the publisher’s good-faith belief that she had the real author’s permission to print will not help the publisher avoid liability in a copyright infringement suit. This partakes of a formal or “property” approach.

On the other hand, the fact-finder (usually the jury) has both latitude and significant normative responsibility in deciding how much similarity amounts to “substantiality.” Although the quantity copied will be important in the determination of substantiality, the inquiry remains remarkably open.42 In any case that involves other than exact

---

42 There are many verbal formulations as to the meaning of “substantial similarity,” also known as “illicit copying,” but none does very much to define the jury’s task. Consider, for example, a classic case regarding music infringement, *Arnstein v. Porter*, 154 F.2d 464 (2d Cir. 1946), *cert. denied* 330 U.S. 851 (1947). First, the court valuably noted that, “Assuming that adequate proof is made of copying, that is not enough; for there can be
copying, a defendant will probably try to argue that his work is not "substantially similar" to plaintiff's copyrighted work.

In addition, both American and Canadian copyright law permit anyone to copy the "idea" from a copyrighted work, so long as the copyist does not also borrow the work's "expression." Since no firm definition of what constitutes an "idea" has ever evolved, the idea/expression dichotomy is fully dependent on a judge's characterization of what constitutes an "idea". By characterizing something as an "idea", a judge is essentially ruling that it is something that cannot be commodified for purposes of private ownership, but rather should be commonly shared. No certain rule has yet been developed capable of exhaustively defining what kinds of human mental product should or should not be commodified. I suggest no such "rule" is even possible.

Most of the literature on commodification admits the difficulty of the questions raised. Thus, consider SPHERES OF JUSTICE, where Michael Walzer makes a convincing case (if one were needed) that monetary criteria should not rule all spheres. Some goods — such as political office, artistic prestige, basic human dignity — should not be made into commodities...

---

'permissible copying,' copying which is not illicit.” Id. at 472. But then it foundered when it had to distinguish rightful from illicit copying:

"The proper criterion on that issue [...] is whether defendant took from plaintiff's works so much of what is pleasing to the ears of lay listeners, who comprise the audience for whom such popular music is composed, that the defendant wrongfully appropriated something which belongs to the plaintiff.”

Id. at 473.

Aside from the reference to the lay (non-expert) audience, this formulation offers no more than vague references to quantity ("so much"), to market value, and to a conclusory notion of wrongfulness.


44 See, e.g., Margaret Jane RADIN, Contested Commodities, Cambridge, Harvard U. Press, 1996 (particularly her discussion of the "double bind.").
for purchase and sale. But Walzer is less clear about what institutions should mediate society’s difficult choices over what behaviors and resources should be placed in the market sphere, and what behaviors and resources should be governed by non-market criteria.

One institutional possibility is for the legislature to set out particular rules, and then for that same legislature to change the rules as circumstances change, or as particular norms come under pressure. This has occurred. Thus, United States rules on conscription have changed and changed again, including at several points a rule of commodification: under the nation’s first conscription law, a draftee could lawfully “hire a substitute in his place.” Under a succeeding statute, Congress “set a flat fee of $300 for exemption from induction.” Later, of course, such monetary exemptions became anathema, and successive rule changes gave other criteria (consider, e.g., educational deferments from the draft) their chance at being a governing norm.

But such rule changes take time to implement. Arguably, the technology that drives modern copyright markets changes too fast — and arguably, the norms concerned are so subtle and the dynamics of power politics too insensitive to the public interest — to rely on legislative rule-making and rule-changing as an optimal route for defining the boundaries between the market and non-market alternatives.

---

45 Cf. Michael WALZER, *Spheres of justice: a defense of pluralism and equality*, New York, Basic Books (arguing that money should not to dominate all spheres of life, and that some way must be found to distinguish the monetary sphere from spheres where other values should predominate.)


47 *Id.*, at 160.

48 Jessica Litman’s studies of the legislative process in the recent U.S. Copyright arena suggest that it has served private interests much more consistently than it has the public interest. See, e.g., Jessica D. LITMAN, “Copyright Legislation and Technological change”, (1989) 68 *Or. L. Rev.* 275; Jessica D. LITMAN, “Copyright, Compromise, and Legislative History”, (1987) 72 *Cornell L. Rev.* 857. This is likely to lead to over-commodification.
A perhaps preferable institutional possibility is for the legislature to adopt vague standards. As has been often noted, vagueness delegates to other decision-makers (largely judges) the power to fill in the details. Congress chose this route when it declared that “ideas” cannot be owned in copyright law.\textsuperscript{49} without ever defining what constitutes an “idea”. As a result, shifting norms and facts have fed into judges’ conceptions of what should be declared an “idea” and thus placed outside the bounds of copyright ownership.

An example of judges’ shifting instincts on this issue can be seen in their treatment of whether “compatibility standards” in the computer field can be protected by copyright. At one point, a defendant was ridiculed for claiming that he should be entitled to copy whatever was essential to the “idea” of producing a computer capable of running standard Apple programs. That was not an “idea”, said the Court.\textsuperscript{50} More recently, a court to the contrary indicated that a defendant is free to copy “elements that might have been dictated by external factors,”\textsuperscript{51} including elements needed to achieve compatibility with “other programs with which [the defendant’s program] was designed to interact.” Such elements were not ownable expression. Presumably the courts were influenced by shifts in the computer industry, and a growing recognition of the social benefit to be gained from fostering network externalities.\textsuperscript{52}

\textsuperscript{49} 17 U.S.C. Section 102(b).

\textsuperscript{50} The court wrote: “Franklin may wish to achieve total compatibility with independently developed application programs written for the Apple II, but that is a commercial and competitive objective which does not enter into the somewhat metaphysical issue of whether particular ideas and expressions have merged.” \textit{Apple Computer Inc. v. Franklin Computer Corp.}, 714 F. 2d 1240 ( 3rd Cir. 1983), cert dismissed, 464 U.S. 1033 (1984). The Apple II operating system was a competitor of the now-dominant DOS system.

\textsuperscript{51} \textit{Computer Associates International Inc. v. Altai Inc.}, 982 F.2d 693 (1992). Note, however, that this was not explicitly a holding on what was or was not an “idea”.

\textsuperscript{52} Basically, if a network grows more valuable to each member when additional members join, then each new member is said to confer an “external benefit” on the existing membership. There is a growing literature applying network externality analysis to the computer field.

For example, if I use a WINDOWS operating system, I will be better off if many other people also use the same system. The more people who use WINDOWS, the more application programs will be written for WINDOWS.
Even in cases where a defendant has committed “substantial” copying of “expression”, he remains able to call upon another important but “muddy” standard in American copyright law: the doctrine of “fair use”. A defendant will have no liability for making a copy which is “fair” in her particular circumstances.

The fair use defense essentially draws the court into deciding the social desirability of the defendant’s copying.

and thus for me; also, the more people who use WINDOWS, the more easily I can communicate with other computer users. Each additional person who buys and installs a WINDOWS operating system thus confers a benefit on me.

53 The imagery of “mud” for standards and “crystal” for rules originates in C. M. ROSE, loc. cit., note 10.

54 See 17 U.S.C. Section 107. As originally enacted, section 107 provided as follows:

“107. Limitations on exclusive rights: Fair use

Notwithstanding the provisions of section 106 [which set out a copyright owner’s exclusive rights of reproduction, adaptation, public performance, and the like], the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include -

(1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;

(2) the nature of the copyrighted work;

(3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and

(4) the effect of the use upon the potential market for or value of the copyrighted work.”

17 U.S.C. Section 107 (1976). The section was recently amended to make clear that the unpublished status of a work should not be determinative.

Although the fair use doctrine appears in the Copyright Act, that statute declines to set out any definite strictures.\textsuperscript{55} Formal line-drawing is rejected. As the legislative history recounts, “[S]ince the doctrine is an equitable rule of reason, no generally applicable definition is possible, and each case raising the question must be decided on its own facts.”\textsuperscript{56} At one point it seemed that the Supreme Court was willing to bring a bit more certainty to the doctrine, for one majority opinion seemed to declare that any commercial use is presumptively unfair, or at least that one of the fair use factors would be presumptively resolved against the defendant.\textsuperscript{57} More recently, however, in the context of a commercial parody, the Supreme Court reinterpreted its prior statement and wrote that “no such evidentiary presumption is available.”\textsuperscript{58} Similarly, when a Supreme Court opinion\textsuperscript{59} was applied as if it had adopted a rule making fair use unavailable for unpublished works, the Congress responded by amending the fair use statute to specify that the unpublished nature of a plaintiff’s work was only one factor among many.\textsuperscript{60} Thus, both the Court and Congress have

\footnotesize
\textsuperscript{55} The legislative history of section 107 indicates that, despite the statutory recognition accorded fair use, the nature of the doctrine remains to be defined by case law: “The bill endorses the purpose and general scope of the judicial doctrine of fair use, but there is no disposition to freeze the statute [...] “H.R. Rep. No. 94-1476, at 66 (1976), reprinted in 1976 U.S.C.C.A.N. at 5680 [hereinafter House Report]. See also S. Rep. No. 94-473, 1st Sess. 62 (1975) [Senate Report]. The courts have recognized their freedom to continue the development of fair use doctrine. See, e.g., Triangle Publications Inc. v. Knight-Ridder Newspapers Inc., 626 F.2d 1171, 1174 (5th Cir. 1980) (“Congress made clear that it in no way intended to depart from Court-created principles or to short-circuit further judicial development [...]”).


\textsuperscript{57} \textit{Sony}, 464 U.S. at 451. The fourth fair use factor is “the effect of the use upon the potential market for or value of the copyrighted work,” sec 107(4). In regard to that factor, the Court indicated that a likelihood of future harm to a plaintiff could be presumed from the commerciality of a defendant’s use. \textit{Id.}

\textsuperscript{58} \textit{Campbell v. Acuff Rose Music Inc., supra}, note 17, at the last page of the opinion. The Court at one stage of the \textit{Campbell} opinion distinguished \textit{Sony} on the ground that \textit{Sony} involved “mere duplication for commercial purposes” while \textit{Campbell} involved a transformative use, namely a parody. But the overall tenor of the opinion went beyond that distinction to reject the notion of certain and sharp rules in fair use.

shown a deliberate preference for an equitable standard over specific rules.

B. Copyright’s unusual mid-range status

At first blush, it is surprising to find American copyright law placing at the virtual center of a plaintiff’s case the distinctively nonformal principles of “substantiality” and “fairness.” After all, copyright is a form of property, and copyists always act volitionally and deliberately. (For example, they know they are publicly performing, or using the photocopy machine, or playing music on their guitar, even if they don’t know that they are copying someone else while doing so.) It seems most logical that such a volitional trespass should be treated under a formal rule, as are other non-accidental violations of property rights. The mere act of nonconsensual copying is arguably like the mere act of stepping onto someone else’s land without permission, and arguably should give rise to similar liability. So why is this not the case?

Economics does not yield an immediate answer. The classic article by Calabresi and Melamed61 tells us that intentional takings of property are prima facie wrongful because people should not depart from the market without a strong justification. Their article tells us that accident law uses a “reasonableness” inquiry because, in accidents, such a justification is present: the participants cannot bargain with each other in advance. Before a driver chooses to drive down a particular street at 35 miles per hour, neither she nor the pedestrian with whom she may accidentally collide on that street, has any reason to know that they need to deal with each other. In the presence of such complete market failure, we are told, the court “mimics the market” through a negligence inquiry, trying to determine whether the parties behaved efficiently and imposing liability accordingly.

---

60 17 U.S.C. Section 107, last sentence.
But in copyright the “substantiality” inquiry and the “fair use” doctrine are both available regardless of whether the copyright owner and the copyist have complete knowledge of each other’s identity and are otherwise able to bargain. Thus, in the most recent “fair use” case to reach the United States Supreme Court, the parties had the requisite knowledge and transaction costs were minimal. Defendant, the rap group “2 Live Crew,” had in fact offered the copyright owners compensation in exchange for permission to make a parody of their song, “Oh, Pretty Woman.” The copyright owners simply refused to give the rap group permission.

This hardly looks like market failure — the two parties were virtually face to face. Yet the Supreme Court indicated that fair use might nevertheless be available to shield the makers of the parody from liability.62 How, then, can such a doctrine be squared, either with usual American patterns of tort and property law, or with economic notions of market failure?

The answer lies in the imperfection of intellectual property as a response to the “public goods” problem. At least in the absence of perfect price discrimination,63 obtaining adequate incentives for production will necessarily involve a price that is set above marginal cost, and thus a quantity produced that is below the quantity that would be produced by a competitive market. Market imperfection is present in even the most pristine copyright transaction. The issue is how legal institutions cure the imperfections caused by excludability without losing the benefits excludability brings.

62 See Campbell v. Acuff Rose Music Inc., supra, note 17. The case was remanded for further proceedings in light of the Supreme Court opinion that, inter alia, made clear the copyright owner’s refusing to license should not weigh against a fair use finding. See id. (parodists are unlikely to be able to obtain consents). See also id. at 585 n.18 (“we reject Acuff Rose’s argument that 2 Live Crew’s request for permission to use the original should be weighed against a fair use finding.”).

Or, one can put the same matter in non-economic normative terms. Copying is not necessarily wrongful. It is how we learn, it can be both harmless and beneficial, and it is the essence of having a common culture. Therefore it would be absurd to make all copiers prima facie liable as infringers.

The trick is to find some means to distinguish wrongful from fair copying. Some of those means are case-by-case, like “substantial similarity” and “fair use.” Some are system-wide rules, like the provisions limiting the duration of copyrights and patents.

But the use of vague standards is probably what stands most in need of explanation. The employment of these standards is, in my view, largely due to the lost promise of inexhaustibility, coupled with the speed of technological change that characterizes the copyright industries in the last century. Since copying of an intangible is often harmless, a “fair use” and “substantial similarity” standard permit socially useful experimentation. And since such experimentation may be blocked by transaction costs, and these costs can change quickly, “fair use” allows the courts to adapt.

---

64 See Benjamin KAPLAN, An Unhurried View of Copyright (1967).

65 If copying is harmless, allowing copying produces a Pareto-superior result: no one is hurt and the copyist and her customers gain.


67 W. J. GORDON, loc. cit., note 34 at 1656-1657; also see American Geophysical Union v. Texaco Inc., 60 F 3d 913 (2d Cir 1995), cert. denied 516 S Ct 1005 (1995) (photocopying held not fair use because, inter alia, licensing through the Copyright Clearance Center would have been feasible).

Note, moreover, that transaction cost barriers between copier and copyright owner are only one of many forms of market failure relevant to fair use. For example, uses such as criticism generate positive externalities, and this is one reason why a critic’s extensive use of quotation is potentially entitled to fair use treatment. “Fair Use as Market Failure,” at 1630-1631. Further, “fair use” is also used to address uses that should not be owned (or commodified) at all, id. at 1631-1632, such as uses that implicate non-monetizable free speech concerns.

The importance of providing a safe harbor for critics may be obvious and stable enough to be recognized in a rule, as the Canadian “fair dealing” statute has done. But is is not so easy to reduce to a rule the many ways in which the
All that being said, however, it is still true that Copyright law places on the public a duty to refrain from certain uses of others’ labor, while Restitution takes the opposite starting point. It will be useful to employ basic economics to understand why Copyright law begins with a presumption of liability (subject to an equitably-expressed exception for “fair” uses) while Restitution begins with a presumption of no liability (subject to an equitably-expressed exception for “unjust” enrichment).

C. Comparing Copyright and Restitution

Restitution doctrine provides that persons whose labor makes others better off will ordinarily have no legal recourse if they labor without advance agreement. Yet intellectual product producers can sue to obtain payment for the “fruits of their labor” from copyists who never agreed to pay. This has led some observers to view some forms of intellectual property as unjustifiable. Since Restitution law contains no presumption that there should be recovery for benefits generated, it forms a useful contrast with copyright.

“fair use” doctrine can be used to investigate issues such as non-monetizability of the interests at stake in a particular case. When the question before the court is the normative inapplicability of the whole market apparatus, a flexible standard is (I argue) the best way to proceed, at least until a societal consensus has formed.

Note that the issues embraced by “fair use” cases can include issues of commodification. This essay broached those issues earlier, in relation to the statutory declaration that “ideas” cannot be owned under United States copyright law. 17 U.S.C. Section 102(b). In the middle of the century, the overlap between “fair use” and the non-ownership of “ideas” was visible in the language courts used, for a defendant who used a copyright owner’s “ideas” was said to be engaged in a “fair use”.

68 Thus Murray Rothbard argues that intellectual property is legitimate only to the extent that it can be analogized to consent. See Murray N. ROTHBARD, Man, Economy and State. A Treatise on Economic Principles, Auburn, Ludwig von Mises Institute, 2nd ed., 1993, 652-660 (1962).

69 Under American copyright law, the work’s creator has a right to exclusively control the rights of reproduction, copying, adaptation, performance and the like, see 17 U.S.C. Section 106. She can extract monies and obtain injunctions when someone does these things without permission.
First, consider a homely example to illustrate the different treatment a laborer without a contract will receive under the two areas of law. First, imagine someone paints the roof of a building while its owner is away. After the owner returns the painter presents himself at the door and says, “pay me for this wonderful benefit I have given you,” pointing at the new paint job which (we will assume) has increased the value of the building. In such a situation, the building owner is entitled to say something quite rude. Next, in contrast, imagine that the homeowner, using her own sweat and paint, does a mural on one of her building’s exterior walls, copying onto it a painting which has a valid copyright. Perhaps the mural increases the value of the building more than a new coat of pure white paint would have; perhaps the mural is an eyesore. In either event, if the owner of the copyright comes to the door and says, “Pay me or I’m going to sue you for a very large amount of money,” the building owner had better be very polite.

In neither situation has the building owner agreed in advance to pay for use of the other’s resource. Yet the photographer’s labor embedded in a visual pattern must be paid for, and the painter’s labor embedded in new roof pigment need not be.

Do not jump too quickly to say that one of the pigments embodies property while the other merely embodies labor. Pretend for a moment there is no property in pictures, and that we have to return to basic common-law techniques to determine whether or not the author of the copied picture should be paid.

To prevail in Restitution, persons whose voluntary actions provide benefits to others must ordinarily show one of a few very narrow justifications for departing from the market: mistake,70 coercion,71 request,72 or a narrow range of exigent situations, such as danger to life and health.73 Even then, their ability to recover will often be further restricted by the courts’

70 Restatement of Restitution, supra, note 2, Sections 6-69.
71 Id., Sections 70-106.
72 Id., Sections 107-111.
73 Id., Section 112.
desire to be sure that the defendant really was benefited and that forcing him to pay or disgorge will not leave him worse off than he would have been in the status quo ante.\textsuperscript{74} Similarly, the Restatement of Restitution is not hospitable to persons who generate benefits as a by-product of self-serving activity. Thus, the Restatement states that: “A person who, incidentally to the performance of his own duty or to the protection or improvement of his own things, has conferred a benefit upon another, is not thereby entitled to contribution.”\textsuperscript{75} For example, a mine owner whose drainage efforts clear both her mine and her neighbor’s mine of waters is not entitled to contribution from the neighbor.\textsuperscript{76}

A person who writes a book and publishes it is certainly operating in the furtherance of his or her own interests. Except as to someone who has bargained with the author for production of the work (such as a patron, granting agency, employer, or contract-publisher), the author is a sort of volunteer. She is voluntarily taking the risk that putting her product on the market will bring her a profit. When a book is mass-marketed, many strangers will come across it. If a stranger makes copies of the book for sale, copyright law will give the author a right of action against the copyist even if the author “volunteered” to send the work into the stream of commerce. Since that right of action will be available whether or not the copyist had a contract with the author promising to refrain from copying, and whether or not the copyist’s actions harm the author,\textsuperscript{77} it is clear that, under copyright law, a unilateral transfer of “benefits” will trigger liability.

\textsuperscript{74} See, e.g., \textit{id.} Section 40, cmt. b, at 109.

\textsuperscript{75} \textit{Id.}, supra, note 43, Section 106. There are situations in which protecting one’s own interests does not bar restitution, but these tend to be associated with coercion, as where a property owner discharges another’s duty when that is the only way to prevent a third party from lawfully taking the property. \textit{Id.}, Section 103.

\textsuperscript{76} \textit{Id.}, Section 106, illus. 2. The result in this situation may vary if the neighbor can be said to have “freely accepted” the benefit. See Peter BIRKS, \textit{Introduction to the Law of Restitution}.

\textsuperscript{77} Sometimes the absence of harm may make it easier to obtain fair use treatment, however. See the fourth factor in 17 U.S.C. section 107, set out at footnote 54.
There are many reasons for the difference between the two fields’ basic rules; one difference obviously lies in the active or passive role of the person using the benefit. That is not simply an issue of autonomy. Activity or passivity also has implications for market formation. To see this, consider what results would follow if a legally enforced right to payment were given to the claimants in the two situations.

In the Restitution context, the active parties are the benefactors, the volunteers. Systematically allowing volunteers to sue for the benefits they have given would reduce their desire to make contracts with those who might want their services. Admittedly, the people who are the best at painting houses would have no desire to sneak around and do it behind the backs of their customers. However, the people who make messy jobs of it would probably start to paint and then ask for money after the fact — and could do so in disregard of whether the building owner preferred a different supplier, thus ruining the market even for those who would otherwise be willing to make contracts. In the volunteer context, then, a rule that encourages contract formation — and thus market formation — is a rule that denies to the benefit-generator (the potential volunteer) any right of recompense independent of contract. If a volunteer thinks the law will not give restitution, then she will seek to make a bargain by asking the potential recipients for contributions before the project begins.

---

78 Even when there is a market failure in the restitution context, so that the potential benefactor and the potential recipient are unable to identify or negotiate with each other, there are only very few circumstances in which payment is ordered through the courts — mistake, request, coercion and a narrow range of emergencies justify recovery. This narrowness of recovery in the restitution context reflects the fact that if we give the benefit-generator a right to legally enforce recompense, it would tend to erode markets. See LEVMORE, loc. cit., note 38.

79 Many examples exist of “internalizing benefits” by contract. Thus, in many shopping malls, where small stores are likely to benefit from the propinquity of large department stores that draw masses of customers, the small stores may be willing to pay extra rent to subsidize the larger stores’ entry. Something like this also happens in oil exploration: neighboring lessees will learn a great deal about whether or not it is worthwhile to drill under their own land from the results of their neighbor’s drilling. So “dry hole contribution agreements” have come into being: contracts by which the neighbor who stands to benefit from the information agrees to pay a share of his neighbor’s drilling costs should the hole come up dry.
In the intellectual property context, the likely impact of a right to recover is quite the opposite. This occurs largely because the identity of the active party — and thus the party who has superior access to information, who is otherwise better able to enter transactions, and who is better situated to respond to the law’s messages — is different there.

In the volunteer context, the recipient may be ignorant until the deed is done. It is the benefactor — like the house painter — who has the greater access to information; he knows where and when he will act. In restitution, the rule of law that speaks to this active party and encourages him to seek out consensual market arrangements is therefore a rule of “no monetary recovery without contract.” In the intellectual property situation, by contrast, the recipient-copyist is the active party: he can better initiate the transaction. After all, the copyist knows what he is copying, whereas the plaintiff-owner may be hundreds of miles away and have no idea copying is being contemplated. The copyist will also find it fairly easy to identify the author or copyright owner from the by-line, while the copyright owner has no such source of information.

Similarly, assume that landowners are likely to benefit from a venture like a resort complex locating nearby, but the resort is so expensive to build that it cannot afford many externalized benefits — e.g., it will come to the area only if it is subsidized by the existing landowners or can itself capture most of the benefits generated. Even in such a case there may be no need to allow the resort complex to sue the benefited owners, after the fact, in restitution. The developers can try to persuade these neighbors, in advance, to pay them something to encourage them to build nearby. Admittedly, there could be hold-out problems and other strategic maneuvering making this difficult. So the active party has another option: the owner of an attraction could quietly buy the land on which the beneficial spillovers will fall. This is apparently what the Disney organization did with Epcot and Disney World: it bought up surrounding land and built on it enough hotels and restaurants to capture much of the benefit their tourist attraction generates. Where this is possible, benefits are again internalized without the need for restitution suits.

Even if the author can identify the potential copyists, she faces strong strategic behavior problems in making them pay. A baseline rule that denied copyright — that gave a benefit-generator no recovery unless he had a prior contract with the copyist — would leave the party with the best ability to contract (the copyist) with little motive to do so. He would probably prefer to free ride. Therefore, the law has to give the benefit-generators (the copyright owners) a right of recovery independent of contract.

Both restitution and intellectual property law give the party who has the information and ability to internalize the incentive to do so. The party simply happens to be different in restitution than in intellectual property. In each case, legal rights are arranged to facilitate the consensual transfers of resources.

81 In a world without intellectual property rights, an author may want to bargain with her audience for payment, but the audience is likely to be a wide and uncertain one, and the benefits are those that will flow from an as yet undisclosed intellectual product. Even if the author could somehow identify all the potential recipients, it would be expensive and awkward to reach simultaneously all of the persons who will eventually want access to the work. Even if this were possible, what would happen when the creator tried to negotiate for a payment from them all in exchange for disclosing the work? Many of those audience members might be tempted to hold back in the hope that others’ monies would be sufficient to draw the work into the marketplace where they could then make a cheap copy. The larger the group of potential purchasers, the better the odds on the gamble may seem. Also, the work’s contents are largely unknown at this stage; the less certain the benefits, the less seems to be risked if the gamble does not pay off. Good odds in favor of winning, and low perceived cost in the event of a loss, make the gamble very tempting. If enough people take this gamble in the hope of taking a free ride, the requisite funds may not be forthcoming. “Chicken,” “prisoner’s dilemma,” and other free-rider games illustrate analogous dynamics.

The presence of a publisher does not much alter the desirability of granting intellectual property rights to resolve potential bargaining stalemates. Admittedly, in a world without intellectual property rights the author may find it easier to deal with a publisher than with an undifferentiated audience (only one party, low transaction costs), but then the publisher must deal with the audience. The author’s problems with information, transaction costs, and free riders would simply be passed on, one step further down the line. How much would a publisher pay for a book that could be lawfully copied by all comers once it appeared on the market? Unless the publisher has a lead-time advantage or some other sort of real-world clout that can discourage copying, the rate the publisher would offer the author in such a world might be too low. If the anticipated rate of payment is low, otherwise-desirable works may not be created.
The party best positioned to alter the use to which a resource is put is required to do so by a systemic choice of a liability or no-liability rule.

The centerpoint of intellectual property liability appears in the copyright and patent statutes as a grant to the proprietor of rights to exclusive use. Note that the rule setting up liability appears to be a sharp-edged and certain grant. As with all sharp-edged rules, it may be over-inclusive. Thus, although the rationale of the rule depends on the potential printer or copyist being an active party who can knowledgeably seek out bargains, copyright law also makes a publisher liable if, in good faith, she prints a copyrighted work which a plagiarizer has submitted. The law might even make a passive recipient liable. For example, assume that a painting duplicating a copyrighted work is put on the exterior wall of a house not by the building’s owner, but by a skilled prankster. In that case, the householder is a passive recipient, rather than an active user (as is more typical in intellectual property cases). Nevertheless, he may be liable. Assume for example that the reproduction draws attention to the house so that the householder can sell the building for an amount in excess of the value of otherwise similar homes. If the owner in fact does sell, he will likely be

---

82 This is a variation of the phenomenon Dean Calabresi referred to as looking for the “best briber.” Guido CALABRESI, The Costs of Accidents; a legal and economic analysis, New Haven, Yale University Press, 1970.

83 Markets can yield efficiency only where resources can practicably be transferred to their highest-valued uses. See R. H. COASE, “The Problem of Social Cost”, (1960) 3 J. L. & Econ. 1.

84 For copyright law in the U.S., see 17 U.S.C. Section 106 (giving exclusive rights over the reproduction, distribution, public performance, adaptation, etc., of the copyrighted work of authorship); for patent law in the U.S., see 35 U.S.C. Sections 101 et seq. (giving exclusive rights over making, using, selling, offering to sell, or importing, the patented invention.)
guilty of infringing the copyright owner’s exclusive rights over the distribution of copies. Such over-inclusion can be costly.

III. Systemic intellectual property rules to minimize market imperfections

Rules are not only used to enhance tradability. They are also used to distinguish between areas where legal protection is desirable and areas in which it is not. Industries may exist where perhaps the need for government to provide exclusion rights is less (because there is no significant market failure under copyright), or uses which are more expensive to restrict, or other areas where the costs of copyright might outweigh its value. Thus, copyright distinguishes among subject matters. Not all beneficial products of human ingenuity are capable of being owned. Differing treatment exists for computer programs, musical compositions, literary works, architecture, recorded oral presentations — each has some specialized rules in the statute.

Further the Copyright Act grants ownership only in works of authorship fixed in a tangible medium of expression, and even among such works the act denies protection to whole classes (such as ideas, and typographic designs) which, presumably,

---

85 Under section 106(3), only the owner of the copyright has the right to distribute copies. Although the first sale doctrine (see section 109) provides an exception that allows the owners of lawfully made copies to distribute them without liability, the house owner in our example does not own a copy that was “lawfully made”. Therefore he could not take shelter under the first sale doctrine, and his sale of the house would constitute an unlawful “distribution” of the copy painted upon it.

Public display is also one of the copyright owner’s exclusive entitlements. 17 U.S.C. Section 106(5). Even without a sale, it is possible that an irritated copyright owner could bring a successful suit merely for the house owner’s continued “public display” of the work. However, I suspect that, in such a case, a court might read into the copyright statute a requirement that the defendant act volitionally. (A suit premised on sale of the house — sale being a volitional act — would not be as vulnerable.)

86 See, e.g., 17 U.S.C. Sections 102, 103.

87 Section 102(6) denies copyright to, inter alia, ideas and processes. As for typographic designs lacking protection, this is a matter of legislative history; the statute itself does not explicitly mention typography. See House Report, supra, note 55, at 53-57. The reason protection was not given to typographic design may have been a fear that such protection might be misused by
would prove more costly than beneficial to propertize.\textsuperscript{88} Similarly, copyright does not give owners of copyright a right to control all uses of their works. Rather, they have control over certain kinds of enumerated uses — those uses whose control is desirably centralized in one entity. Thus, for example, the composer of a song has exclusive rights over reproduction and public performance, but not over private performance.\textsuperscript{89}

There are also system-wide rules that work to decrease the system’s costs, and to take better advantage of intangibles’ inexhaustability. The key example here is that of duration.\textsuperscript{90}

To see this, we must backtrack to consider a conceptual matter. At one point, lawyers foundered when asked how to assess, even conceptually, the value of an intellectual property system. The empirical questions are hard enough, but there appeared to be a paradox where deadweight loss was concerned. True, the exclusive right that the copyright or patent owner receives from the law confers a kind of monopoly power (of varying effectiveness, depending on the competing intellectual products available to the audience.)\textsuperscript{91} Also true, this monopoly power can then cause deadweight loss as the intellectual property owner imposes a price above marginal cost and the quantity effectively available to the public is reduced. But lawyers were hard put to assess the significance of this deadweight loss, because the work to which access was being reduced by the intellectual property law might never have come into existence without that very law.

---

\textsuperscript{88} Industry pressures undoubtedly play a role here as well.

\textsuperscript{89} For more on the role of copyright’s particular provisions, see W. J. GORDON, “Merits of Copyright”, \textit{loc. cit.}, note 3, and W. J. GORDON, “Fair Use as Market Failure”, \textit{loc. cit.}, note 34, at 1605-1615.

\textsuperscript{90} In the discussion of duration that follows I am indebted to the work of Stanley Liebowitz. See, \textit{e.g.}, S. J. LIEBOWITZ, \textit{loc. cit.}, note 32, at 183-188.

\textsuperscript{91} That is, the person who owns copyright in a particular book will have a monopoly over that book, but not over competing titles by other authors.
The way out of this apparent paradox was to make a conceptual distinction between those works which needed the intellectual property law to induce their authors to create them, and those works which did not. As to the former items, any production is due to the legal regime’s grant of intellectual property rights. As to the latter items, the intellectual property law merely functions as a restriction.92

Thus, one would credit the intellectual property system with all the value of the works that would not have come into existence without intellectual property rights. As to these works, any production of the item would count as a positive value, and restrictions on quantity would be irrelevant. Then, from this aggregate positive value would be deducted the deadweight loss in markets for works that would have been produced even in the absence of intellectual property rights.93

The value of the intellectual property system is the net of these two numbers. As Professor Stanley Liebowitz has made clear,94 imposing system-wide durational limits — limiting how long particular types of intellectual property rights will last — can serve to maximize this net value.

First, as duration increases, the number of new works attributable to an additional period of protection will grow smaller. The usual law of diminishing marginal utility would seem to govern; as the duration of a copyright or patent is made longer and longer, the incentive effect of additional length is likely to decrease.95 To illustrate: extending the duration of

---

92 The discussion here puts aside what Edmund Kitch calls “prospect effects,” namely, those positive effects on ease of exploitation that can occur when property rights are centralized in one entity. See Edmund KITCH, “The Nature and Function of the Patent System”, (1977) 20 J. Law & Econ. 265.

93 S. J. LIEBOWITZ, loc. cit., note 32.

94 His graphical representation is particularly helpful. See id. at 187.

95 This point is probably made most wittily by Lord Macaulay, in his speeches before the British parliament protesting their extending the duration of copyright. Lord Macaulay argued that while copyright might be necessary to ensure a “supply of good books,” the monopoly that it imposed was at best a necessary evil. “For the sake of the good we must submit to the evil; but the evil ought not to last a day longer than is necessary for the purpose of securing the good.” Thomas MACAULAY, “Speech Before the House of
copyright from one year to five is likely to so increase the expected rewards of writing a book or designing a poster that the increase will induce some new works to be made which would not otherwise be created. By contrast, extending the duration from 101 years to 105 is not likely to have as strong an effect, so such further extensions will bring less to the “plus” side of the ledger.

Second, as duration increases, more and more works will not “need” the extra years to come into existence. As the reader will recall, years of protection that are not needed for incentives constitute unnecessary restrictions, and as to them “deadweight loss” should be counted on the debit side of the social ledger. For example, works which were called into existence by the promise of a 56-year reward are “pluses” to be credited to the copyright system for only their first 56 years. The copyright system’s grant of exclusive rights for years 57 and following would deserve no credit for those works’ creation. Equally importantly, any deadweight loss in markets in years 57 and following becomes a cost attributable to copyright.

Thus, as duration grows longer, the incentive value of an added durational restriction grows less, and the deadweight loss grows larger. The economic goal in choosing a durational limit is to maximize the difference between the two measures, incentive value and deadweight loss. Thus, duration can be custom-designed, providing a set of categorical sharp rules that limit the periods of protection for intangibles protected by a given regime, such as the seventeen to twenty years that U.S. law gives to utility patents, as compared with the fourteen years of design patents, the “life plus seventy years” that inheres in most copyrights, or the ten years of protection applicable to semiconductor chip mask works. These various

---

98: 17 U.S.C. Sections 302-305. There is a different duration for works made for hire, etc.

Commons” (Feb. 5, 1841), in 8 The Works of Lord Macaulay 195, 199 (Lady Trevelyan ed. 1866) (discussing a bill which would have extended the duration of copyright protection).
systemic limits can function to minimize the cost attributable to deadweight loss.

IV. Using standards to minimize market imperfections: turning to the fair use doctrine and the question of parody

Sometimes markets do not evolve for a particular creative work or use — say, for example, that bargaining is impeded by problems such as externalities, or high transaction costs in identifying or communicating with the copyright proprietor. If the copyright laws prohibited copying in that area it could simply be preventing copying without yielding creators any monetary advantage. That would be undesirable. Not only would copyright then fail to perform its primary function, but if users cannot reach market deals with creators, copyright would impose more costs and generate less benefit than would a regime without copyright. For though incentives may be low in a world without copyright, at least copyists and other users would have access to whatever works happened to be created; by contrast, in a world where there is copyright but no markets, incentives are low and the public has no access. Therefore, as discussed in the initial sections of this article, the ability of users to form markets is crucial to copyright’s economic mission of encouraging the production and use of new work.

This observation has implications for policy in individual cases. If a defendant faces market failure in the face of copyright, then, in his case, the economic foundation for copyright has crumbled. That is a good argument (if not a complete one) for not enforcing the copyright against him. Further, it can be argued that “fair use” has evolved as an equitable response to market failure, to ensure that socially desirable uses will not be blocked.¹⁰⁰

¹⁰⁰ I have advanced this argument in W. J. GORDON, “Fair Use as Market Failure”, loc. cit., note 34, 1614-1615, 1627-1641. For further development, see id. at 1614-1657 (proposing a 3-part test for fair use, and comparing such test with the case law results) and W. J. GORDON, Private Censorship, loc. cit., note 1, 1042-1043 (1990) . See also, e.g., W. M. LANDES & R. A. POSNER, loc. cit., note 32; Sheldon LIGHT, “Parody, Burlesque and the Economic Rationale for Copyright”, (1979) 11 Conn. L. Rev. 615.
For example, consider photocopying by individual scholars. The transaction costs in contacting a copyright owner for permission to photocopy might well outweigh the benefit the scholar expects to reap. In such a case, enforcing the copyright would merely eliminate the photocopying, rather than generate any license fees for the copyright owner. In such an event, granting “fair use” treatment to the scholar will not impair the copyright owner’s potential income stream, and will allow a socially beneficial use to go forward that the transaction costs barrier would otherwise have blocked. High transaction costs are, of course, a classic cause of market failure.

The market failure approach is consistent with the great bulk of “fair use” precedent, and in recent years this sort of argument has even found its way into the courts’ explicit arguments. For example, in a recent fair use case involving corporate photocopying of scientific journals, the courts clearly had a market failure model in mind. The opinions of both the District Court and Court of Appeals discuss what economists identify as “transaction costs,” and examine the extent to which the defendant’s employees — if required to stop photocopying — could find other avenues through which to obtain the desired material.

Similarly, in the most recent fair use case before the Supreme Court, the opinion indicated that “fair use” can be justified in part as a response to situations in which copyright owners are unlikely to give permission at virtually any price.

---

101 That is, even if the scholar were willing and able to pay whatever price the copyright owner demanded, the scholar might not be willing to both pay that price plus bear the time delay, hassle, and secretarial costs involved in securing a permission.

102 Note that this analysis is dependent upon the relative size of the applicable transaction cost barrier. If a clearinghouse or compulsory license system exists which reduces the transaction costs, then the scholar may not require fair use treatment in order to allow her use to go forward.


105 See Campbell v. Acuff Rose Music Inc., supra, note 17. In assessing the plaintiffs’ claim that the parody would impair their potential market, the Court
This position, advanced in a case involving a song parody, might strike the reader as inconsistent with the usual assumption of neoclassical economics that one must take preferences as a given.

If one takes this assumption seriously — it is sometimes known as the assumption of “consumer sovereignty” — then it seems the Court should have accorded to the copyright owner’s desire not to be parodied as much respect as any other value. After all, in theory, an unwillingness to sell or license merely indicates that the potential buyer/licensee is not the highest-valued user. And in many countries, such as Canada, an unwillingness to allow one’s work to be copied in a distorted manner is given more, rather than less, respect than a refusal to sell motivated by ordinary commercial reasons.106

---

106 Canada’s moral rights statute is complex. Analysis can fruitfully begin at Copyright Act, supra, note 19, section 14.1 (1985) (Can.) which provides:

“Moral rights

14.1 (1) The author of a work has, subject to section 28.2, the right to the integrity of the work and, in connection with an act mentioned in section 3, the right, where reasonable in the circumstances, to be associated with the work as its author by name or under a pseudonym and the right to remain anonymous.”

By contrast, the United States copyright law provides a right of integrity to only a narrow class of visual artwork, usually originals. See 17 U.S.C. Section 106. The U.S. moral right is subject to “fair use” in a way that the Canadian moral right does not seem to be subject to “fair dealing.”

To parody would be to distort, and, since it might very well injure an author’s reputation, might violate the Canadian moral right of integrity. The statute provides:

“Nature of right of integrity

28.2 (1) The author’s right to the integrity of a work is infringed only if the work is, to the prejudice of the honour or reputation of the author,
So it may seem wrongheaded of the United States Supreme Court to suggest that it may be appropriate to give a parodist — a disappointed licensee — the liberty to copy for free on the ground that the owner would not sell him a license. Is the Court under-valuing the owner’s preferences? Not necessarily; there are several explanations of the Court’s approach that are consistent with the traditional economic deference to individual preferences. In fact, economics suggests that a rule like Canada’s that defers to an author’s copyright or her supposed “moral right of integrity” will yield results that are (administrative costs aside) inferior to a case-by-case approach which tolerates some parodies — however insulting the parodies might be to an artist whose work is being intentionally distorted and mocked.

Canada, like the United States, gives fair use or fair dealing to works of criticism or review. I shall argue that parodies fall into the same analytic class, and should be treated similarly.

When a copyright owner refuses to let someone adapt her work for purposes of parodying it, or refuses to give an ideological opponent permission to quote lengthy passages, or insists on suing anyone who quotes passages of her memoirs that reflect unfavorably on her, she is using her copyright as a

---

(a) distorted, mutilated or otherwise modified; or
(b) used in association with a product, service, cause or institution.

Where prejudice deemed
(2) In the case of a painting, sculpture or engraving, the prejudice referred to in subsection (1) shall be deemed to have occurred as a result of any distortion, mutilation or other modification of the work.

When work not distorted, etc.
(3) For the purposes of this section,
(a) a change in the location of a work, the physical means by which a work is exposed or the physical structure containing a work, or
(b) steps taken in good faith to restore or preserve the work shall not, by that act alone, constitute a distortion, mutilation or other modification of the work.”

R.S., 1985, c. 10 (4th Supp.), s. 6.
tool of suppression. The question of whether authors should be entitled to refuse permission to those users of whom they disapprove is a complex one. On which side of the issue would economics weigh in? If the proper way to look at these problems is economic, then, as mentioned, the principles of consumer sovereignty would seem to dictate that governmental decision-makers should not question why someone refuses to sell or license. Economics "assum[es] that man is a rational maximizer of his ends in life," and a desire to suppress would seem to be as rational an end as a desire for fame or fast cars.

Additionally, Ronald Coase has persuasively emphasized the importance of transaction costs by showing that, in their absence, the ultimate allocation of a resource will be efficient regardless of how entitlements are initially assigned.

107 Similar instances also appear in the corporate realm. For example, when a newspaper expanded its TV coverage it told its readership about the extended service in an advertisement that pictured a copyrighted TV Guide cover for purposes of comparison. TV Guide then sued for copyright infringement. Presumably the suit was motivated by something other than a desire for license fees. The comparative advertising was held to be a fair use. See Triangle Publications Inc. v. Knight-Ridder Newspapers Inc., 626 F. Supp. 1171 (5th Cir. 1980).

108 For example, it can be difficult to distinguish suppression from an attempt to direct the work into the most valuable derivative work markets. See, e.g., Paul Goldstein, Copyright Vol. I, at 571-573 (rights over derivative works can affect the direction of investment and the type of works produced).

Similarly, in regard to unpublished works, it can be difficult to distinguish cases of suppression from cases of economically motivated refusals to license. An author accused of suppression may be simply trying to keep the work out of the public eye temporarily until it reaches its mature form and can be published.

Even if some practical means existed to distinguish all dissembling "suppressors" from those copyright owners who are genuinely motivated by financial return, some cases will present instances of truly mixed motives. For example, the owner of copyright in an out-of-print collection of letters might sue a biographer who extensively quotes the letters, not only out of a dislike for the biographer's message or perceived inaccuracies, but also out of a desire to preserve the reprint market for the letters. See Meeropol v. Nizer, 417 F. Supp. 1201, 1208 (2d Cir. 1976), rev'd and remanded, 520 F. 2d 1061 (2d Cir. 1977), cert. denied, 434 U.S. 1013 (1978).


110 See R. H. Coase, "The Problem of Social Cost", loc. cit., note 83. The Coase Theorem is effective at least in the absence of factors such as
So long as the parties can meet face to face, as in copyright a copyright owner and potential parodist or critic could often do, why should there be any need for the judiciary to do anything but enforce whatever property right is before it?

Whether suppression would or would not be economically desirable will depend in most cases on empirical analysis of the particular fact pattern. But some general observations can indicate preliminarily why, when copyright owners seek to use the copyright law to avoid criticism or ridicule, neither consumer sovereignty nor the Coase Theorem suggest that judges should give the owners formal deference.

At least four reasons suggest that the market cannot always be relied upon to mediate attempts at suppression and that it might be economically desirable to refuse authors an entitlement to suppress. The four reasons are the “suppression triangle”: pecuniary effects; managerial discretion; and endowment effects. The four reasons are interrelated, and to explicate them let me begin with the “suppression triangle.”

102

LES CERTITUDES DU DROIT — CERTAINTY AND THE LAW


111 Even if one interprets copyright’s economic goal as being solely the use of incentives to “promote knowledge,” so that satisfying the copyright owner’s personal tastes would not count as an independent value, the empirical answer to suppression questions would not be easy: in a given case enforcing any particular type of suppression would both keep some knowledge secret, and yield long-term incentives that could aid knowledge in the long run (because authors who can suppress have a copyright worth more than authors who cannot). Cf. Frank I. MICHELMAN, “Property, Utility & Fairness”, (1967) 80 Harv. L. Rev. 1165 (the effects of demoralization on productivity). Which of the two potential effects on knowledge would be greater (the loss from enforcing suppression or the gain from long-term incentives) cannot be determined a priori.

112 For a fuller discussion of this issue, see Wendy J. GORDON, “The Right Not to Use” (unpublished manuscript on file with the author).

113 Additional reasons might include, e.g., the potential nonmonetizability of first amendment values. See W. J. GORDON, “Fair Use as Market Failure”, loc. cit., note 34, at 1631-1632.
In all these examples, remember: so long as there is the possibility that the social interest will be better served by refusing to enforce the owner’s copyright, an economic case is made for using a nonformal, reasonableness mode of inquiry. One would then need to compare the costs of the extra suppression that would result from adhering to a formal pro-owner result, with the administrative and other costs that would be necessary in employing a “reasonableness” or “fairness” standard. In the United States, with its strong history of prizing free speech, the costs of improper suppression of news or cultural material is viewed as very high.

A. Suppression Triangle

I use the term “suppression triangle”\textsuperscript{114} to point to the fact that in cases involving the suppression of information or other intellectual products,\textsuperscript{115} at least three parties are affected: (1) the person who seeks or threatens to make the contested use (for example, the potential parody), (2) the copyright owner who wants to keep the material from being copied or adapted (the potential suppressor), and (3) the person or persons who would want to see the material (the potential recipients). This is the triangle of affected interests. Yet in the suppression transaction typically only two parties are present: the potential user (such as a parodist), and the copyright owner. Whether an attempt to suppress is likely to be value-maximizing will depend, inter alia, on how well the interest of the omitted third party, the class of potential recipients, is represented by the two immediate participants.


I am indebted to Warren Schwartz for suggesting the potential relevance of the blackmail literature to this problem.

\textsuperscript{115} Information can implicate different issues from literary expression and other intellectual products; for purposes of this very general discussion, however, I shall group all together under the rubric “information.”
Theoretically, the more valuable the parody or other use is to the public, the more the public should be willing to pay for it, and the more the parodist should be willing and able to bid for permission. Thus, the notion of the Invisible Hand expects that any market participant will be in a position to reflect the interests of affected third parties (that is, the public audience). Nevertheless, the Invisible Hand often falters, and the possibility of misallocation remains.

Consider a hypothetical novelist or movie maker who wants to keep the world from knowing what a hostile critic or parodist has to say about his work. Assume also that the critic or parodist wants to quote from the work or use its imagery, and that use of the quotation or imagery is somehow essential to the comprehensibility or believability of the criticism or parody. If the law required the critic or parodist to purchase licenses to quote or paraphrase, how sure could we be that the “highest-valued” use would ensue?

For purposes of mathematical example, assume that the critic or parodist stands to earn at most a thousand dollars profit from even the best-written product. Assume that the novelist or film-maker would lose fifty thousand if the criticism or parody is published. Since the copyright owner would charge at least fifty thousand for a license to criticize or ridicule his work and the critic or parodist stands to gain only one thousand from publishing, it may look like the copyright owner holds the “highest valued” use when compared with the parodist or critic. But that may be an illusion resulting from the fact that the third party (in the owner/user/public triangle) is not being counted as part of the deal.

---

There is another factor that may be at work here as well: the idea/expression dichotomy. Since under current law copyright owners cannot prevent others from using their ideas, it could be argued that little suppression of note could occur; it might be suggested that a critic deprived of the privilege to quote could nevertheless communicate effectively.

For simplicity’s sake, therefore, assume that in the following examples, whatever the defendant has taken from the first artist’s work could be considered copyrightable expression rather than simply “idea” and that the use of the copyrighted expression is somehow essential to the effectiveness of the planned derivative work.
The publishing of the review or parody might benefit the public (who would thus be warned off from, let’s say, a much-hyped romance novel that doesn’t really excite anyone who reads past page five) to the tune of that same fifty thousand, or perhaps even more. On these hypothesized facts, requiring the publisher to buy a license from someone who would not sell it is a bad idea, and giving the publisher (the critic or parodist) free use is a good idea. And both are consistent with economic measures of value. If the critic had been able to capture the full value that the review gave to the audience, then the novelist’s fifty thousand minimum asking price would have been met.

A parodist may similarly be unable to capture the full value that the work holds for the audience. This can occur for many reasons. There may be significant positive externalities and surplus in the market for parodies, for example. There also may be other complications in the markets for reviews and parodies, such as pecuniary losses that diverge from societal economic losses.

B. Pecuniary losses

Much of the loss that can come from a critical review will often be merely pecuniary, reflecting not a net loss to society but rather a shifting of revenues from one novelist to another and possibly better one. It is as if the triangle now were a geometric figure with four points (the criticized novelist, the critic, the public, and the better novelist). If one could add to the price offered for the “license to criticize” an amount

117 As economist Michael L. Katz writes of the similar problem in the research and development area:

“In the absence of perfect discrimination, the firm conducting the R & D will be unable to appropriate all of the surplus generated by the licensing of its R & D, and the firm will sell its R & D results at prices that lead to inefficiently low levels of utilization by other firms.”


118 See Richard A. POSNER, “Conventionalist Defenses of Law as an Autonomous Discipline” (September 21, 1987) (unpublished manuscript, on file with the University of Chicago Law Review) (using pecuniary effects to explain why landowners who create certain positive spillovers are not entitled to payment from those who benefited).
reflecting the monies that the better novelist would reap, it might be enough to make the difference. Since this cannot happen, mere pecuniary losses may take on an importance they should not have and they might prevent socially desirable licensing.

C. Managerial discretion

Another possible complication has to do not with the potential buyer’s inability to raise the appropriate amount of capital, but with the potential licensor’s potential inability to know even a good deal when it comes along. This complication I will label managerial discretion, by which I mean to embrace all those things that may make managers in complex corporations sometimes arrive at decisions that are less value-maximizing than they could be. I would include here, for example, personal risk aversion, bureaucratic structure, group dynamics, and laziness. Thus, the officials of a company that owns a given copyright may refuse to license simply because the license is in an unfamiliar field and their particular bureaucratic structure penalizes unlucky risk takers more than it rewards lucky ones. When critical, parodic, or otherwise controversial licenses would be at issue, the human desire to “play it safe” might prevent value-maximizing transfers from occurring. Managerial discretion is just one of many agency

---

119 Journalistic ethics undoubtedly prohibit reviewers from accepting subsidies for doing hostile reviews.

120 There is a fairly extensive literature on the controversial question of whether managerial discretion exists and if so what impact it has and what should be done about it; all I mean to suggest here is the simple possibility that managers in complex corporations do not always make the same decisions that an individual owner of a business would.

121 In an individual, a taste for risk or laziness might be a legitimate part of her utility curve, but a manager is supposed to act unselfishly on the part of the corporation. There is a large literature on these agency problems.

122 It might be argued that tastes for laziness or risk aversion are simply preferences that deserve the same respect under the notion of consumer sovereignty as other desires. However, we are not talking here about the risk aversion or laziness of the copyright owner, but of some person who is fortuitously placed within the licensor organization to be able to control licensing decisions. Whether gratifying such a person’s taste in regard to laziness or risk serves greater economic ends (as, e.g., a form of compensation) is itself complex.
problems that can prevent the parties from dealing with each other like the unitary participants in the classic Coasian transaction.

D. **Endowment or wealth effects: pricelessness**

All of the above are reasons why socially desirable “licenses to be critical” are not likely to be granted if left solely to the devices of copyright owners.\(^{123}\) One additional and probably most important factor remains to be discussed: the difference between willingness to pay and willingness to sell, sometimes identified with “endowment” or “wealth” effects.\(^{124}\)

The concept here basically refers to the fact that giving someone an entitlement makes that person richer, and this may change how the holder values both the entitlement and other resources, and this in turn may affect how entitlements are eventually allocated once bargaining between that person and other persons is completed.\(^ {125}\) Wealth effects do not retard resources from moving to hands in which, given a particular entitlement starting-point, they have the highest-valued use. Nor are they often strong enough to make a difference; in instances where fungible commodities are sold in markets populated by many buyers and sellers, “buy” prices and “sell” prices probably

\(^{123}\) Of course, such licenses might be granted; I offer here only an abstract analysis which would need to be empirically verified.

\(^{124}\) Wealth effects are, roughly, the impact on one’s preferences brought about by a change in wealth, including the change brought about by being given, or being denied, an entitlement. See, e.g., E.J. MISHAN, “The Postwar Literature on Externalities: An Interpretive Essay”, (1971) 9 J. Econ. Literature 1 (the allocative impact of wealth effects illustrated at 18-21, though not explicitly in the context of the Coase theorem).

\(^{125}\) For an excellent numerical example, see *id.* at 18-21. It is well recognized that a divergence often exists between the price that a potential buyer would be willing to pay for a resource he does not own, and the price that the same person would demand before he would sell that same resource if the law had initially awarded its ownership to him. What is less clear is what terminology, explanations, and characterizations are best employed for discussing the phenomenon. For a valuable discussion suggesting, *inter alia* that traditional “wealth effects” do not fully explain divergence between willingness-to-accept and willingness-to-pay, see Elizabeth HOFFMAN & Matthew SPITZER, “Willingness to Pay vs. Willingness to Accept: Legal and Economic Implications”, (1993) 71 Wash. U.L.Q. 59.
tend to converge. But, when wealth effects do have an impact, they have the potential of rendering the meaning of “highest-valued” use indeterminate in the sense that the location of the highest-valued use is not independent of the law. Where wealth effects are strong, everything depends on the legal assignment of entitlements that form the transaction’s starting point. As a result, in such cases the search for the highest-valued use cannot provide a good basis for assigning initial entitlements.

Professor Coase showed that in a world without transaction costs, resources will be traded to their highest-valued uses, so that, as between any two users of a resource, if A can use the resource more productively than B, A will end up with it. Therefore, many scholars argue, in a real world full of transaction costs that can impede bargaining, it often makes sense to “mimic the market” and assign legal rights to the highest-valued user in the first instance. This is a core insight of Law and Economics.

Yet the Law and Economics argument largely depends on there being a stable highest-valued user. The injunction to “seek efficiency by mimicking the perfect market” only makes normative sense if the perfect market allocation is a constant. If the allocation of rights significantly affects the monetary valuation that parties place on a resource, then there may be no stable economic reality for the law to seek to mimic.

There is indeed a rare class of goods which lack this stability. These are the precious, personal, irreplaceable, crucial goods one thinks of as “priceless.” Examples are many: the Dead Sea Scrolls; family heirlooms; one’s children; one’s health; one’s reputation; one’s peace of mind. The monetary value a person places on one of these goods may well depend on

126 For a dramatic hypothetical example, see Alfred C. YEN, “Restoring the Natural Law: Copyright as Labor and Possession”, (1990) 51 Ohio St. L.J. 491, 518-519 (“flip flop” of rights).
127 See R. H. COASE, loc. cit., note 83.
128 See, e.g., G. CALABRESI & A. D. MELAMED, loc. cit., note 61, for a classic explanation of market-mimicry.
129 For further exploration, and for citation to relevant literature, see E. HOFFMAN & M. SPITZER, loc. cit., note 125.
whether the person has a legal entitlement to it (whether she “owns” it) or whether she must purchase it.

Consider health, for example. It is plausible that most people would be unwilling to sell their organs at any price, so that Jane Smith might turn down an offer of five million dollars from Billionaire X for one of her kidneys. Similarly, if Jane Smith has kidney failure and one of her dying relatives wills her a healthy kidney, she might well be unwilling to take the billionaire’s five million dollars in exchange for her entitlement to it. If so, Jane Smith looks like the kidney’s “highest-valued user.”

But should she have no entitlement to the kidney from the recently-deceased person (perhaps because the relevant jurisdiction does not recognize such bequests as enforceable), Jane Smith’s own budget and health insurance will place a limit on how much she can spend pursuing the transplant. It is highly unlikely she will be able to outbid Billionaire X for the kidney. If so, Billionaire X will appear to be the “highest-valued user.” One can draw from such a pattern no reliable information about whether the resource has its highest value in the hands of the billionaire or Jane Smith. This phenomenon might be called the “pricelessness effect,” and its presence may be one reason why our society resists giving private ownership in body parts and other items that are “priceless” in this sense.

The pricelessness effect is a subset of the category that economists call “endowment effects” which in turn is related to wealth effects: since assigning an entitlement to someone makes that person wealthier, it can affect the valuation the person puts on resources. Often a person’s “willingness to buy” price will differ significantly from the price at which she is “willing to sell”. Many people hedge the Coase Theorem by noting it does not apply when significant wealth or endowment effects are present. But usually the wealth or endowment effect is so minor that it does not impair the reliability of using a market mimicry approach to model efficiency.

---

130 Experimental evidence on this point has been collected by Matthew Spitzer and others.

131 The impact of endowment or wealth effects is sometimes exaggerated. See R. H. COASE, “Notes on the Problem of Social Cost”, loc. cit., note 110, at 170-
The “pricelessness effect” deserves having its own name precisely because the subcategory of effects it denotes are likely to be significant. The “pricelessness effect” comes into play when the entitlement at issue pertains to a good that (1) an individual or group values very highly and (2) which is virtually irreplaceable, and (3) when it is the allocation of that very good\textsuperscript{132} which is at issue. As to such items, the initial placement of the entitlement is likely to have a sharp effect on the price and allocation of the resource, even in the absence of transaction costs.

In cases of parody or criticism — both areas where “fair use” treatment tends to be awarded to defendants — reputation may be at issue. To many, reputation is priceless in the sense we have been discussing. For example, a novelist who fears that a journalist will use extensive quotations from her book to bolster a hostile review will be most unlikely to sell the journalist a license to copy those quotations — regardless of the price offered. But that does not mean the author’s preference is the “highest-valued use” in any meaningful sense, since that same author may be unable to buy silence if the law gives the journalist a “fair use” liberty right to publish. A similar analysis can be made of parody: since most people intensely dislike being ridiculed, the legal right may determine where the highest-valued use lies.\textsuperscript{133} In such cases, the market is nearly

\textsuperscript{174} (discussing arguments re the presumed effect of changes in legal position on the distribution of wealth and on the allocation of resources).

Professor Coase argues that the impact of wealth effects can be overstated because, among other things, if the legal rules are known in advance, the prices of applicable resources will likely alter in a way that minimizes such effects; in addition, he suggests, contractual provision for contingencies may be available to mitigate some changes in legal rules. See \textit{id.}, at 157. See also \textit{id.}, at 170-174. Neither of these devices are likely to eliminate the wealth effect — here “pricelessness” — in the context of authorial suppression of embarrassing criticism, however.

\textsuperscript{132} That is, while I predict that the law’s assignment of rights in organs is likely to have a distinct effect on a kidney’s allocation, it is a more complex question whether the law’s assignment of rights in organs will have much of an effect on the allocation of other resources.

\textsuperscript{133} These points are also explored in W. J. GORDON, “Private Censorship”, \textit{loc. cit.}, note 1, at 1042-1043; also see W. J. GORDON, “Fair Use as Market Failure”, \textit{loc. cit.}, note 34, at 1632-1636 (anti-dissemination motives).
useless as a guide\textsuperscript{134}, and formal deference to owners’ market powers is inappropriate.

For example, assume A is a novelist, a copyright owner who has an entitlement not to license and who is otherwise financially comfortable; she has perhaps $4000 in the bank and a two-year old car and a prospect of steady royalties. A may be tempted by B’s offer of, say, $10000 for a license to use her work, but she can afford to say no without altering her lifestyle. If B’s project is an ordinary commercial project and A will not be sacrificing more than $10000 from foregoing alternative uses of the work, she will probably license. (It might also happen that B’s project would not require an exclusive license and would not otherwise interfere with A’s other licensing opportunities. If so, granting B permission to go forward would have no opportunity cost at all for A. She would be even more likely to license such a use.) However, if B’s project is hostile toward A’s work, A may well refuse the license, either to protect her long-term economic interest (which may be a mere pecuniary loss, remember), her aesthetic reputation, or her feelings.

If however the law gave novelist A no entitlement to prevent B’s use, then she would have to persuade B not to publish (cf., blackmail payments.) The most she could offer B to persuade B not to make the critical use planned is the amount in her bank account, plus whatever she could sell her car for, plus whatever she could borrow on the strength of her expected royalty stream. The total may well be less than $10000, and A will probably demand a price in excess of $10000. Give A the entitlement and the highest-valued use of the contested expression is in her hands; give B the entitlement and the highest-valued use is in that licensee’s hands. The locus of the “highest-valued use” has shifted as a result of where the law places its entitlement.

\textsuperscript{134} However, if the market were to yield the same result under \textit{either} allocation of the legal right, then that result could be used as a guide at least to where lies the highest economic value (that is, value as measured by ability and willingness to pay).
In such cases, looking to the results of consensual transactions will not give us any information about who “should” have the right.

Another way to put the point is this: Economics is sometimes used as a normative guide for good social policy. When it is used in this fashion, its primary claim to legitimacy stems from the links between economics and utilitarianism. The more that income distribution restricts the expression of individuals’ preferences, the more shaky the link between economics and utility becomes. This linkage has the potential for completely breaking down in cases of “pricelessness.” Though in such cases the parties’ preferences may remain constant, both in their objects and in their intensity, a shift in who owns the entitlement may effectively disable one of those parties from effectuating that preference. Thus a legal regime that is committed (even in part) to utilitarian consequentialism would be unwise to rely upon a money-bound market model for normative guidance in cases of pricelessness.

In sum, refusing to allow a copyright owner to suppress a hostile use of the copyrighted work, in a case where the “pricelessness effect” is likely to make a determinative difference, does not necessarily contravene economic principles. In such an instance, it is appropriate for even an economically-oriented court to refuse to defer to the copyright owner, and instead make an individualized weighing of how enforcing the copyright in the given instance would affect welfare, and any other relevant consequentialist or nonconsequentialist policies.

Conclusion

This essay has examined the dynamics behind the key systemic choice made by copyright law, which is to reverse the presumption of Restitution law that people who refrain from

---

135 I am indebted here to Alan Feld.
136 This belief is rather controversial. See, e.g., such classic sources on the debate as the “Symposium on Efficiency as a Legal Concern”, (1980) 8 Hofstra L. Rev. 485 and Richard A. POSNER, The Economics of Justice (1987) for further discussion of the question of whether utilitarianism and economics are truly linked in this way.
crossing others’ tangible boundaries are free to take advantage of each other’s labor. In intellectual property law, this presumptive freedom is replaced by a duty not to copy. Such a duty impels potential copyists to identify themselves and seek contracts with authors and other creators. In this way, the authors are given monetary incentives to continue creating, and works are disseminated to the public. However, all this comes at a significant price: some people will not have access to works for which they would be willing to pay the marginal cost of production. This loss of access is particularly significant because the works may not be simply fungible goods, but may be irreplaceable, unique, and important artifacts crucial to comprehending and participating in a culture.

In order to keep access open to the most important cultural components, copyright law allows judges to declare that something constitutes an “idea” and is therefore not ownable. This flexible if muddy concept keeps the private property system permeable to the society’s changing preferences regarding commodification.

“Ideas” are not ownable at all. But some cultural artifacts must be owned, if authors are to have incentives in a world without extensive subsidy. Given the high costs of an erroneous judgment to privatize, U.S. copyright law provides, inter alia, two standards that provide liberty, particularly to members of the public who make creative use of others’ copyrighted work. These two standards are “substantial similarity” and “fair use.” Both allow some copying of expression, particularly by persons who add a great deal of their own talents to what has been copied, transforming the originals.

Canada’s treatment of certain transformative users — namely, parodists — is significantly different from that of the U.S. Many explanations present themselves. One is the United States’ romance with free speech. We value the iconoclast more than do most nations.137 But another obvious explanation is that

in each country copyright may serve a different purpose. In the U.S., the Constitution explicitly grants Congress power to pass copyright and patent laws to serve a public purpose, namely, furthering of the “progress of science and the useful arts.” Canada’s copyright law may follow more of a continental model, in which authors’ rights stand on their own as a valid reason for copyright. Given Canada’s dual heritage, from England and France, it seems likely that Canada would be more inclined than would the U.S. toward the French traditions of honoring the droit d’auteur.

But to the extent Canada is aiming at maximizing public benefit, and to the extent that economics is a reliable guide to that benefit, this essay has suggested that an automatic deference to authors over parodists cannot be justified.


139 I have suggested elsewhere that even an “authors’ rights” approach would yield far less protection to authors than do current copyright statutes, see W. J. GORDON, “Property Right in Self-Expression”, loc. cit., note 54. Admittedly, in that essay I was operating out of an Scottish/English “natural rights” tradition rather than a Continental one.
La légitimité des sources du droit pénal
(réflexions d’un agnostique sur les certitudes fondamentales du droit répressif)

André Jodouin*

Introduction................................................................. 117

I. Les vices dans la production des normes pénales......................................................... 117

A. L’incurie du Parlement................................. 118
   1. Les lacunes de la première codification........ 118
   2. L’échec de la réforme................................. 122
   3. La législation par réaction ....................... 129

B. Le zèle des tribunaux................................. 132
   1. L’évolution du rôle créateur de normes...... 132
   2. L’évolution des justifications du pouvoir de créer des normes ....................... 135
   3. Les tribunaux et la règle des trois unités ...... 136

II. Les tares congénitales des normes répressives : l’obscurité et les légitimations artificielles ......................................................... 139

A. Obscurité et légitimité........................................ 140
   1. Un cas de figure!: la mens rea ..................... 140
   2. L’obscurité et ses séquelles ....................... 144

* Professeur titulaire à la Faculté de droit de l’Université d’Ottawa, section de droit civil.
B. Les légitimations artificielles .......................... 145
   1. Légitimation et motifs surabondants .......... 145
   2. Légitimation et préambules ...................... 146

Conclusion..................................................................................... 149
Cette réflexion sur la légitimité des sources du droit pénal présente deux séries de constatations, d’abord que la production des normes pénales au Canada est entachée de vices systématiques qui mettent en cause sa légitimité. Ensuite que les normes produites par ce système sont tarées: elles sont obscures et les producteurs de normes ont recours à des moyens artificiels pour atteindre une légitimité qui ne s’impose pas d’elle-même.

I. Les vices dans la production des normes pénales

On raconte qu’à la cour de Louis XV, un marquis, rentrant inopinément dans ses appartements, y trouve sa marquise en flagrant délit avec un prélat de la cour. Sans sourciller, le marquis se rend à la fenêtre, l’ouvre toute grande et, se penchant dehors, se met à esquisser de grands gestes de bénédiction. L’évêque, interrompant ses coupables ébats, lui demande le sens de son manège. «Monseigneur», répond le marquis, «puisqu’il vous vous chargez de remplir mes devoirs, je prends sur moi de m’acquitter des vôtres».

On retrouve une situation analogue parmi les producteurs de la norme pénale au Canada, mais l’étonnement produit n’est pas source d’hilarité. Par un curieux renversement des rôles, le Parlement semble avoir abandonné aux tribunaux le choix des grandes orientations du droit pour se contenter de fonctions plus modestes et ponctuelles. De leur côté, les tribunaux (en clair, la Cour suprême) se sont affranchis des limites que leur imposait la tradition de la common law et s’adonnent allègrement et sans détours à des activités qui relevaient naguère des législatives. Ces espèces de saturnales juridiques posent de graves problèmes sur le plan de la légitimité des normes.

Dans notre système, la production des règles juridiques relève du Parlement et des tribunaux, notamment de la Cour suprême. Une bonne partie des difficultés actuelles résultent d’un mauvais partage des responsabilités entre ces deux producteurs. Le Parlement ne légifère pas comme il devrait le faire, en dessinant les grandes orientations de la politique criminelle et en proposant aux exécutants des principes généraux aptes à guider l’exercice de leurs pouvoirs...
d’application. Trop souvent, l’activité législative n’est que réactive: le Parlement réagit à l’opinion publique, aux interprétations que font les tribunaux de la Charte, aux pressions des groupes consacrés à la défense d’intérêts divers.

De leur côté, les tribunaux vont tenter de combler les lacunes de la dogmatique pénale. Alors que leur structure et leur méthode de travail conviennent particulièrement bien à l’interprétation et à l’application des normes, ils s’adonnent de plus en plus énergiquement à les créer.

Il faut dire que le manque d’initiative du Parlement est singulièrement conforté par les habitudes intellectuelles de la tradition juridique dominante au Canada, celles de la common law. Dans ce système, les principes généraux du droit pénal sont beaucoup plus le résultat d’une élaboration progressive par les tribunaux que l’effet d’une politique législative expresse. C’est ce qui ressort de l’histoire du droit pénal anglais; c’est de cette histoire que le droit pénal canadien est tributaire et le laconisme de la mère des parlements a donné naissance à l’incurie du nôtre. Au départ, le Parlement canadien est l’héritier de la tradition juridique dominante.

A. L’incurie du Parlement

1. Les lacunes de la première codification

Le premier Code criminel du Canada s’est fondé sur plusieurs modèles, notamment le Digest de Stephen\(^1\), celui de Burbidge\(^2\) (adapté du premier, avec la permission de l’auteur) et le projet de Code criminel de la Commission royale impériale\(^3\) présenté, mais sans succès, au Parlement britannique. Dans toutes ces sources, l’influence de Stephen prédomine; il n’est donc pas inutile d’examiner la conception de la

---

codification qui ressort de son oeuvre. Or, cet examen révèle que Stephen avait une vue tronquée de ce qui constitue un code\(^4\).

Ainsi, l’ouvrage charnière qui sert de modèle au projet anglais de codification (et partant, avec les aménagements qui se sont imposés, au nôtre) contient intentionnellement des lacunes:

\[\text{Chapter II., [sic] entitled «General Exceptions», intentionally omits an examination of certain legal doctrines, too vague to be thrown into the form of propositions [...]. I have said nothing as to the doctrine of intention, as to the sense of the word «malice», or as to the maxim «Non est reus nisi mens sit rea.» These topics will find a more appropriate place in a commentary on the law than in a statement of the law itself[...].}\(^5\)

Selon Stephen, la codification du droit criminel «\textit{when fully understood, means only its reduction to an explicit systematic shape, and the removal of the technicalities and other defects by which it is disfigured}\(^6\). Le code qu’il propose correspond donc au type de codification «palliative», destinée, «au premier chef, à atténuer les maux engendrés par le désordre...»


\(^5\) J.F. STEPHEN, \textit{op. cit.}, note 1, p.xxix. Le chapitre III intitulé «General Exceptions», contient des dispositions visant la minorité pénale, l’aliénation mentale, la présomption que toute personne est saine d’esprit, l’ivresse, la présomption de contrainte en faveur de la femme mariée, la défense de contrainte, la défense fondée sur l’état de nécessité, l’ignorance de la loi et l’ignorance de fait.

et la surabondance des normes»

Il ne révèle aucune volonté «organisatrice», dans la mesure où les principes généraux de la responsabilité pénale n’y sont pas exposés de façon systématique. Une partie intitulée «Justification and excuse for acts which would be otherwise offences» aborde certaines défenses comme celles fondées sur la minorité pénale ou l’aliénation mentale, mais les conditions générales de la responsabilité demeurent implicites. Par contre, la possibilité pour les tribunaux d’appliquer les défenses de la common law à des situations nouvelles est expressément réservée.

Il est important de s’interroger sur les motifs de ce laconisme des modèles de notre première codification. Il s’agit en effet d’un choix délibéré des légistes de l’époque. Pour Stephen, il n’est pas souhaitable d’articuler les principes de la responsabilité en ce qui concerne «the principles which declare what circumstances amount to a justification or excuse for doing what would be otherwise a crime, or at least would alter the quality of the crime».

L’opposition de Stephen repose sans doute sur une conception de la règle de droit propre à la tradition juridique de common law. D’ailleurs, la conviction de l’auteur quant à la supériorité de son droit national s’est exprimée en d’autres lieux. Ainsi, on retrouve dans son A History of the Criminal Law of England les deux thèmes de la grandeur du droit anglais et de la misère des autres droits!: le droit anglais

represents [...] the results of the labours of the most powerful legislature and the most authoritative body of judges known to history. In no other country in the world has a single legislature exercised without dispute and without rival the power of legislating over a

---

7 Cf. Alain François BISSON, Aspects contemporains de la codification, conférence publiée dans Ordre juridique et espace marchand, Actes des colloques Ottawa-Nantes, 1994-95, à paraître prochainement chez Wilson et Lafleur.

8 BRITISH PARLIAMENTARY PAPERS, Royal Commission Select committee and other reports on the Criminal Law with Proceedings Minutes of Evidence appendix and index 1847-79, Shannon, Irish University Press Series.

9 Id., Draft Code, art. 19.
compact and yet extensive nation for anything approaching to so long a period as the parliament of England. In no other country has a small number of judges exercised over a country anything like so extensive and compact the undisputed power of interpreting written and declaring unwritten law, in a manner generally recognized as of conclusive authority [...] Any code which was not founded upon and did not recognize these characteristics of the law of England would give up one of its most valuable characteristics.10

À l’autorité morale et intellectuelle de ce système, on doit comparer le caractère facile et superficiel du droit des autres nations:

The generality of language which is characteristic of the foreign codes would be wholly unsuited to our own country, and it would necessitate the re-opening and fresh decision of a great number of points which existing decisions have settled. There is no doubt something attractive at first sight in broad and apparently plain enactments. Further acquaintance with the matter shows that such enactments are in reality nothing but simple and therefore descriptive descriptions of intricate subjects.11

Le Code criminel du Canada, adopté en 1892 et entré en vigueur en 1893, reflète en très grande partie la conception de Stephen!: même absence d’énoncés des principes généraux du droit, même recours à la voie judiciaire pour parfaire les excuses, justifications et défenses. Depuis son adoption, le Code a subi un certain nombre de changements et de mises à jour, notamment en 1955. Toutes les réformes ont été ponctuelles, visant soit une ou plusieurs infractions, soit une ou plusieurs règles de la partie générale!: par exemple, la règle relative au seuil d’âge de la responsabilité pénale, ou la définition de l’ancienne aliénation mentale, devenue les «troubles mentaux».

---

10 J.F. STEPHEN, op. cit., note 6, p.355.
11 Id., pp.355 et 356.
2. L’échec de la réforme

L’adoption du *Code criminel* de 1892 représente un progrès quant à la situation antérieure et on ne peut vraiment blâmer le Parlement canadien d’avoir adopté une loi qui n’avait d’un code que le titre. Par contre, l’échec récent de la recodification ne mérite aucune indulgence. Les années 70 et 80 ont été marquées, en effet, par un mouvement de réforme du droit pénal. La Commission de réforme du droit, créée en 1971, s’était penchée sur l’état du droit pénal au Canada. Dans son troisième rapport au Parlement, la Commission avait proposé un vaste réaménagement de ce droit.

Sur le plan du fond, on proposait que le droit pénal n’intervienne que lorsqu’un comportement répréhensible enfreignait les valeurs fondamentales de la société. Ce principe de modération entraînait une restriction de la portée du pénal et une décriminalisation des comportements où la sanction pénale n’était pas à la fois nécessaire, efficace et appropriée. Il postulait la nécessité d’une revalorisation de la notion de faute personnelle, essentielle pour la légitimité morale de la sanction. Il limitait l’emploi du procès traditionnel et proposait la déjudiciarisation des cas moins graves. Il restreignait le recours à l’incarcération et recommandait «l’utilisation de sanctions plus créatrices et plus constructives».

En ce qui concerne la législation pénale elle-même, la Commission, dès le départ, a affirmé la nécessité de sa rationalisation. Le Code est mal structuré, il ne contient pas de principes directeurs, il ne fait pas de distinction entre les infractions réglementaires et les crimes véritables. Il contient un excès de détails «qui brouille le message clair, simple et direct du Code». De là, la nécessité de légiférer.

---

13 *Id.*, p.31.
14 *Id.*, p.37.
15 *Id.*, p.32.
En 1979, le gouvernement de l’époque a admis que le régime des modifications ponctuelles du Code avait donné lieu à une politique criminelle incohérente et inefficace; le ministre de la Justice a fait état de la nécessité d’une révision en profondeur du Code. La même année, les ministres provinciaux responsables de l’administration de la justice et le gouvernement fédéral ont convenus que cette révision du Code constituait une priorité législative. On a élaboré un plan d’action!: recherche fondamentale, analyse et formulation de recommandations par la Commission de réforme, révision subséquente des recommandations par les ministères de la Justice et du Solliciteur général, et rédaction sous forme de projet de loi; ces deux étapes prévoient de vastes consultations avec les provinces, les barreaux, les associations de policiers et le public16. Enfin, le Parlement serait saisi du projet17.

On connaît la suite!: rapports de la Commission de réforme au Parlement, consultations, élaboration d’un livre blanc, proposition de modification de la Partie générale du Code. Le Canada, 100 ans après l’adoption et l’entrée en vigueur de son premier Code, allait enfin se doter d’une loi moderne, cohérente, claire!: un véritable code.

Comme chacun le sait, le projet a fait long feu. La réforme en profondeur du code pénal a cessé d’être une priorité pour le gouvernement. Que s’est-il passé?

Plusieurs facteurs semblent avoir convergé pour contrer la réforme. Facteurs politiques, sans doute, au sens électoral du terme. Comme l’a écrit un auteur particulièrement bien placé pour connaître les aléas de la réforme,

---

16 Il est bien évident que le processus adopté comportait un dédoublement des efforts, particulièrement quant à la consultation des intervenants non gouvernementaux. L’hypothèse du conflit de territoire entre la Commission et le ministère de la Justice est tout à fait plausible. Celle d’un sabotage du projet n’est pas non plus exclue.

il n’existe pas, il faut bien le reconnaître, de district électoral pour la réforme du droit criminel. Personne n’a encore été témoin de défilés sur la colline parlementaire pour réclamer, pancartes et slogans à l’appui, un nouveau Code criminel canadien.18

Dans la mesure où le projet de recodification représente une amélioration de la juristique pénale, on voit mal, en effet, comment un gouvernement pourrait convaincre la population canadienne de la nécessité d’une telle réforme. L’elegantia juris ne semble pas figurer parmi les préoccupations les plus urgentes de la société canadienne19.

La recodification s’est également heurtée à des obstacles institutionnels et méthodologiques. La Commission de réforme du droit était un organisme indépendant, relevant directement du Parlement du Canada. Or, la responsabilité de préparer la législation pénale revenait traditionnellement au ministère de la Justice. Le Ministère a insisté sur son obligation de reprendre le travail de la commission de réforme. Ainsi, dans le cadre de l’engagement du gouvernement de repenser le droit pénal, on a formé au ministère de la Justice un groupe de travail chargé d’étudier la situation dans le système de justice pénale de l’aliénation mentale (renommée «désordre mental» et éventuellement «troubles mentaux»). Le groupe de travail a produit un document de consultation et a entrepris des consultations auprès des ministères provinciaux, des autorités de la santé, des associations de police et de tous ceux que la question intéressait. Il en est résulté un vaste projet de loi sur le traitement par le système de justice pénale des personnes


19 À cet égard, on a raté l’occasion rêvée de lier le projet à un événement symbolique important! le centenaire de l’adoption (1992) ou de la mise en vigueur (1993) du premier Code criminel. Pour des raisons qui échappent à la logique, le chiffre 100 a un pouvoir d’entraînement qui équivaut quasiment à une dispense de preuve.
atteintes de troubles mentaux. La loi\textsuperscript{20} a été adoptée en 1991, comme modification du \textit{Code criminel}. La plupart de ses dispositions portent sur la procédure, encore que l’une d’entre elles modifie le verdict qui doit être rendu lorsqu’on détermine que l’accusé, au moment de la commission de l’infraction, était atteint de troubles mentaux. Ce n’est plus un verdict d’acquittement pour cause d’aliénation mentale, mais de non-responsabilité criminelle pour cause de troubles mentaux\textsuperscript{21}. Il s’agit donc d’une loi qui a des répercussions théoriques importantes pour la Partie générale du droit pénal.

On avait également entrepris, à l’époque même où la «réforme en profondeur» était engagée, une étude complète de la sanction pénale, sous l’angle de la détermination de la peine. La Commission Archambault, chargée de cette étude, a déposé en 1987 un rapport qui a inspiré d’importantes modifications, entrées en vigueur en 1994\textsuperscript{22}. Pourtant, on voit mal comment on peut scinder la réforme des principes généraux de la responsabilité pénale de celle des peines, l’une étant logiquement tributaire de l’autre\textsuperscript{23}.

Parmi les causes de l’échec on peut compter, de plus, une certaine hostilité de la part de beaucoup de juristes canadiens à l’endroit de l’adoption d’une forme de droit perçue comme étrangère. En effet, on a pu constater que le principe même de la codification ne fait pas l’unanimité chez les

\begin{itemize}
\item \textsuperscript{20} \textit{Loi modifiant le Code criminel (troubles mentaux) et modifiant en conséquence la Loi sur la défense nationale et la Loi sur les jeunes contrevenants, L.C. 1991, c. 34; Cf. pour l’historique de cette loi, Edwin A. TOLLEFSON et Bernard STARKMAN, \textit{Mental Disorder in Criminal Proceedings}, Scarborough, Carswell, 1993.}
\item \textsuperscript{21} \textit{Code criminel, L.R.C. (1985), c. C-46, modifié par L.R.C. (1985), c. 2 (1er supp.), art. 672.34.}
\item \textsuperscript{22} La loi modificatrice n’a malheureusement pas repris certaines des recommandations les plus importantes de la Commission, notamment la réduction des peines maximales démesurées et l’adoption de «lignes directrices» destinées à éliminer les disparités dans les sentences imposées par les tribunaux. Néanmoins, la loi a articulé bon nombre de principes élaborés par la jurisprudence anglaise et canadienne.
\item \textsuperscript{23} Paul FAUCONNET, \textit{La responsabilité: étude de sociologie}, Paris, Félix Alcan, 1920! «Une théorie de la responsabilité supposerait une théorie des sanctions.» (p. 16).
\end{itemize}
professionnels du droit. Ces résistances m’apparaissent, en partie, d’ordre culturel: la culture juridique dominante au Canada est celle de la common law. Or, traditionnellement, dans cette culture, le dispensateur du droit n’est pas le législateur, mais le juge. Le destinataire de la norme n’est pas le justiciable, mais l’avocat. On le voit bien dans la genèse du Code criminel canadien de 1892: le point de départ du Code est un condensé des décisions judiciaires. On le voit, dans le rejet du projet de Code de la preuve de la Commission de réforme du droit. On le voit dans le style de rédaction de certaines lois, inintelligibles sauf pour certains spécialistes. On le voit à de multiples indices, dans les réactions de la communauté juridique au projet de recodification.

24 Voir G. LÉTOURNEAU, loc. cit., note 18, 84!: «[L’]exercice de codification présuppose un engagement politique et doit pouvoir s’appuyer sur un consensus social. À cet égard, la résistance des juristes canadiens au principe même de la codification est déjà bien connue.»


26 Une des réactions les plus bizarres est celle de Brian Hogan dans A. M. LINDEN, loc. cit., note 17, 67, où il écrit!: «By and large, as it seems to me, law reform agencies are confined to the reform of «lawyer’s law». That is not a precise phrase, but it refers to the systematisation or codification of the criminal law, the law of contract, the law of tort and so on. Perhaps this is a laudable object but I do sometimes wonder if this is where the shoe is really pinching. Lawyers love debating what is, or what ought to be, the mens rea of murder but very few citizens ever find themselves charged with murder and fewer still find that their case exercises the minds of the judges on their supreme court. The shoe pinches, as it always does, down at street level where vastly more people find themselves stopped and questioned, or searched, or arrested. It has always seemed to me that law reform agencies should have designated police powers a first priority on their programme and relegated the reform of the substantive law of crime to a poor third.»

Ces remarques paraissent assez saugrenues dans le cadre d’un ouvrage commémorant la mémoire de Jacques Fortin, un des principaux artisans du projet de code de la Commission de la réforme du droit du Canada. On remarquera, d’ailleurs, la conception des «codes» que semble évoquer l’auteur: la systématisation des torts, des contrats, du droit criminel. C’est, à
L’appartenance à une culture juridique a des conséquences politiques dans un pays comme le Canada où deux systèmes juridiques se côtoient. Si on peut qualifier le système juridique québécois de «mixte», ou de «hybride», on ne peut porter le même jugement sur le droit du reste du Canada, ce qui est à peu près normal, ni sur le droit fédéral, ce qui l’est moins. Le droit de source fédérale reste résolument rattaché à ses racines de common law; ainsi en est-il du droit pénal. Il en résulte que les juristes de formation civiliste n’auront de dialogue avec leurs homologues de common law qu’en acceptant un discours commun, celui de la common law\textsuperscript{27}. Le projet de recodification de la Commission et, dans une moindre mesure, celui du Ministère auraient introduit un autre langage dans le discours juridique canadien, celui du système civiliste, avec ses énoncés de principe, son mode de rédaction concise, sa volonté de cohérence\textsuperscript{28}. Le pari de la recodification n’était pas gagné au départ.

La recodification, comme le disait l’auteur, était un obstacle considérable à l’établissement d’un véritable code: «But while a code can supersede the common law, it cannot be supposed that it would ever oust the common law technique. We are unlikely ever to adopt the civil law approach to codes where the code is regarded as the authoritative, comprehensive and exclusive source of law. I would not claim for a moment that the common law technique is any better than that of civil law systems, but it is ours and it works.»

L’auteur est anglais, mais il me semble traduire des sentiments fréquemment exprimés au Canada. À noter, l’ineffable possessif ours, en italique dans le texte.

\textsuperscript{27} Jacques Fortin se plaisait d’ailleurs à répéter qu’un collègue ontarien l’avait qualifié de «best common lawyer in Canada»!

\textsuperscript{28} Cf. pour un exemple de résistance systémique fondée sur l’appartenance au système juridique dominant, Christopher NOWLIN, «Against a General Part of the Canadian Criminal Code», (1993) 27 U. B. C. L. R. 291, 296, 301 et 312. A. F. BISSON, op. cit., note 7, impute charitahtablement les résistances de cet auteur «à l’air du temps»; il s’agit plus probablement d’une manifestation de l’ignorance endémique des common lawyers à l’endroit du système civiliste, sinon d’un réflexe de défense déclenché par la crainte d’être dépossédé du pouvoir que confèrent des connaissances ésotériques. Nowlin affirme que l’adoption d’une nouvelle partie générale aurait pour conséquence d’enchaîner dans le Code criminel des contre-valeurs telles la protection de la propriété et des droits sexuels des hommes («male property rights, male sexual interests»). L’auteur n’explique pas pourquoi ni comment ces contre-
Enfin, l’absence de volonté politique favorable à un changement en profondeur du droit pénal canadien peut aussi s’imputer au rejet de la philosophie de la modération qui a inspiré les travaux de la Commission de la réforme du droit. Le discours actuel sur la politique criminelle n’invoque plus ce principe. Au contraire, le discours politique reprend volontiers celui des médias qui, prétendument au nom de l’opinion publique, réclament un renforcement de la répression. Les associations pour la défense d’intérêts divers, depuis les droits des victimes jusqu’à ceux des animaux, en passant par les associations de policiers, tous réclament l’incrimination de comportements actuellement impunis ou l’aggravation des sanctions prévues pour des incriminations existantes. Le phénomène est d’autant plus mystifiant que les statistiques criminelles révèlent plutôt une baisse qu’une augmentation de la criminalité.

En quoi la volonté répressive s’oppose-t-elle à l’élaboration de principes généraux légiférés? L’idéologie répressive a tendance à nier les distinctions morales sur lesquelles se fondent ces principes. Ainsi, Fauconnet définit les règles de la responsabilité comme

valeurs qu’il a découvertes dans le système actuel seraient renforcées par l’adoption d’un véritable code. Croyait-il que la codification entraîne la rigidité des normes? Estime-t-il que les sociétés régies par des normes codifiées sont plus acharnées que les autres à défendre le patriarcat? Son traitement de l’argument selon lequel la clarté de la loi pénale favorise son appréhension par le public est d’ailleurs révélateur: «[...] public accessibility [sic] to the criminal law is a genuine concern, then institutional barriers to understanding criminal legal theory and practice (such as academic preconditions for entering law school) would have been dismantled long ago.» C’est là que l’auteur révèle ses convictions inconscientes: le droit, le seul vrai, est un lawyers’ law. Pour le comprendre, il faut avoir fait son cours de droit!

Parmi les comportements nouvellement érigés en infraction, on retrouve le «tourisme sexuel». À l’heure actuelle, le ministère de la Justice envisage d’augmenter les peines prévues pour l’infraction de cruauté envers les animaux. Cette infraction rend présentement son auteur passible d’un emprisonnement de 6 mois et d’une amende de 1000 $. La peine maximale passerait à 5 ans de pénitencier et 5000 $ d’amende. On peut, sans être favorable à la promotion de la pédophilie ou insensible aux souffrances des animaux, douter que les moyens adoptés constituent la solution la plus efficace à ces problèmes.
celles qui prescrivent comment doivent être choisis, à
l’exclusion de tous autres, les sujets passifs d’une
sanction et comment cette sanction doit être modifiée,

dans sa grandeur ou dans sa nature, pour s’appliquer à
ces patients, la modification pouvant aller jusqu’à
l’annulation de la sanction.30

Dans un système qui privilégie la répression, la seule
mesure de la sévérité de la sanction est le préjudice que cause
l’infraction, évalué généralement dans la perspective de la
victime. Or, dans cette perspective, il n’y pas d’infraction
triviale. Toutes les infractions sont graves; elles peuvent être
plus graves, elles ne peuvent l’être moins. On constate cette
tendance, par exemple, dans les revendications des groupes de
victimes qui font de la sévérité de la peine l’étalon de la valeur
que la société attache à leur personne!: une peine qui tient
compte de circonstances atténuantes dans le comportement de
l’accusé est interprétée et dénoncée comme une négation des
souffrances de la victime ou de son droit d’être protégée31. La
tendance répressive va donc nier la pertinence des nuances
qu’une théorie générale bien faite permet d’introduire dans le
droit pénal32.

30 P. FAUCONNET, op. cit., note 23, p.111.
31 Voir, par exemple, la réaction de certains groupes de défense des intérêts des
161, 185 W.A.C. 161. La récupération de la victime par l’idéologie répressive
est un phénomène assez récent. Il ne faut pas oublier que l’établissement du
droit pénal établit a d’abord passé par la négation du droit de la victime de
se faire justice à elle-même. L’article 141 du Code criminel incriminant
le fait de composer avec un acte criminel, est un rappel d’une vérité
historique!: l’action pénale appartient à la Couronne. À l’heure actuelle
encore, même si le système répressif exploite abondamment l’image de la
victime, elle ne lui accorde aucun pouvoir de décision quant à la poursuite
du crime. Cf. pour une réflexion intéressante sur l’identification de la répression
du crime, Antonine GARAPON, Le gardien des promesses, Paris, Éditions
32 Le professeur Milsom soutient que le niveau de développement d’un droit se
mesure à sa capacité d’appréhender le plus grand nombre possible de faits,
S.F.C. MILSÔM, «Law and Fact in Legal Development», (1967) 17
U.T.L.J.1.
3. La législation par réaction

Même si le gouvernement a ajourné *sine die* ses projets de codification, l’activité législative se poursuit allègrement. Cependant, au lieu de dominer les questions, au lieu de chercher à structurer la matière pénale, au lieu de mettre en œuvre une politique criminelle cohérente, le Parlement travaille à la pièce. En matière pénale, ce n’est plus le Parlement qui dirige; le Parlement suit. Si on cherche à qualifier l’activité législative depuis quelques années, l’idée qui vient inévitablement à l’esprit est que le Parlement réagit. Par un curieux retour des choses, le Parlement est comme un tribunal qui n’agit qu’après avoir été dûment saisi d’un cas particulier.

À quoi réagit-il? D’abord, aux faits divers. En Saskatchewan, un homme est acquitté d’un homicide parce que sa victime n’est pas décédée dans l’an et le jour qui a suivi l’agression. La ministre de la Justice n’a eu de cesse qu’elle n’eût présenté au Parlement du Canada une modification pour corriger cet anachronisme. Pourquoi un anachronisme? Parce que la règle était prétendument justifiée par l’incapacité de la science d’autrefois d’établir avec suffisamment de certitude qu’un décès survenu après ce délai avait bien été causé par l’acte délictueux. Notons par ailleurs la règle générale selon laquelle les actes criminels sont imprescriptibles. Pourquoi? Parce que les poursuites criminelles étaient des actions royales et, comme chacun le sait, *<tempus non occurrit regi>*. Pourtant, personne ne songe à remettre en question le principe de l’impresscriptibilité des actes criminels. Y-a-t-il des bons anachronismes et des mauvais?

Un autre exemple est la réaction du Parlement aux rivalités meurtrières des gangs à Montréal, lorsqu’un enfant est mort dans une explosion destinée à des membres d’un gang rival. Le Parlement a modifié le *Code criminel* par une loi qui range parmi les meurtres au premier degré ceux qui sont commis au moyen d’explosifs. Le préambule de la loi rappelle

---

les faits divers qui l’ont inspirée: «Attendu [...] que l’usage de la violence par les gangs a tué ou blessé plusieurs personnes, dont certaines sont des victimes innocentes [...]»34 [!]


Il réagit aussi beaucoup à la jurisprudence constitutionnelle. Si la Cour suprême déclare inconstitutionnelle une disposition du Code limitant le contre-interrogatoire des victimes d’agression sexuelle36, le Parlement réplique en modifiant les conditions de la défense d’erreur quant au consentement37. Si la Cour suprême juge que la défense d’ivresse extrême peut être invoquée à l’encontre d’une accusation d’agression sexuelle38, le gouvernement propose au Parlement le Projet de loi C-3339; le ministre qui le présente annonce fièrement!: «C’est la réponse du gouvernement et du

35 Loi modifiant le Code criminel (prostitution chez les enfants, tourisme sexuel impliquant des enfants, harcèlement criminel et mutilation d’organes génitaux féminins), L.C. 1997, c. 16.
39 Devenu l’article 33.1 du Code criminel, précité, note 21. La loi modificatrice, Loi modifiant le Code criminel (intoxication volontaire), L.C. 1995, c. 32, contient un préambule qui, il va de soi, ne fait pas partie du Code. Le procédé est assez curieux et, me semble-t-il, assez contestable dans la mesure où un code a pour but de réunir et de rendre accessibles les diverses sources de droit. Chaque fois que l’on utilise ce procédé, le Code devient un peu moins complet et sa consultation de plus en plus insatisfaisante. Voilà une preuve de plus que le droit pénal canadien appartient à une tradition dans laquelle la norme s’adresse au lawyer plutôt qu’au justiciable.
Notons que la majeure partie des réactions du Parlement vont dans le sens d’un renforcement de la répression. À cet égard, on semble considérer la Charte canadienne comme un empêchement de danser en rond: les law officers de sa Majesté qui défendent les projets ministériels devant les comités de la Chambre assurent les députés que leur projet est «Charter risk-proof», c’est-à-dire qu’il n’est pas susceptible de faire l’objet d’une déclaration d’inconstitutionnalité de la part des tribunaux. Par ailleurs, on a l’impression, en lisant la législation récente, que les lois répressives tentent d’occuper tout l’espace constitutionnel disponible41.

B. Le zèle des tribunaux

1. L’évolution du rôle créateur de normes

Peut-on croire que, de mémoire d’homme, il y a eu une époque où les juristes reprochaient aux tribunaux, et notamment à la Cour suprême, une excessive timidité dans la création judiciaires de normes? Pourtant, au début des années 60, bon nombre de professeurs et de praticiens pénalistes estimaient que la Cour suprême faisait montre d’une trop grande prudence à l’égard des sources de droit qui circonscrivaient son action. D’abord, la Cour ne prenait pas suffisamment ses distances vis-à-vis de la Chambre des lords, dont elle continuait à reconnaître le rôle prééminent d’oracle de la common law. Ensuite, elle ne tirait pas assez parti de la disposition du Code criminel qui maintenait en vigueur les défenses de common law. Enfin, au nom de la déférence constitutionnelle, elle hésitait à assumer tous les pouvoirs de contrôle des lois qu’une interprétation plus audacieuse de la Déclaration canadienne des droits aurait pu autoriser.

40 Débats des Communes, 16 juin 1995. Au ministère de la Justice, on avait surnommé le projet de loi le «in your face! amendment».
41 Voir à cet égard la Loi réglementant certaines drogues et autres substances, précitée, note 24, qui évite, mais de justesse, les écueils sur lesquels l’ancienne Loi sur les stupéfiants, S.R.C. 1970, c. N-1 s’était échouée.
La situation s’est complètement renversée, au point où, maintenant, la Cour légifère sans détours. Bon nombre de facteurs ont contribué à cette révolution. D’abord, la Loi constitutionnelle de 1982 qui affirme sans ambages l’inopérance des lois incompatibles avec les valeurs qu’elle protège. Il est clair que les nouveaux pouvoirs constitutionnels de la Cour lui ont montré une nouvelle voie et pas seulement en ce qui concerne l’interprétation de la Charte. En premier lieu, cette dernière déborde largement dans les autres disciplines du droit, notamment en droit pénal. Ensuite, l’exercice du pouvoir créateur dans un domaine où ce dernier est largement perçu comme légitime crée des habitudes de jugement qui persistent dans des domaines autres que le domaine constitutionnel42.

Il ne faut pas oublier, non plus, que d’autres facteurs institutionnels ont joué pour transformer tant la perception qu’on a du rôle des tribunaux que les moyens à la disposition de ces derniers pour jouer un rôle plus actif. Depuis le début des années 70, les juges de la Cour suprême bénéficient de l’aide de secrétaires juridiques, de stagiaires.

Les stagiaires des juges n’ont sans doute pas, au Canada, l’influence idéologique que peuvent avoir leur homologues américains43. Néanmoins, à d’autres niveaux, de façon moins évidente et peut-être plus pernicieuse, cette influence peut être considérable44. Citons à cet égard un ancien stagiaire:

Because clerks are often responsible for the precise wording of judgments or portions of judgments, they have some influence over the «form» these judgments


43 Encore que certains d’entre eux fassent état, dans leur curriculum vitae, de jugements à la rédaction desquels ils ont participé de façon plus que mineure...

take. The first area in which this impact is felt is in the length of decisions. Unless given specific guidance [...] most clerks tend to write lengthy judgments [...] Second, clerks tend to have had their legal education [...] more influenced by law reviews [...] This academic style seeps into their approach to writing at the Court.45

Les stagiaires serviraient donc de courroie de transmission entre le monde universitaire et le pouvoir judiciaire.

Cette idée mérite qu'on s'y arrête. De quel monde universitaire s'agit-il? Bon nombre de professeurs canadiens ont fait leurs études aux États-Unis. Les principales orientations idéologiques que l'on constate dans le milieu universitaire canadien, qu'il s'agisse d'analyse économique, de théorie féministe, ou de critical legal studies ont été, en grande partie, articulées aux États-Unis. Même si les arrêts de la Cour suprême citent des Canadiens deux fois plus souvent que des Américains46, la différence peut être plus apparente que réelle si l'auteur canadien formule les problèmes juridiques de la même façon que ses homologues américains47.

En ce qui concerne le fond, certains des stagiaires sont «result-oriented» et croient que la Charte «can accomplish policy ends». Les stagiaires n’ont pas, selon Stossin, une «significant and direct influence on the decisions of the

45 On pense à la critique que Milsom adressait à la jurisprudence anglaise!: «Longer judgments cite more cases to settle smaller questions less clearly.» S.F.C. MILSOM, loc. cit., note 33, 217.


À légibilité des sources du droit penal

Auteur compare la Cour suprême “élargie” (c’est-à-dire juges et stagiaires confondus) au Cabinet du Premier ministre: les conseillers et techniciens qui entourent le Premier ministre contribuent à l’élaboration des grandes orientations du gouvernement, mais seul le Premier ministre en prend la responsabilité politique.

La redéfinition du rôle des tribunaux, jointe aux changements institutionnels qui ont modifié leurs habitudes de travail et leur mode de rédaction ont créé des problèmes de légitimité pour les normes judiciaires. Les initiés commencent à formuler des doutes. Ainsi en est-il de l’avocat qui prétend que la jurisprudence est devenue inintelligible, du juge qui déclare ne plus la lire, du professeur qui capitule devant la quasi impossibilité de l’enseigner. La grogne n’est pas encore la révolte, mais on constate une insatisfaction croissante à l’égard de la façon dont le droit se construit. Inévitablement, les interrogations se portent sur le rôle de la Cour suprême parce que, à l’heure actuelle, c’est elle qui est sans doute le plus prolifique des producteurs de normes et celui dont l’activité est la plus visible.

L’évolution des justifications du pouvoir de créer des normes

À quelle aune mesurer la légitimité des normes prétorienennes? Aucun texte constitutionnel ou légal ne fonde directement le pouvoir des juges de créer des règles de droit. On peut toutefois le déduire de certains textes constitutionnels48, à titre de présomption d’intention législative et de certains textes de loi, comme la Loi d’interprétation49. Même sans support textuel, personne aujourd’hui ne met en doute le pouvoir des juges de créer des règles. Il est très important, toutefois, de tenter d’identifier le fondement théorique de ce pouvoir, afin d’en délimiter la légitimité. Or, l’examen du fondement de la

---


croyance révèle une évolution dans les motifs que l’on allègue pour justifier le rôle de créateur de règles qu’assument de plus en plus ouvertement les tribunaux et notamment la Cour suprême.

Le fondement traditionnel du pouvoir est la nécessité inhérente à la fonction même de juger. Le juge qui tranche un litige invoque une règle pour justifier son jugement, pour démontrer que ce dernier n’est pas arbitraire. Pour la même raison, c’est-à-dire pour éviter l’arbitraire, la même règle est appliquée aux affaires subséquentes, à moins que ces dernières soient suffisamment différentes pour justifier un jugement différent. Quand il n’y a pas de règle établie, la nécessité de juger devient la nécessité de créer. Dans ce système, le jugement est le principal et la règle l’accessoire. La règle est un sous-produit du jugement. Il s’ensuit que la règle judiciaire ne sera légitime que dans la mesure où elle aura été nécessaire pour conduire le tribunal à son dispositif. C’est la norme qu’on peut appeler la norme décisionnelle.

La nécessité découlant de la fonction de juger est en voie d’être remplacée par une autre nécessité!: celle, générale et abstraite, du besoin de règles. Les partisans de cette théorie font état d’une vie sociale de plus en plus complexe, d’acteurs sociaux déboussolés qui ne savent plus ce qui est défendu ou permis, voire de groupes de justiciables tenus à l’écart du pouvoir et, partant, incapables de bénéficier de la bienveillance du législateur. Parce que ce besoin de règles existe et parce que les instances législatives ne sont pas en mesure de le combler, les tribunaux doivent colmater la brèche et produire des règles. La norme judiciaire cesse d’être un sous-produit de la nécessité de juger et devient un aspect autonome de la fonction judiciaire50. Elle cesse d’être accessoire et devient une des raisons d’être du jugement. Selon cette conception, la règle

50 Notons que cette conception de la création judiciaire de normes n’est pas rattachée à une tradition juridique plutôt qu’à une autre. Voir, notamment, Frédéric ZENATI, Méthodes du droit!— la jurisprudence, Paris, Dalloz, 1991. On ne peut prétendre non plus que le rule enrichment model est typique de la common law. Il est même plutôt atypique puisqu’il rejette la discipline du stare decisis fondée sur l’analyse du jugement par le juge interprète et sur la possibilité pour ce dernier de distinguer entre l’obiter dictum et la ratio decidendi; voir Matthieu DEVINAT, «L’autorité des obiter dicta de la Cour suprême», (1998) 77 R. du B. can. 1.
judiciaire n’est plus rattachée à sa matrice factuelle. Elle relève du pouvoir que le juge dérive de sa désignation comme juge: c’est la norme charismatique.

3. **Les tribunaux et la règle des trois unités**

Le remplacement de la norme décisionnelle par une norme charismatique aura des conséquences importantes sur la façon dont on fait le droit. L’ancienne norme forçait le tribunal à respecter une sorte de règle judiciaire des trois unités: unité de lieu, de temps, d’action. La décision était rendue dans un lieu prescrit, soit dans l’enceinte d’un tribunal. Le pouvoir judiciaire agissait à compter de l’introduction de l’instance jusqu’au jugement final, après quoi ses détenteurs étaient *functi officio*. La question à laquelle on devait répondre était celle qui avait donné lieu à la saisine du tribunal.

La norme charismatique élimine, à la rigueur, toute nécessité de respecter la règle des trois unités. Ces dernières n’ont plus leur raison d’être puisque l’autorité de la règle découle de la nécessité sociale de la produire et de la qualité du personnage qui la produit. Ainsi, en ce qui concerne le respect de l’unité de lieu, on pense tout de suite à l’affaire *Askov*51, où un juge de la Cour suprême, lors d’un colloque juridique à Cambridge, se plaint publiquement que les tribunaux inférieurs ont mal compris l’interprétation que la Cour suprême a donnée à la disposition constitutionnelle garantissant la tenue d’un procès sans délai abusif.

À partir du moment où le juge Cory fait cette déclaration, y a-t-il un seul pénaliste au Canada qui croit vraiment que l’arrêt *Askov* fait encore autorité? Qu’on le veuille ou non, à compter de la conférence de presse de Cambridge, le droit est changé. De fait, la Cour suprême confirme elle-même cette perception par son arrêt *Morin*52, rendu quelque temps plus tard.

---

Ce n’est pas le seul exemple. L’arrêt Stillman\textsuperscript{53} de la Cour suprême a donné lieu à des interprétations divergentes de la part des tribunaux inférieurs. Le juge Cory, toujours aussi soucieux de clarté, a mis les pendules à l’heure dans une conférence prononcée à Halifax, où il a précisé le sens que la Cour suprême voulait donner à la règle dans l’arrêt Stillman. La Cour d’appel de l’Ontario se fonde sur cette glose, qu’elle qualifie de «extra-judicial statement», dans un jugement (R. c. Lewis)\textsuperscript{54} qui applique l’arrêt Stillman. La Cour des Territoires du Yukon reprend l’explication donnée par le juge Cory en affirmant qu’il est légitime de le faire puisque la Cour d’appel de l’Ontario l’a maintenant incorporée dans une décision\textsuperscript{55}.

On note également des violations de la règle prescrivant l’unité de temps\textsuperscript{56}. Ainsi, dans l’arrêt Delaronde\textsuperscript{57}, la Cour suprême rend une décision le 30 janvier 1997, rejetant un pourvoi d’une décision de la Cour d’appel du Québec. La Cour motive le rejet par un renvoi aux motifs du juge Otis de la Cour d’appel\textsuperscript{58}. Mais, surprise! La note suivante apparaît au dossier:\textsuperscript{59} «L’addenda suivant du juge en chef Lamer a été déposé le 27 février 1997.» L’ajout en question précise la portée de l’adhésion de la Cour suprême à la décision de la Cour d’appel.

Dans la même veine, on retrouve la séquence des arrêts Feeney dont le premier introduit de nouvelles exigences de mandat en cas d’arrestation dans une maison d’habitation, le second accorde une «période transitoire» avant l’entrée en vigueur générale de cette règle (qui a pourtant bénéficié à Feeney), et le troisième, une prolongation du sursis d’exécution.

\textsuperscript{54} R. c. Lewis (1998), 13 C.R. (5th) 34 (C.A. Ont.).
\textsuperscript{55} R. c. Davies, [1998] Y.J. no. 64.
\textsuperscript{56} Voir: Hugo CYR, «Développements constitutionnels au Canada en 1997!—chronique» [à paraître]. L’auteur commente deux tendances qu’il constate à l’heure actuelle dans la jurisprudence de la Cour suprême!: l’usage de la suspension d’exécution et l’extension de la saisine.
«jusqu’au 19 décembre 1997 ou jusqu’à la date de la sanction royale du projet de loi C-16...»59.

Enfin, quant à l’unité d’action, on n’y pense même plus, depuis l’interprétation extravagante que les tribunaux (y compris la Cour suprême) ont donné à l’arrêt Sells60, les exemples abondent où les tribunaux statuent ultra petita. On a déjà pu prétendre que «A case is only authority for what it really decides»61; ce n’est plus la règle. Ainsi, dans l’affaire Delaronde, l’ajout du juge en chef porte sur «une question qui n’était pas en litige mais qui a été discutée devant notre Cour et qui pourrait éventuellement être soulevée dans une autre cause...»62. Comme on le voit, la manoeuvre avait pour but d’ajouter un obiter...

L’arrêt Hibbert63 est sans doute l’un des exemples les plus notoires de cette tendance. La question en litige est la défense fondée sur la contrainte par menaces. L’arrêt déborde largement cette question pour s’interroger sur le fondement théorique de ce moyen de défense, ses rapports avec d’autres moyens de défense comme la légitime défense, et sur la possibilité que la contrainte et l’intention criminelle puissent coexister. Le jugement se penche longuement sur les modes de participation à l’infraction et propose une lecture pour le moins originale de l’article 21 du Code criminel. La jurisprudence subséquente s’est inspirée de cet aspect de la décision64.

59 R. c. Feeney, [1997] 2 R.C.S. 13; R. c. Feeney, [1997] 2 R.C.S. 117; R. c. Feeney, [1997] 3 R.C.S. 1008. Que veut-on dire lorsqu’on parle de la suspension d’une règle prétorienne? Est-ce qu’on refuserait à un justiciable qui se trouve dans la même situation que Feeney le droit d’invoquer le même argument constitutionnel que lui? Comment concilier cette suspension avec la croyance naïve mais nécessaire à la légitimité que le droit s’applique également à tous?
Comme on le voit, ni le Parlement ni les tribunaux ne jouent leur rôle traditionnel en matière de création de règles. Le Parlement refuse de prendre ses responsabilités à l’égard des grandes orientations du droit. Il se confine à un rôle passif où il réagit aux mouvements de l’opinion publique. La Cour suprême, quant à elle, rejette la discipline de la tradition. Elle occupe tout l’espace que l’incurie du Parlement lui concède.

II. Les tares congénitales des normes répressives : l’obscurité et les légitimations artificielles

Il ne faut pas se surprendre si les normes créées sans égards pour les règles qui devraient régir cette création présentent des tares congénitales. Sur le seul plan technique, le Parlement est mieux situé que les cours pour dessiner les grandes orientations du droit. Il dispose de ressources beaucoup plus abondantes. Ses méthodes de travail et la façon dont les questions lui sont présentées lui permettent, en théorie du moins, d’envisager une multiplicité de solutions, alors que les tribunaux éprouvent la double restriction de la saisine par l’accusation et la procédure contradictoire. Il a à son service des légistes expérimentés, capables de produire des règles claires, cohérentes, conformes à l’idéal de notre société. Les tribunaux n’ont pas ces avantages. Il est donc normal de s’attendre à ce que les règles fondamentales du système pénal soient produites par l’organisme législatif.

À cet égard, il ne faut pas oublier que c’est la procédure et la preuve pénales qui font l’orgueil de la common law et non la théorie générale de la responsabilité. Dans ces domaines intimement liés à la production du jugement, les tribunaux ont la compétence voulue. On ne peut dire, toutefois, que la théorie pénale produite dans le cadre des droits de la common law est une des réalisations marquantes de l’esprit humain. Le droit pénal général canadien méritait amplement les critiques que les partisans de la réforme lui adressaient dans les années 70 et 80; comme la réforme a fait long feu, on présume qu’il les mérite encore. De fait, la situation s’est aggravée, notamment parce que la mission confiée au pouvoir judiciaire par la Charte a accéléré le processus de création judiciaire des normes fondamentales du droit répressif.
A. Obscurité et légitimité

1. Un cas de figure : la mens rea

Les normes produites selon les modes que nous avons décrits ne sont ni cohérentes ni claires. Or, en matière répressive, l’absence de clarté est plus qu’un défaut de forme : c’est une violation du principe de la légalité, principe à la fois politique et moral. Qui plus est, l’obscurité des normes crée une distance entre le droit et les justiciables.

L’exemple le plus éloquent de l’obscurité des normes fondamentales du droit répressif est le phénomène de la mens rea. Je dis bien phénomène car la mens rea est bien plus qu’une simple notion juridique. Elle est à la fois un mécanisme organisateur des infractions pénales, un principe moral et un champ de bataille idéologique.

Pour ceux qui ne sont pas des adeptes du gai savoir qu’est le droit pénal, l’expression « mens rea » aurait été employée la première fois par Saint Augustin, au sujet du parjure : « Lingua non facit ream, nisi sit mens rea. » En d’autres termes, pour mentir, il faut vouloir tromper. En droit anglais, « mens rea » est le nom qu’on a donné à la composante psychologique de l’infraction pénale, composante qui confère en même temps à cette dernière une signification morale. Par exemple, la différence entre un accident et un crime réside dans l’intention de l’auteur de l’acte. De même, la différence entre un meurtre et la négligence criminelle causant la mort se situe dans l’esprit qui a animé l’acte : l’auteur du premier a fait exprès, l’auteur du second a manqué gravement à son devoir de prudence.

Les difficultés que soulève cette notion fort simple et fort utile viennent de ce qu’on lui fait jouer un rôle démesurément important dans la dogmatique pénale de la

---

65 Encore que l’arrêt R. c. Nova Scotia Pharmaceutical Society, précité, note 25, ait considérablement émasculé ce principe en réduisant les exigences de clarté en matière pénale au critère de la justiciabilité. Il demeure que dans un droit pénal fait pour les justiciables plutôt que pour les professionnels du droit, les gens devraient avoir une bonne idée des principes fondamentaux de la responsabilité pénale.
common law. Dans les droits de l’Europe continentale, les infractions pénales comportent une composante psychologique appelée «élément moral». Par ailleurs, dans ces systèmes, la responsabilité pénale fait aussi appel à d’autres notions, comme l’illicéité et le blâme, auxquelles se rapportent des moyens de défense classés sous les rubriques de la justification et de l’excuse. Par conséquent, l’analyse de l’élément moral de l’infraction n’est qu’une étape préliminaire de la démarche conduisant à établir la culpabilité criminelle.

Dans les droits issus de la common law, la *mens rea* est l’alpha et l’oméga de la dogmatique pénale. Traditionnellement, la plupart des défenses s’expriment en termes d’insuffisance de la preuve de la *mens rea*. Ainsi, l’*error factis*, la provocation ou l’ivresse ont leur fondement théorique dans cette notion. Même si le droit canadien connait l’excuse et la justification, ces notions n’ont ni l’importance théorique ni l’efficacité qu’elles prennent dans d’autres systèmes.

Pourquoi la *mens rea* est-elle surutilisée dans les systèmes dérivés de la common law? Tout simplement parce que, dans ces systèmes, le procès pénal est le cadre dans lequel le droit se fabrique. Les matériaux de cette fabrication sont l’accusation, la procédure et la preuve. La partie générale du droit pénal est quasiment embryonnaire en common law. L’analyse du libellé de l’infraction constitue le seul mécanisme permettant au tribunal de tenir compte des variations de la culpabilité morale de l’accusé. Le juge doit donc se rabattre sur la partie spéciale et tenter de dégager les règles générales de la responsabilité à partir de la définition des infractions.

C’est pour cette raison que les deux grandes problématiques du droit pénal anglo-canadien tournent autour de cette question. D’abord, dans quelle mesure la *mens rea* est-elle une composante de l’infraction, ensuite, quelle est la nature exacte de cette composante? Avant l’adoption de la Charte, en 1982, la *mens rea* et, partant, la possibilité d’invoquer une défense logée à son enseigne, repose ou bien sur le libellé de l’infraction, ou bien sur le postulat qu’elle était voulue par le

---

législateur. Ensuite, il s’agit de déterminer en quoi la mens rea consiste. S’agit-il du dol ou de la faute? Quelle gradation exacte de l’intention, de l’insouciance ou de la négligence la Couronne doit-elle prouver? L’infraction exige-t-elle la connaissance ou l’aveuglement volontaire, ce qui permet la défense d’erreur pourvu qu’elle soit sincère, ou la mens rea en est-elle une d’imprudence, auquel cas l’erreur, en plus d’être sincère doit être raisonnable. Mais attention! Il pourrait s’agir d’une infraction de responsabilité stricte, qui dispense le ministère public de la preuve de l’élément moral, tout en permettant l’exonération de l’accusé si ce dernier arrive à prouver par prépondérance qu’il a fait diligence.

L’adoption de la Charte a fait de la mens rea une exigence constitutionnelle. La situation est-elle plus simple? Depuis les premiers arrêts qui ont classé l’exigence de la mens rea parmi les principes de justice fondamentale, la Cour suprême a tellement multiplié les règles d’interprétation qu’Occam aurait plutôt besoin d’une faux que d’un rasoir pour y mettre de l’ordre. Ainsi, l’infraction exige-t-elle qu’on prouve une mens rea de prévisibilité subjective ou simplement la mens rea minimale? Seules les infractions comportant des stigmates particulières exigent, constitutionnellement, une mens rea plus que minimale. Quelles sont les infractions suffisamment stigmatisantes? Le meurtre et la tentative de meurtre, certes, et peut-être le vol, mais pas l’agression sexuelle, qui entraîne «un important degré d’opprobre moral», mais dont les stigmates n’ont pas le même effet constitutionnel. Pour ces infractions insuffisamment stigmatisantes, une mens rea de prévisibilité objective suffit, encore qu’il pourrait toujours s’agir d’un «critère objectif modifié». Qu’est-ce qu’un «critère objectif modifié»? Le juge en chef, dans l’arrêt Creighton, en propose une définition, légèrement différente de celle de madame McLachlin, dans le même arrêt67.

À ce foisonnement de distinctions et de subtilités, qui serait moins déplacé dans des textes de doctrine (qu’on peut à la rigueur ignorer), que dans des arrêts qui constituent des sources de droit, on peut opposer la solution élégante et nuancée du Code pénal français dont l’article 121-3 énonce:

Il n’y a point de crime ou de délit sans intention de le commettre.

Toutefois, lorsque la loi le prévoit, il y a délit en cas d’imprudence, de négligence ou de mise en danger délibérée de la personne d’autrui.

Il n’y a point de contravention en cas de force majeure.

L’article du Code pénal français liquide élégamment en trois phrases un contentieux qu’entretient depuis plus de 100 ans la jurisprudence anglo-canadienne.

2. L’obscurité et ses séquelles

Quelle est la signification de cette obscurité pour la légitimité de la loi pénale?

En premier lieu, elle occulte les bases de la responsabilité pénale. Le principe selon lequel on ne doit pas punir un innocent, c’est-à-dire celui qui n’a pas volontairement mal agi, ne transparaît pas à travers les décisions judiciaires, essentiellement fondées sur une vision analytique des infractions criminelles. Par exemple, la règle selon laquelle un acteur ivre est considéré irresponsable si l’infraction dont on l’accuse en est une « d’intention spécifique » est inintelligible pour la plupart des juristes, ne parlons pas du grand public. On peut dire la même chose de la règle concernant l’ivresse extrême; la réaction du public à la décision Davault, à peu près universellement hostile, paraît symptomatique d’une incompréhension profonde. L’incompréhension entraîne l’indignation et c’est moins la norme elle-même que l’impossibilité pour ses producteurs de la justifier qui cause scandale.

Ensuite, l’obscurité de la norme rend nécessaire des explications qui, sous prétexte d’éclairer, génèrent des quantités d’énoncés théoriques; or, ces énoncés portent en eux-mêmes la promesse de conflits futurs. En effet, comme on n’arrive plus à distinguer ce qui est règle de droit de ce qui est opinion ou doctrine, ces énoncés trouveront éventuellement un contexte où ils paraîtront obscurs et généreront à leur tour d’autres tentatives d’explications. Plus les règles se multiplient, plus la norme se
La légitimité des sources du droit pénal

complexifie, moins elle risque d’être comprise et plus le fossé se creuse entre les justiciables et la justice.

L’obscurité entraîne aussi des séquelles moins évidentes. Comme une partie de plus en plus considérable du discours juridique est consacrée à la dissiper (ou du moins à la commenter!), le débat est moins orienté sur le sens de la responsabilité pénale que sur d’autres objets plus ésotériques et spécialisés. Ainsi, en changeant l’objet du débat, en insistant, par exemple, sur le contenu conceptuel de l’infraction plutôt que sur la situation humaine qui est à l’origine de l’accusation, on supprime toute possibilité d’arriver à d’autres solutions au problème de la responsabilité. Par exemple, toujours en matière d’ivresse, le droit pénal ne fait aucun effort pour composer avec la notion d’alcoolisme même si l’alcoolisme, dans un autre contexte, est considéré justiciable de traitements médicaux. Si les seuls mécanismes permettant d’appréhender l’alcoolisme sont ceux qui ont été élaborés dans le cadre d’une accusation criminelle, on ne doit pas être surpris de leur manque de nuance: après tout, un procès pénal n’a que deux issues.

B. Les légitimations artificielles

Parce que les normes ne sont pas claires et parce que la forme qu’elles revêtent empêche qu’elles le soient, leur légitimité n’est pas manifeste. C’est pourquoi les producteurs de la norme tentent de leur fabriquer une légitimité artificielle. De plus en plus, un discours de légitimation accompagne la norme. Ce discours prend des formes différentes selon l’instance qui le prononce.

En toute justice, le refus de considérer l’alcoolisme comme une condition pertinente dans l’appréciation de la responsabilité de l’acteur ivre n’est pas propre au droit canadien. Des droits dotés d’un appareil théorique beaucoup plus élaboré n’ont pas trouvé non plus de solution à ce problème. Voir, pour le droit pénal français, Georges LEVASSEUR, Albert CHAVANNE, Jean MONTREUIL, Bernard BOULOC, Droit pénal général et procédure pénale, 13e éd., Paris, Sirey, 1998, n° 198, p.74. Néanmoins, les chances de le faire sont meilleures dans un système où la notion d’excuse est plus développée que dans un système où la majorité des moyens de défense se réduisent au concept de la mens rea.
1. Légitation et motifs surabondants

Ainsi, la jurisprudence contient de plus en plus souvent des références à des légítimations externes. Bon nombre de ces légítimations débordent le domaine strictement juridique et s’inspirent de la philosophie, de l’histoire, des sciences humaines. Or, en avançant des justifications non juridiques, en invoquant des raisons enracinées dans des disciplines que ni leur expérience ni leur formation ne leur ont permis de maîtriser, les juges prétent le flanc aux critiques des spécialistes et finissent par diminuer la crédibilité de leurs jugements: les mauvaises raisons chassent les bonnes.

On connaît les critiques que certains historiens ont formulées au sujet de l’utilisation de l’histoire par les tribunaux; d’autres auteurs ont signalé l’usage intempestif par la Cour des citations de philosophes, critiquables tant sur le plan de la pertinence que sur celui du respect du contexte. À l’heure actuelle, ce sont les sciences humaines qui sont à l’honneur. On a de plus en plus recours à des arguments fondés sur des théories socio-psychologiques pour justifier les normes pénales. Or, l’utilisation que font les tribunaux des thèses scientifiques est dénoncée avec vigueur par bon nombre d’auteurs. La difficulté vient de ce que bon nombre de ces thèses ne font pas l’unanimité chez les chercheurs. En les adoptant comme justification, les tribunaux se trouvent à créer par ukase une sorte de science officielle. Le moins qu’on puisse dire de ce procédé, c’est qu’il est anti-scientifique.

---


2. Légitimation et préambules

On retrouve le même souci de légitimation artificielle du côté de la législation pénale, où il se manifeste par la prolifération des préambules. Ce genre spécialisé de la littérature juridique apparaît souvent dans la législation «réactive» où le gouvernement cherche à se rallier la sympathie du public. Certains les jugent inutiles, voire trompeurs!

Quand il y a un préambule dans une mesure législative, c’est généralement un écran de fumée politique dressé par le ministère qui a rédigé la mesure législative en question, en principe le ministère de la Justice.\textsuperscript{72}

Au cours de la dernière décennie, le Parlement du Canada a gratifié les justiciables d’une importante quantité de préambules, notamment dans le domaine pénal. Ces préambules, dont la \textit{Loi d’interprétation} nous apprend qu’ils «[font]...partie du texte et en constitue[nt] l’exposé des motifs (en anglais!: «shall be read as part of an enactment intended to assist in explaining its purport and object»), contiennent des attendus qui nous apparaissent assez peu efficaces comme guides de l’interprétation. Ils nous apprennent surtout que le Parlement a le coeur sensible et que le gouvernement a écouté les groupes de pression qui lui ont demandé d’adopter une loi. Ainsi que «les cas de violence et d’exploitation sexuelles préoccupent sérieusement le Parlement du Canada […]\textsuperscript{73}, que «la violence au sein de la société canadienne préoccupe sérieusement le Parlement du Canada»\textsuperscript{74}, que «les cas de violence et d’exploitation sexuelles au sein de la société canadienne continuent de préoccuper sérieusement le Parlement du Canada»\textsuperscript{75} (les lois précédentes n’ayant sans doute pas eu l’effet escompté), ces préoccupations, toutes électorales qu’elles soient, font honneur au Parlement, mais n’éclairent pas…

\textsuperscript{72} Intervention de M. John BRYDEN, Journal des débats de la Chambre des communes, 7 octobre 1997. Ce qui est étonnant dans cette intervention, c’est que le député qui la fait est du parti ministériel.

\textsuperscript{73} \textit{Loi modifiant le Code criminel (agression sexuelle)}, précitée, note 37.

\textsuperscript{74} \textit{Loi modifiant le Code criminel (intoxication volontaire)}, précitée, note 39.

\textsuperscript{75} \textit{Loi modifiant le Code criminel (communication de dossiers dans les cas d’infraction d’ordre sexuel)}, L.C. 1997, c. 30
l’interprète qui cherche à préciser le sens et la portée de la nouvelle loi\textsuperscript{76}.

En réalité, en plus de démontrer la bonne volonté des gouvernements, les préambules ont une mission plus ou moins occulte: ils cherchent ( naïvement, à mon sens) à immuniser la loi contre une éventuelle contestation constitutionnelle tout en dispensant le gouvernement d’invoquer l’article 33 de la Charte constitutionnelle. C’est peut-être la raison d’être de toutes ces <<préoccupations>>, qui sont certainement <<urgentes et réelles>> au sens de l’arrêt \textit{Oakes}\textsuperscript{77}.

C’est sans doute pourquoi les préambules font souvent allusion aux valeurs constitutionnelles que la loi prétend protéger et qu’on oppose généralement aux garanties juridiques de la Charte!: notons les second et sixième attendus de la loi abrogeant la règle de l’arrêt \textit{Daviault}, qui citent expressément les articles 7, 11, 15 et 28 de la Charte.

Le procédé est grossier et se révèlera, espérons-le, inefficace. Il est inconcevable que le Parlement du Canada puisse résister à une interprétation de la Cour en déclarant unilatéralement qu’une loi ne viole pas un principe de justice naturelle, \textit{contrairement à ce que la Cour suprême a déjà affirmé}, ou qu’elle devrait être protégée par l’article premier, \textit{ce que la Cour suprême a déjà nié}. Si le Parlement estime sage, ou vertueux, ou rentable d’abolir la règle de l’arrêt \textit{Daviault}, il n’a qu’à adopter une loi qui déroge expressément à la Charte aux termes de l’article 33. Encore faudrait-il que le gouvernement assume ses responsabilités en présentant une telle loi à la Chambre des communes. C’est sans doute là l’enjeu véritable de certains préambules et, notamment, celui de la loi modifiant la défense d’ivresse!: dispenser le gouvernement d’invoquer l’article 33. La manoeuvre est transparente et il serait surprenant que la Cour suprême, qui aura sans doute à se prononcer un jour sur la validité constitutionnelle de la modification, risque sa


réputation d’indépendance judiciaire en accepte de tirer les marrons du feu.

L’utilisation de plus en plus fréquente des préambules entraîne des difficultés techniques considérables lorsque ces préambules chapeaute des modifications au Code criminel. Qu’arrive-t-il au préambule d’une loi modificatrice lorsque la modification est adoptée? Le préambule et la loi modifiée se disent adieu; les modifications se retrouvent au Code et le préambule dans les recueils annuels des lois du Canada. Ce résultat, celui de la multiplication des sources matérielles du droit, est précisément ce que la codification cherche à éviter. Un code, au bas mot, est un document qui facilite la consultation des sources du droit. Par conséquent, chaque nouveau préambule, qui oblige à des efforts de recherche dans des sources autres que le Code, est un clou dans le cercueil de la codification.

Conclusion

Cette réflexion sur la légitimité des sources du droit pénal nous amène à des conclusions assez pessimistes. Il semble exister de sérieux problèmes de légitimité, tant en ce qui concerne le droit légiféré que le droit prétorien. De plus, comme ces problèmes paraissent reliés à l’appareil de production des normes, on voit mal comment la situation va s’améliorer. Au contraire, puisque c’est la production même des normes qui est en cause, on peut s’attendre à ce que les difficultés se reproduisent de façon constante.

La constitutionnalisation des principes généraux va, elle aussi, contribuer à perpétuer les dysfonctions constatées. Ainsi, le tribunal qui innove va très souvent revendiquer l’autorité de la Charte à l’appui de sa création. La norme prend donc une pesanteur que les principes généraux n’avaient pas dans le système antérieur. L’interprétation de la loi par la Cour suprême devient la loi suprême. Il en résulte une situation très délicate pour le Parlement et pour les interprètes subséquents de la jurisprudence constitutionnelle. Comment, en effet, révoquer une norme à laquelle on a reconnu un statut constitutionnel? Si la création par le Parlement d’une nouvelle partie générale du Code a déjà pu être envisagée, on peut se demander si l’entreprise est aujourd’hui réalisable.
Pourtant, même si les normes créées par les tribunaux participent au statut de la loi suprême, la profusion de normes entraînée par la nouvelle perception du rôle des tribunaux diminue leur vie utile. Qui invoque aujourd’hui les arrêts qui, hier, affirmaient les principes traditionnels du droit pénal? Il y a certainement là un paradoxe puisque ce sont ces sources traditionnelles qui constituent les principes de justice fondamentale de l’article 7. On dirait que la marge de manœuvre que les tribunaux se sont reconnue pour créer des normes les autorise à ignorer, le cas échéant, leurs créations.

Si la situation change, ce ne sera ni pour des raisons d’*elegantia juris* ni pour des soucis de légalité ou de légitimité. Ce sont des motifs pratiques qui ont été à l’origine des grandes réformes de nos systèmes. La confusion des sources a été le moteur de bon nombre de codifications, y compris celle du Bas Canada et celle de 1892, en droit pénal. Si le droit en vient au point où il risque de crouler sous ses sources, ses arrêts, ses opinions, ses lois ponctuelles, ses annexes et ses préambules, en un mot, sous le poids écrasant de son discours, il y aura changement. En attendant, faut-il espérer que les artisans de la règle continuent à pratiquer la politique du pire?
The Plain Meaning Rule and Other Ways to Cheat at Statutory Interpretation

Ruth Sullivan*

Introduction ........................................................................................................... 153
I. Basic Assumptions .............................................................................................. 153
II. The Plain Meaning Rule (PMR) ................................................................. 155
III. Advantages of the Plain Meaning Rule ....................................................... 156
IV. Fidelity to Legislative Intent (FLI) ............................................................... 157
V. Advantages of Fidelity to Legislative Intent .................................................. 159
VI. The Promise of Certainty .............................................................................. 159
VII. The Reality .................................................................................................... 160
VIII. How to Interpret a Text While Pretending Not To .................................. 161

A. PMR Trick # 1 — Artful Text Selection ....................................................... 161
B. PMR Trick # 2 — Elastic Co-Text ................................................................. 164
C. PMR Trick # 3 — The Shifting Meaning Game ........................................... 168

* Professor of Law, University of Ottawa.
D. PMR Trick # 4 — It Must Be Plain To You If It’s Plain To Me .............................. 170
E. PMR Trick # 5 — The Inherent Meaning Illusion.................................................. 172
F. PMR Trick # 6 — Abandoning Ship......... 174

Conclusion.......................................................... 177

IX. How To Be Faithful to the Legislature Without Giving Up Your Freedom ........ 177
A. FLI Trick # 1 — Presume That Nothing Worth Mentioning Has Changed..................... 177
B. FLI Trick # 2 — More Meaning Games ........................................................................ 178
C. FLI Trick # 3 — Distinguish Sloppy Drafting From Legislative Error ..................... 180
D. FLI Trick # 4 — Presume That the Legislature Wants What You Want ............. 184

Conclusion.......................................................... 186

What Should Be Done?.............................................. 187
Anyone who spends time studying language and interpretation soon comes to appreciate the complexity and uncertainty inherent in any effort to interpret a text. While in some contexts (the study of poetry, for example) complexity and uncertainty are valued, when it comes to disputes about the interpretation of legislative texts, these features are a cause of dismay!— and a certain amount of dissembling. It is this last response to complexity and uncertainty that I focus on in this presentation. I am interested in the ways superior courts, which are the official interpreters of legislative texts, attempt to disguise the complexity and uncertainty of interpretation with doctrines like the plain meaning rule and fidelity to legislative intent. Although these doctrines do not fully or accurately reflect what judges must do to resolve interpretation disputes, they generate a reassuring rhetoric. If we can’t have simple messages and certain outcomes, reassuring rhetoric is the next best thing.

I. Basic Assumptions

There are two key assumptions relied on by courts to explain and justify their work in statutory interpretation. One is the assumption that meaning inheres in legislative texts and that at least some of the time meaning is “plain”!— that is, clear and certain, not susceptible of doubt. This assumption is the necessary basis for the plain meaning rule. The other assumption is that legislatures have intentions when they enact legislation and these intentions are knowable by courts when called on to interpret legislation. This assumption is the necessary basis for the doctrine of fidelity to legislative intent.

Although these assumptions are repeatedly challenged by academics like myself, the courts are not inclined to give them up, for they are actually quite useful. The assumptions tell us that interpretation is rooted in something definite!— the text!— which has been fixed once and for all by the legislature. They thus support a view of statutory interpretation that fits the current mythology of western democracy. In a democracy such as ours, persons and their activities are governed not by the whim of rulers but by the rule of law. We know that the rule of law is legitimate because law is made by the elected representatives of the people, at the end of a democratic process.
We know that the rule of law is fair and efficient because law is issued in the form of texts which are fixed and stable, published in advance and applied equally to everyone under the supervision of the courts. Finally, we know that courts are well placed to supervise the application of law because of their impartiality and their interpretive expertise.

It is obvious to anyone who reads the newspapers or otherwise follows public affairs that rule of law and the related ideals of democracy, fairness, equality and efficiency are highly prized by Canadians. These are the virtues in which government tries to dress itself and there is little incentive for those who are part of government to challenge the mythology or to suggest that those in power are not wearing clothes. Like the other branches of government, the judiciary is reluctant to draw attention to possible discrepancies between the rule of law ideal and its own practice in interpreting and applying legislation. Furthermore, there are dangers in too much iconoclasm. If law or its enforcement is perceived to be illegitimate or unfair, or if compliance with law is perceived to be pointless or impossible, there will be little incentive to obey the law. And despite our large police force and crowded jails, the rule of law in western democracies is fundamentally self-enforcing. This is why the rhetoric of law is important. Any attack on the assumptions and doctrines which sustain our ideals must therefore proceed with care.

In this presentation I argue that the plain meaning rule and the doctrine of fidelity to legislative intent are misleading in certain ways and force the courts to rely on a number of “tricks” which disguise the important role played by discretion in statutory interpretation. I identify and describe these tricks and give some examples of how they are done. I complain that the courts’ reliance on these tricks is a form of “cheating”. However, mindful of the dangers of too much iconoclasm, I end by suggesting that there is really no need for the courts to cheat. With fairly minor revisions to their basic assumption, it would be possible for courts to acknowledge the role of discretion in statutory interpretation without putting the virtues of democracy and rule of law in jeopardy.
II. The Plain Meaning Rule (PMR)

The plain meaning says that if the meaning of a legislative text is plain, the court may not interpret it but must simply apply it as written. The court may resort to the rules and techniques of interpretation only if the text is ambiguous.

This rule presupposes that there is an important difference between the first-impression meaning of a text and its post-interpretation meaning. First impression meaning is meaning that spontaneously comes to mind when a person reads a text relying on nothing but the text and her own linguistic competence. Post-interpretation meaning is meaning that is constructed by a person through interpretation relying on factors other than the text itself—factors like the imagined purpose of a text, or its possible consequences, or extrinsic aids like legislative history. According to the plain meaning rule, when a person sets out to resolve a dispute about the correct interpretation of a legal text, the first thing she must do is read the text and form an impression of its meaning based on reading alone. She must then judge whether this meaning is plain. A text has a plain meaning if a competent reader would judge, on the basis of reading alone, that her first impression meaning is the only meaning the text can plausibly bear. A text is ambiguous if a competent reader could plausibly read it in more than one way.

When a legislative text has a plain meaning, the courts are prohibited from interpreting it; they are bound by the words of the text. Given this fundamental rule, it is perhaps surprising that even when the meaning is plain, the courts may look at non-textual evidence of legislative intent (like purpose or consequences or extrinsic aids) and may rely on these other factors to support the plain meaning. What they may not do is rely on these factors to vary or contradict the plain meaning of the text.

This approach to extra-textual evidence is convenient if you are a proponent of the plain meaning rule: heads you win, tails you still win. The plain meaning rule has many such tricks.

What is distinct and important about the plain meaning rule is its claim that the court’s first responsibility is to give
effect to the apparent meaning of the text whenever this meaning is plain. If there is a conflict between what the text appears to say and what the legislature seems to have intended, the text wins. Intention governs only when the text itself is ambiguous.

III. Advantages of the Plain Meaning Rule

The great advantage of the plain meaning rule is that, in theory at least, it creates a zone of certainty!—an interpretation-free zone, in effect. It tells the public that if the text is plain, it means what it says and it is safe to rely on it. The courts won’t come along and trick you at the last minute by importing unsuspected qualifications or implications into the text, even if these qualifications or implications were probably intended by the legislature. This emphasis on text at the expense of intention ensures that the law is certain and that the public has fair notice, both of which are prerequisites for effective law.

A second advantage of the theory is that it supports formal equality. If the courts are bound by plain meaning, and plain meaning is the same for everyone, it follows that a rule whose meaning is plain expresses the same law for everyone and will be applied in the same way, to the same effect, to everyone. Plain meaning thus creates a level playing field on which we all have equal opportunity to score.

A third advantage of the plain meaning rule is that it can be used as an apparently neutral proxy for strict construction. It is no coincidence that the plain meaning rule is applied most persistently and enthusiastically to fiscal legislation, with penal legislation running a close second. The courts have always been reluctant to assist the other branches of government in limiting the freedom of subjects or taking away their property. Historically in interpretation these judicial values have been given bite primarily through the doctrine of strict construction. However, in many instances a court can achieve the same result by relying on the plain meaning rule without having to invoke the common law values. By claiming that the text is plain and does not require interpretation, a court purports to base the outcome in a case on linguistic competence alone. The appeal to judge-made law and policy is replaced by an exercise in technical expertise.
IV. Fidelity to Legislative Intent (FLI)

Although the plain meaning rule is often evoked, it is not accepted by everyone. For many judges, the touchstone of interpretation is the intention of the legislature. The legislature may reveal its intentions directly, for example by explaining them in a preamble or a purpose statement. Some would argue that Ministerial statements addressed to the legislature during the passage of a bill also provide direct evidence of legislative intent. Generally speaking, however, legislative intent is discovered indirectly through inference: having regard to xxx, a person who said aaa in context yyy when speaking to audience zzz must have meant bbb. To draw inferences of this sort, even competent speakers must rely on a wide range of contextual factors, both textual and extra-textual.

If it appears that a text has a clear precise meaning, an intentionalist judge would normally attach significant weight to that factor, but she would still be obliged to consider other factors such as the legislative evolution of the text or its real world consequences. If this extra-textual evidence of legislative intent turned out to be sufficiently compelling, the intentionalist would have to accept it. What is distinct and important about intentionalism is that it gives top priority not to the apparent meaning of the text, but to the meaning it was intended to have. The intentionalist does her best to carry out the real intentions of the legislature.

Intentionalism treats legislation as if it were an ordinary speech act, that is, an act of communication occurring at a particular place and time in which a speaker attempts to evoke particular meanings in the mind of an audience. It assumes that speaker and audience are operating within a context that is more or less shared or is at least accessible to both. This context consists of shared linguistic, social and cultural conventions and shared historical experience. Intentionalism further assumes that the speaker is aware of the context in which the audience will carry out its interpretation and ensures accurate communication by anticipating the inferences the audience will draw when reading different combinations of words in that context. Although there is always the possibility of a misunderstanding or mistake, the speaker relies on the audience to interpret cooperatively.
A cooperative interpreter makes a good faith effort to reconstruct the actual intentions of the speaker, sorting out contradictions when necessary, correcting the speaker’s mistakes, looking for the implications implicit in her choice of words. If I say “it’s hot in here”, I may intend simply to report on the temperature in my immediate surroundings; or I may intend to communicate my discomfort; or I may “really” be asking the other person in the room to open the window or turn on the fan. It’s even possible that I wanted him to turn on the space heater and my use of “hot” was just a slip; I meant to say “cold”. A cooperative interpreter attempts to piece together the speaker’s actual meaning from all the available evidence.

Judges who accept the doctrine of fidelity to legislative intent are driven to accept a number of corollary doctrines including the original meaning rule, the doctrine of presumed intent and the distinction between interpretation and amendment. It is easy to see why an intentionalist approach to interpretation requires something like the original meaning rule. If a legislative text is designed to communicate a particular meaning, one that was actually intended by the legislature on the occasion of enactment, that meaning must be fixed once and for all at the moment of enactment; it cannot subsequently change. It is also easy to see why the doctrine of presumed intent is important to the intentionalist approach. This doctrine is needed to establish the legislature’s connection to the context in which judicial interpretation occurs. A court can infer the meaning that was actually intended by the legislature only if its inferences are grounded in assumptions, values and conventions that were actually shared by the legislature at the time of enactment. By presuming the legislature intended to comply with whatever policies, values and conventions are relied on by courts from time to time, this connection is assured. Finally, it is easy to see why an intentionalist draws a sharp distinction between correcting or reading down a text so that it conforms to the actual intention of the legislature and reading in or filling gaps so as to supplement or change that intention. Legislative intent is a fact waiting to be discovered; and the job of the court is to discover it, not make it up.
V. Advantages of Fidelity to Legislative Intent

The main function of intentionalism, and its chief advantage, is that it explains the role of courts in interpretation without challenging our popular conceptions of democracy. Most of us accept that for law to be legitimate it must be democratic, and for law to be democratic it must be made by the elected representatives of the people. If this is so, then the only legitimate source of law is the legislature; judge-made law is undemocratic and therefore illegitimate. While this reasoning has the potential to cause some serious problems, under the doctrine of fidelity to legislative intent, the question of judicial law-making is answered before it can arise. Judges do not make law; they only apply it. Being impartial and unbiased, they use their legal interpretation skills to discover and give effect to the intention of the legislature. They have neither the need nor the mandate to engage in political choice.

A second advantage of the intentionalist doctrine is that it supports the positivist view of law, which is a major bulwark of certainty. Positivism tells us that law is one thing and not-law is something else entirely and we can easily tell them apart by applying a simple rule of recognition. What the legislature enacts is law. Legal values, preferred policies, judicial assumptions – these things are not law and therefore play no role in the governance of people’s lives.

VI. The Promise of Certainty

By suggesting that outcomes are dictated by a fixed text or a fixed legislative intent, the plain meaning rule and the doctrine of fidelity to legislative intent both promise certainty. The rules are embodied in black and white; and within the interpretation-free zone at least, they are the same for everyone. What you see is what you get.

Even when the meaning is not plain and interpretation is required, the judges who carry out this interpretation are constrained by the fixed intention of the legislature. Upon enactment, the content of the law is fixed once and for all; through interpretation the judges discover that content!— not change it, not create it!— and ensure that it applies equally to all.
At least that is how it goes in rule of law heaven. What are things like here on earth?

VII. The Reality

Here on earth it is easy to appreciate the rhetorical value of the plain meaning rule and the doctrine of fidelity to legislative intent. It is also easy to appreciate the importance of rhetoric in a system that relies primarily on hegemony rather than force to induce compliance. Nonetheless, despite the benefits they bestow, I object to these doctrines because in my view they force the courts to “cheat”. Since this is strong language, I should take a moment to explain what I mean by cheating.

When a student cheats on a paper or exam, she takes someone else’s answer and tries to pass it off as her own; she tries to claim the credit for someone else’s work. Cheating in this sense is a combination of stealing and misrepresentation. I don’t suggest that judges cheat in this way. In fact, what judges do is rather the reverse. They take their own answer!— the outcome in a statutory interpretation dispute!— and try to pass it off as someone else’s. They try to deny responsibility for their work. Cheating in this sense is not a form of stealing, but it is a form of misrepresentation.

When judges invoke the plain meaning rule, they deny responsibility for the outcome by saying that the text made them do it. “The text is clear, its meaning is plain, I have no choice. If you don’t like the outcome, tell the drafter who wrote the text or the government officials who told her what to write or the legislature which enacted the text into law.” When judges invoke the doctrine of legislative intent, they transfer responsibility to the legislature. “I don’t make the laws, I only apply them. I am a legal technician, a mere servant of my legislative master. In democracies like Quebec or Canada, the legislatures call the shots.”

These doctrines are objectionable because they discourage judges from probing and discussing the real basis for outcomes in statutory interpretation disputes. They encourage judges to deny or misrepresent the choices they are obliged to
make and to avoid responsibility for outcomes. In my utopia, judges acknowledge their responsibility and attempt to explain and justify their choices.

These complaints about “cheating” are hardly new. Complaints of this sort have been made over and over again, off and on, for most of this century. For the most part they are ignored by the courts. The question that interests me in this presentation is how the courts get away with it. Usually when someone is caught cheating, they are obliged to take a different approach. But the courts go right on invoking the same doctrines and using the same rhetoric year after year, more or less ignoring academics like myself. I think I know why they do it. They do it for the reassuring rhetoric. By relying on these doctrines, they provide credible resolutions to interpretation disputes without upsetting any applecarts.

The question I want to consider in greater detail have is how they do it. I want to draw attention to the techniques used in statutory interpretation to make choice look like choicelessness, to make discretion look like constraint, and to make judge-made law look like the intention of the legislature. I will start with the plain meaning rule, which in my view is sly and pernicious and deserving of much criticism. I will then look at some of the problematic aspects of intentionalism, in particular the original meaning rule, the doctrine of presumed intent, and the equivocal distinction between drafting mistakes and legislative oversights. My presentation will end with a call for a better rhetoric, one that doesn’t force judges to cheat.

VIII. How to Interpret a Text While Pretending Not To

A. PMR Trick # 1 — Artful Text Selection

There are many “texts” in statutory interpretation. The entire body of legislation produced by a legislature constitutes a text, as do particular statutes and particular provisions of statutes. The first step in applying the plain meaning rule is to identify the text-to-be-interpreted. This consists of the words whose meaning has been put at issue by an attempt to apply legislation to particular facts.
In *R. v. McCraw*,¹ for example, the facts were that an accused wrote letters to young women which said (using cruder and more aggressive language), “I’m going to have sex with you even if I have to rape you”. The legislation was s. 264.1 of the *Criminal Code* which said:

*Everyone commits an offence who, in any manner, knowingly utters ... a threat*

(a) *to cause death or serious bodily harm to any person*;

(b) *to burn, destroy or damage real or personal property*; or

(c) *to kill, poison or injure an animal or bird that is the property of any person*.

In applying this language to these facts, everyone agreed that the young woman was a “person” and that the letter contained a “threat” that was “knowingly uttered”. Everyone also agreed that the threat was not a threat of “death” or of interference with property in any of the forms mentioned in the text. The only thing the parties were prepared to argue about was whether the letter was a threat of “serious bodily harm”. These then were the words in issue, the text-to-be-interpreted.

If you pay close attention, you’ll notice that sometimes identifying the text-to-be-interpreted involves choosing between alternative texts which favour different outcomes. When this happens, the outcome of the dispute cannot really be blamed on the meaning of the text, for it actually depends (at least in part) on the initial choice of text.

To illustrate the point, we can look at the analysis of LaForest J. in *R. v. Audet*.² This case concerned a 22 year old man who, shortly after his teaching contract ended, had sexual contact with a 14 year old girl whom he had taught the previous year in eighth grade. At the time of the sexual encounter the accused was aware that his contract would be renewed the

following year and that the girl might become his student again. He was accused of sexual exploitation contrary to s. 153 of the Criminal Code:

153. Every person who is in a position of trust or authority towards a young person ... and who for a sexual purpose touches ... the body of the young person ... is guilty of an indictable offence....

At trial the accused was acquitted on the ground that he did not occupy a position of trust or authority at the time of the sexual contact. Both the trial judge and the Court of Appeal for New Brunswick focused on the words “position of”; however, in the Supreme Court of Canada LaForest J. focused on the word “authority”. To determine the meaning of this text, he looked up the words “autorité” and “authority” in French and English dictionaries and reached the conclusion that the term was broad enough to include any exercise of power, not just those stemming from a formal relationship. He wrote:

In the absence of statutory definitions, the process of interpretation must begin with a consideration of the ordinary meaning of the words used by Parliament.... [T]he meaning of the term [“authority”] must not be restricted to cases in which the relationship of authority stems from a role of the accused but must extend to any relationship in which the accused actually exercises such a power. As can be seen from these definitions, the ordinary meaning of the word “authority” or “autorité” does not permit so restrictive an interpretation.3

Although La Forest J. implies that the fate of the accused was sealed by the meaning of the word chosen by Parliament, in fact he had a choice. He could have focussed on the words “in a position of”. He could have pointed out that Parliament put these words in for a reason, then pointed to the usual reference of the words, namely a formally defined or officially created status.

---

3  Id. at 192-193.
I am not suggesting that the outcome preferred by a majority of the court was wrong. My point is that the outcome was chosen, not dictated by the plain meaning of “authority”.

If the outcome of a dispute can differ depending on which words receive attention and the judge does not have to explain the basis for her choice, or even acknowledge that she’s made one, the plain meaning rule’s promise of certainty is rather empty in my view. This does not mean that outcomes are the result of whim or accident or that texts are meaningless and do not constrain at all. But it does suggest that there may be more to reading a text than proponents of the plain meaning rule are prepared to acknowledge.

B. PMR Trick # 2 — Elastic Co-Text

“Co-text” is a term from linguistics that refers to the text immediately surrounding the text-to-be-interpreted. More precisely, it is the amount of surrounding text that a reader needs to hold in his or her short term memory in order to make sense of the words. If you come across the word “shower” in isolation, you can’t know what kind of shower is referred to. You can’t even know whether the word is a verb or a noun. You need some co-text to form a definite impression of its meaning. In reading the following statements, notice how you use the co-text around “shower” to draw inferences about its meaning, and notice the different sorts of knowledge you rely on in drawing those inferences.

(1) a shower  Since there is an article in front of the word, it must be a noun.

(2) He needed a shower.  It’s probably not a meteor shower since this type of shower is normally not “needed”.

(3) The farmer needed a shower.  It’s probably not a wedding or baby shower since these types of showers are normally associated with brides or new parents, not farmers.

(4) The farmer needed a shower after all that planting.  It’s either a rain shower (needed to encourage the seeds to germinate) or the kind you take indoors (needed to wash away dirt and sweat).  But which?
(5) The farmer needed a hot shower after all that planting. He was cold and tired. Aha! It’s the kind you take indoors. Outdoor showers are usually not “hot” and usually not relied on to warm a person up or soothe tired muscles. Conversely, indoor showers often are hot and used for these purposes.

For me, the meaning of “shower” was uncertain until I got to statement (5). But at that point I became pretty certain that “shower” means the kind taken indoors in a bathtub or a shower stall. My certainty was the result of having eliminated the other possibilities as implausible given what I know about language, farmers, planting, and who needs or receives different kinds of showers under different circumstances. Each word in the co-text triggered an additional set of associations, assumptions, experience and the like.

For purposes of the plain meaning rule, the co-text is the amount of surrounding text that a reader takes into account in deciding (1) the first impression meaning of the text-to-be-interpreted and (2) whether it is plain. If you pay close attention, you’ll notice that in judgments that rely on the plain meaning rule, the co-text is marvellously elastic. Sometimes it includes a single section. Sometimes it expands to take in the entire statute. But more often it shrinks to nothing at all, when the court insists that the text-to-be-interpreted must be looked at in isolation.

The McCraw case offers an example of a text with no co-text. The issue in the case, you’ll recall, was whether a threat of rape was an offence under s. 264.1. of the Criminal Code. The text-to-be-interpreted was “serious bodily harm”. Cory J. determined the meaning of this text by adopting the definition of “bodily harm” in the Code and looking up “serious” in the dictionary. He then refused to look at anything else – including the other words of the section. Counsel for the accused argued that because “serious bodily harm” was coupled with “death” in the relevant provision, the sort of harm the legislature had in mind must be serious indeed. Cory J. refused even to entertain this argument. He wrote:

The appellant urged that serious bodily harm is ejusdem generis with death. I cannot accept that contention. The principle of ejusdem generis has no application to this case. It is well settled that words contained in a statute
are to be given their ordinary meaning. Other principles of statutory interpretation only come into play where the words sought to be defined are ambiguous. The words "serious bodily harm" are not in any way ambiguous.\footnote{R. v. McCraw, supra, note 1, at 80.}

In this passage Cory J. implies that to determine the plain meaning of “serious bodily harm”, the text must be read in isolation, ignoring the rest of the sentence in which it appears or at least not letting the rest of the sentence affect the reader’s understanding.

In \textit{R. v. McIntosh} Lamer C.J. took the same approach as Cory J. The text to be interpreted consisted of the introductory words of s. 34(2) of the \textit{Criminal Code}. Lamer C.J. conceded that if you look at that text in light of ss. 34 and 35 together, it is contradictory and confusing. But if you look at it “in isolation” its meaning is plain.\footnote{See \textit{R. v. McIntosh}, [1995] 1 S.C.R. 686, 697-698.} However, contrast this with what Lamer C.J. said in \textit{Ontario v. C. P. Ltd.}, a case decided the very same year!:

\begin{quote}
\textit{the first task of a court construing a statutory provision is to consider the meaning of its words in the context of the statute as a whole. If the meaning of the words when they are considered in this context is clear, there is no need for further interpretation.}\footnote{[1995] 2 S.C.R. 1028, at 1050. My emphasis.}
\end{quote}

[...]

Although the word “use” is somewhat ambiguous when considered on its own, the expression “for any use that can be made of [the natural environment]” has, in my view, an identifiable literal or “plain” meaning when viewed in the context of the E.P.A. as a whole, particularly the other subsections of s. 13(1).\footnote{\textit{Id.}, at 1054. My emphasis.}
Does the size of the co-text matter? Is it really worth bothering about a technicality of this sort? In my view, the answer is yes. It is through the subtle manipulation of such technicalities that the courts manage to do one thing while appearing to do another. The size of the co-text often affects its meaning, and it is even more likely to affect judgements about ambiguity. Let’s return to the provision considered in McCraw:

264.1 Everyone commits an offence who, in any manner, knowingly utters ... a threat

(a) to cause death or serious bodily harm to any person;

(b) to burn, destroy or damage real or personal property; or

(c) to kill, poison or injure an animal or bird that is the property of any person.

Suppose the question is whether a threat to rape is a threat to cause serious bodily harm and you are allowed to look at the words “serious bodily harm” but no others. How will you answer? Can you answer with any degree of assurance? Now suppose you are allowed to take in a certain amount of co-text, consisting of rest of paragraph (a) – a threat to cause death or serious bodily harm. Does that affect your answer? Would you feel more or less inclined to convict? Now let’s expand the co-text a bit more, to include paragraphs (b) and (c)? Does this make any difference? In my view, it does. In my view, any heightened sense of seriousness created by the reference to “death” in para. (a) is thoroughly flattened by the reference to such relatively unserious matters as damaging property and injuring pets in the remainder of the provision.

In the C.P. case, Lamer C.J. concedes that co-text matters. He says that when he considers the word “use” in isolation it’s ambiguous, but when he considers it in the context of the statute as a whole, it’s not. Given the way co-text works, this is hardly surprising. The more co-text you have, the easier it is to narrow down the possible meanings of a text. The thing to
notice here is the way the courts can affect the meaning of a
text, or affect how “plain” it appears to be, by shrinking or
expanding the co-text. The plain meaning rule cannot provide
certainty if the co-text is manipulable in this way.

C. PMR Trick # 3 — The Shifting Meaning Game

Anyone who reads statutory interpretation cases soon
notices that the courts have a dozen and more different ways of
referring to meaning. There’s ordinary meaning, literal
meaning, common sense meaning, ordinary and grammatical
sense, natural sense, and so on. These terms have no fixed or
precise reference. Sometimes they are used as synonyms for
“plain meaning” but it is also clear that different judges mean
different things by them. This ever-shifting terminology is a
shell game. When I say plain meaning, I might mean any of the
following:

(1) “dictionary” meaning. Dictionary meaning is a-
contextual word meaning.

(2) “literal” or “facial” meaning. Literal meaning is a-
contextual sentence meaning, i.e., meaning that can be
derived from the words alone and their arrangement in
a sentence having regard to the rules of language alone
and disregarding extra-textual factors like purpose or
consequences.

(3) “intended” meaning (i.e., the meaning intended by
the legislature). Linguists call this utterance meaning. It
differs from literal meaning in that it depends on
context, which is defined for this purpose as including
all the knowledge that is stored in the minds of speakers
and audiences. Drafters rely on their knowledge (of
everything — not just language and not just law, but
geography, economics, social convention, cause and
effect — everything ) to predict how the readers of
legislation will interpret their words. These readers rely
on their own encyclopedic knowledge to reconstruct
what the drafter (and legislature) must have intended
when she wrote (and it enacted) those words.
(4) “audience-based” meaning. This is the meaning that is in fact understood by the audience for which the legislation was written – whether the public at large or a particular portion of the public. When the audience is taken to be the public at large, audience-based meaning is usually called “ordinary” meaning. When the audience is a specialized sub-group, audience-based meaning is so-called “technical” meaning. When the audience is the specialized sub-group consisting of lawyers and judges, audience-based meaning is “legal” meaning.

(5) “applied” meaning. Applied meaning is the meaning of a text in relation to particular facts. If your job is to solve an interpretation dispute, then strictly speaking, this is the only meaning that matters.

When judges use the expression “plain meaning”, what kind of meaning are they talking about? Dictionary meaning? Literal meaning? Applied meaning? If you pay close attention, you’ll discover that there’s no consistency in their usage. “Plain meaning” can refer to any of the types of meaning defined above!— or to none in particular. Often the “plain” or “ordinary” meaning of a text is equated with the dictionary definition of its words, or at least the portion of the definition that strikes the court as appropriate. (Given that dictionaries set out the entire range of meaning that words are capable of bearing, it is ironic that dictionaries are the tool of choice when a court wants to put responsibility for an outcome on the plainness of the text.) The court’s dictionary research may or may not be supplemented by textual or contextual or purposive analysis. It may or may not be supplemented by speculation about the legislation’s intended audience and how it might understand the words.

In practice the courts have no fixed understanding of what they mean by “meaning” and no standard approach to establishing it in particular contexts. They are free to use the dictionary or ignore it; to consider the audience for which the legislation was written or ignore it; to appeal to context or purpose or stick to the literal meaning, all as they see fit. These options create choice. But because the same vague terminology is indiscriminately applied to all the different possibilities, the choices being made are not apparent.
Along with artful text selection and elastic co-text, the vagueness of the judicial concept of meaning ensures that courts have significant discretion when it comes to determining first impression meaning but they are not obliged to acknowledge this discretion and indicate the factors governing its exercise.

D. PMR Trick # 4 — It Must Be Plain To You If It’s Plain To Me

The distinction between plain and ambiguous, which lies at the heart of the plain meaning rule, requires judges to determine whether a text can plausibly bear more than one meaning. The judicial notion of plausibility is as elusive as the judicial concept of meaning. The courts make no attempt to give content to the notion. It appears that plausibility is a purely linguistic judgement. Or is there a legal dimension here as well? What factors are relevant to judgements about plausibility? Does perspective matter? Are there degrees of plausibility? If so, how are the degrees measured and classified? These questions are never addressed. The courts simply pronounce on the matter: this proposed meaning is plausible; that one is not.

So what happens when Judge A says “the meaning is this and this is the only plausible meaning” and Judge B says “pardon me, but you’re mistaken, the meaning is plainly that.” If we assume (as we must) that both judges are competent speakers of language and that neither is operating in bad faith, then it is impossible to say purely as a matter of language, that only one of them is right. It would seem to follow that in such a case the text cannot be plain, it must be ambiguous – in which case neither of them is right. In practice, however, proponents of the plain meaning rule generally ignore such empirical evidence of ambiguity. They prefer to rely on their personal linguistic intuitions.

A striking example of this disregard for empirical evidence can be found in the judgment of Lamer C.J. in Ontario v. C.P. Ltd. In that case both Gonthier J. for the majority and Lamer C.J. for the dissent defined their task as determining the meaning of the words “any use that can be made of it” in the following provision of Ontario’s Environmental Protection Act:
13.(1) [...] no person shall deposit [...] a contaminant [...] into the natural environment [...] (a) that causes or is likely to cause impairment of the quality of the natural environment for any use that can be made of it [...] 

Both considered the meaning of this text in light of the same co-text, the subsection as a whole and both carried out a bit of textual analysis. Lamer C.J. concluded!:

When these factors are taken into account, it can, I believe, be concluded that the literal meaning of the expression “for any use that can be made of [the natural environment]” is “any use that can conceivably be made of the natural environment by any person or other living creature”.

After reading the same text in the same co-text, Gonthier J. concluded!

The choice of terms in s. 13(1) leads me to conclude that polluting conduct is only prohibited if it has the potential to impair a use of the natural environment in a manner which is more than trivial.

In other words, Gonthier J. thought that the meaning of the expression “for any use that can be made of [the natural environment]” is “for any non-trivial use [...]”.

Unlike Gonthier J., Lamer C.J. concluded that the meaning he identified was the only meaning the text could plausibly bear. Notice what this says about Gonthier J.’s analysis. It says that the meaning identified by Gonthier J. (and accepted by a majority of the Court) is implausible; it implies that Gonthier J. and the judges who concurred with him lacked linguistic competence. One must presume that Lamer C.J. did not mean to imply this, that he would readily acknowledge in his colleagues a linguistic competence comparable to his own.

---

9 Supra, note 6, at 1055.
10 Id., at 1081.
So what then does he mean? By declaring his preferred meaning plain and invoking the plain meaning rule, Lamer C.J. purports to settle the case on linguistic grounds. The preferred meaning of Gonthier J. is ostensibly rejected on linguistic grounds. Yet if challenged on linguistic grounds, Lamer C.J. would surely back away.

This purported reliance on linguistic plausibility is perhaps the most objectionable aspect of the plain meaning rule. By declaring the meaning plain on the basis of her own intuition, even though that intuition is not shared by others, a judge is able to dismiss the arguments that work against her without ever having to deal with them. Competing interpretations are summarily dismissed as being linguistically implausible; and competing arguments need not be considered because plain meaning governs, even in the face of evidence that the legislature intended something else.

E. PMR Trick # 5 — The Inherent Meaning Illusion

The plain meaning rule is based on the fundamental proposition that once a text is written, it “contains” a meaning that does not change, but remains stable regardless of the context in which the text is read and regardless of the different knowledge and expectations readers bring to the text. Recent work in linguistics has discredited this view.\(^{11}\) Texts do not “contain” meanings; rather they invite readers to infer meanings from the words chosen, the arrangement of the words, the circumstance of utterance and other aspects of context. The text of legislation may be fixed once and for all, but the context cannot be.

Context consists of the linguistic and social conventions which the reader has internalized and the broad range of cultural assumptions absorbed through family, school, religion, work and leisure activities like reading and watching television. It includes everything the reader knows or thinks she knows, everything that is “stored” in her mind at the time of reading.

---

\(^{11}\) For an introduction to the massive body of relevant literature, see G. GREEN, *Pragmatics and Natural Language Understanding*, 2nd ed., Mahwah, New Jersey Lawrence Erlbaum Ass., 1996, p.10.
While much context is shared with other readers, context also varies from one time or place to the next, and one reader to the next. These differences in context explain how different readers can look at the same text and infer different meanings. As the linguists Yule and Brown explain:

"However objective the notion of “text” may appear [...] the perception and interpretation of each text is essentially subjective. Different individuals pay attention to different aspects of texts. The content of the text appeals to them or fits into their experience differently. In discussing texts we idealise away from this variability of the experiencing of the text and assume what Schutz has called “the reciprocity of perspective”, whereby we take it for granted that readers of a text [...] share the same experience (Schutz, 1953)."

If the linguists are right (and judges are not really in a position to challenge them), the assumption that texts contain a fixed meaning actually is false; it is merely a convenient fiction. If the meaning of a text is not in the text, but must be constructed through inference that depends in part on open-ended contexts, the distinction between reading and interpretation cannot be sustained.

This is not good news for proponents of the plain meaning rule, for the rule’s promise of certainty is rooted in and depends entirely on the fixed meaning assumption. If we are to have certainty, the meaning of a text must be the same for everyone who reads it. If meaning is not fixed by the text, but can vary depending on context, then it does not have to be the same for everyone. The certainty promised by the plain meaning rule turns out to be an illusion.

Notice that getting rid of the notion of a fixed inherent meaning does not destroy meaning or preclude communication. It is still possible for readers to experience particular meanings as clear or plain. And if these readers are working within similar

---

contexts, they will probably experience the same particular meanings as being plain. This is what happens in successful communication. Getting rid of the notion of a fixed inherent meaning does not preclude successful communication, but it should require interpreters to acknowledge the complexity and uncertainty that are unavoidable in attempts to communicate.

F. PMR Trick # 6 — Abandoning Ship

The final trick relied on by proponents of the plain meaning rule is perhaps the most disquieting. If it looks like the plain meaning of a text is going to take a court where it definitely does not want to go (despite the possibilities offered by artful text selection, adjusting the co-text or switching the meaning of meaning), all is not lost!— the court can always abandon the rule. This is usually done quietly, without drawing attention to the strategic withdrawal, but on occasion it is done with bravura. Here are some examples from judgments written by proponents of the plain meaning rule on the Supreme Court of Canada:

In *R. v. McIntosh* Lamer C.J. refused to abandon the stark, literal meaning of s. 34(2) of the *Criminal Code*. He wrote!

> where, by the use of clear and unequivocal language capable of only one meaning, anything is enacted by the legislature, it must be enforced however harsh or absurd or contrary to common sense the result may be [...]. The fact that a provision gives rise to absurd results is not, in my opinion, sufficient to declare it ambiguous and then embark upon a broad-ranging interpretive analysis.13

In *R. v. Michaud* he relied on purposive analysis to fill a small gap in s. 187 of the *Criminal Code* which expressly prohibited the opening of sealed evidence packets. In that case he wrote!

13 *Supra*, note 5, at 704.
a stark, literal reading of the provision would appear to suggest that the court must rule on such a motion while turning a blind eye to the contents of the packet [...]. In my view, the provision should be interpreted as permitting a judge to examine the contents of the packet in private for the restricted purpose of adjudicating a s. 187(1)(a)(ii) application. The confidentiality interests underlying the provision are simply not triggered when a competent judicial authority examines the contents of the packet in camera.\textsuperscript{14}

The following passage is from the judgment of Iacobucci J., speaking for the Court in Canada v. Antosko. In this passage he asserts the primacy of the text:

\textit{While it is true that the courts must view discrete sections of the Income Tax Act in light of the other provisions of the Act and of the purpose of the legislation, and that they must analyze a given transaction in the context of economic and commercial reality, such techniques cannot alter the result where the words of the statute are clear and plain and where the legal and practical effect of the transaction is undisputed [...].}\textsuperscript{15}

In \textit{Rizzo v. Rizzo}, however, Iacobucci J. offers a different analysis. In the following passage, the governing factor is not the text but the intention of the legislature:

\textit{Consistent with the findings of the Court of Appeal, the plain meaning of the words of the provisions here in question appears to restrict the obligation to pay [...] However, with respect this analysis is incomplete.}

[...]

\textit{Although the Court of Appeal looked to the plain meaning of the specific provisions in question in the present case, with respect, I believe that the court did}


\textsuperscript{15} [1994] 2 S.C.R. 312, at 327. My emphasis.
not pay sufficient attention to the scheme of the [Act],

its object or the intention of the legislature; nor was the context of the words in issue appropriately regarded.16

In Canada v. Friesen, and indeed in most cases, Major J. is a staunch supporter of the plain meaning rule. In Friesen he writes:

the clear language of the Income Tax Act takes precedence over a court’s view of the object and purpose of a provision [...].

Therefore, the object and purpose of a provision need only be resorted to when the statutory language admits of some doubt or ambiguity.17

But when the apparent meaning of a text produces results that are truly unacceptable, even he abandons the rule. This is what says in Canada v. Schwartz:

Section 3(a) ostensibly permits taxation of income from any source. The argument of the Minister, which is supported by the literal wording of the section, is that “office, employment, business and property” are only examples of sources which may be taxed [...].

However, a literal adoption of this position would arguably constitute a dramatic departure from established tax jurisprudence [...]. Despite the inclusive language of ss. 3(a) and 56, [...]. Canadian courts have always recognized that monies which do not fall within the specifically enumerated sources are not subject to tax.18

These are not isolated examples. They are the tip of a mighty iceberg. Based on my reading of the relevant case law, I think it is fair to say that the plain meaning rule applies if the text seems plain – unless the judge would rather not apply it, in which case it doesn’t apply. This is not my idea of certainty. Nor is it my idea of rule of law.

Conclusion

At first blush, the claim that some texts have a plain meaning seems straightforward, even obvious. But when you look more closely at what proponents of the plain meaning rule do, you discover things like arbitrary text selection, elastic definition of the co-text, shifting and uncertain conceptions of meaning, highly subjective judgements about plausibility, reliance on a discredited view of language, and a lack of commitment to the rule itself. In my view, these practices undermine the credibility of the plain meaning rule and in particular the claim that if the meaning is clear, the outcome is determined by the fixed meaning of the text. There is no such thing as fixed meaning, and even if there were the courts are clearly not bound by it.

IX. How To Be Faithful to the Legislature Without Giving Up Your Freedom

A. FLI Trick # 1 — Presume That Nothing Worth Mentioning Has Changed

If legislation is regarded as an ordinary speech act (i.e., a purposeful attempt to communicate a particular meaning), it follows that legislative meaning is fixed once and for all at the time of enactment and that courts, if they want to be faithful to the intentions of the legislature, must seek to discover that original meaning. However, given the role of context in inferring intended meaning, this is not necessarily an easy thing to do. Some statutes were enacted more than a century ago; some were enacted in other jurisdictions; some have been amended many times. The context in which a statute was enacted may bear little resemblance to the context in which it now operates or in which it is to be interpreted. And in truth, the
courts lack the resources required to reconstruct and imaginatively inhabit contexts other than their own. The original meaning rule may be rhetorically necessary, but it is often impossible to apply.

This problem is avoided in practice by presuming that the context in which the legislation was enacted is not significantly different from either the context in which it now operates or the context in which it is being interpreted. The burden of proving some significant divergence in these contexts, one that would affect meaning, falls on the person alleging it. In the absence of such an allegation, current meaning is presumed to be the same as original meaning.

This approach works well in practice for it allows the courts to carry out the essential work of adapting ageing statutes to changing circumstance without having to give up the advantages of a fixed legislative intent. The only drawback to the approach is that, like the doctrine of presumed intent examined below, presumed absence of change is not necessarily true. In cases where these presumptions are false, the real work that courts do in interpreting legislative texts is hidden by the fiction.

B. FLI Trick # 2 — More Meaning Games

On occasion the original meaning rule is expressly invoked and relied on to justify the outcome in a statutory interpretation case. The courts take the rule seriously, but the concept of meaning relied on for purposes of the rule turns out to be as equivocal in this context as it is in the context of the plain meaning rule. Sometimes “original meaning” refers to the original connotation of the words used by the legislature – their abstract sense or their dictionary definition. At other times “original meaning” refers to their original denotation – the facts to which the words would have been applied when the legislation was first enacted. Obviously the connotation of words is much less dependent on context, and therefore less affected by change, than their denotation. By shifting between these two understandings of “original meaning”, the courts are able to ensure appropriate outcomes without appearing to abandon their commitment to legislative intent.
This trick is nicely illustrated by the contrasting judgment of Lamer J. (as he then was) with the judgment of L’Heureux-Dubé J. in Hills v. Canada.\footnote{[1988] 1 S.C.R. 513.} The issue in the case was the meaning of “financing” in s. 44(2) of the Unemployment Insurance Act (as it then was). Section 44(1) disqualified claimants from receiving insurance benefits if their loss of employment was due to a labour dispute at their place of work. However, s.44(2) created an exception for claimants who could prove that they were not “participating in or financing or directly interested in” the dispute that caused the work stoppage. The court had to decide whether a claimant could be said to be “financing” a dispute simply because a portion of his regular union dues was paid into a central fund from which strike pay was subsequently paid to workers striking at his place of work.

L’Heureux-Dubé J. relied on the original denotation of “financing” in concluding that the indirect and involuntary form of payment at issue here was not a form of “financing” within the meaning of the section. As she explained:\footnote{\textit{Id.}, 554-555.}

\begin{quote}
The original "financing" provision, enacted in 1935 and re-enacted in 1940, was drafted at a time when very different social conditions prevailed, particularly in the area of labour relations [...]. In 1935 and until 1944, labour unions were purely voluntary organizations. Individuals would join unions on a voluntary basis and would make their financial contributions in the same manner. They were therefore presumed to be intentionally financing the union’s activities within the meaning of the disentitlement provision.\end{quote}

She therefore concluded that the original meaning of “financing” was “giving direct and voluntary financial support” of the sort that would have been contemplated by the legislature when the provision was first enacted in 1935.

Lamer J. relied on the connotation of “financing” which he discovered by consulting the dictionary:
"Financing" means [TRANSLATION] "obtaining the capital necessary to operate" (Petit Robert 1 (1986)),

[TRANSLATION] "paying, providing money" (Grand Larousse de la langue française (1973)).

As Lamer J. pointed out, understood in this way the term is broad enough to cover any contribution, voluntary or involuntary, direct or indirect, from one person to another!: “it is impossible to find in the word ‘financing’ used by itself a requirement of active and personal participation or a direct link between the claimant’s contribution and the immediate labour dispute”. Lamer J. completed his analysis by connecting this broad definition of “financing” to the legislature’s original intent:

As L’Heureux Dubé J. has indicated, the wording of s. 44(2)(a) has received little or no alteration since the Unemployment Insurance Act was adopted in 1940 [...]. Though aware of the changes that have occurred in labour relations, Parliament has not felt it necessary to limit the application of a word of general import undoubtedly because it intended to cover all situations to which the word might apply. We cannot assume that this is a mere oversight.

For Lamer J. fidelity to legislative intent means that the court must continue to apply the original connotation of “financing” to whatever facts meet the defining criteria (regardless of consequences) unless and until the legislature signals some other intent.

C. FLI Trick # 3 — Distinguish Sloppy Drafting From Legislative Error

Fidelity to legislative intent requires that courts correct any errors in a legislative text that are owing to the inattention or lack of skill of the drafter. Such errors must be corrected to bring the text in line with the meaning actually intended by the

---

21 Id.
22 Id.
THE PLAIN MEANING RULE

legislature. However, if the text contains errors or anomalies for which the legislature itself is responsible, fidelity to legislative intent requires that courts do nothing. To cure a legislative oversight would be a form of amendment and an impermissible encroachment on the legislative sphere.

In principle this distinction is coherent enough, and in practice there are some errors for which there is real evidence of a drafter’s error. But there is also a large category of errors for which it is impossible to assign responsibility. Errors in this category are up for grabs!: depending on its preferred outcome, the court can notionally rewrite the text or piously refuse to touch it!— both in the name of fidelity to legislative intent. These competing possibilities are nicely illustrated in the majority and dissenting judgments of the court in R. v. McIntosh.

The McIntosh case was concerned with the self-defence provisions of the Criminal Code, which read as follows!:

*Self-defence against unprovoked assault*

**34.(1)** Every one who is unlawfully assaulted without having provoked the assault is justified in repelling force by force if [...] [no requirement to retreat].

*Extent of justification*

(2) **Every one who is unlawfully assaulted** and who causes death or grievous bodily harm in repelling the assault is justified if [...] [no requirement to retreat].

*Self-defence in case of aggression*

**35.** Every one who [...] has without justification provoked an assault on himself by another, may justify the use of force subsequent to the assault if [...].

(c) he declined further conflict [...] or retreated from it as far as it was feasible to do [...].

The issue was whether an accused who provoked an assault on himself could rely on s.34(2). The Crown argued that
the less onerous conditions for self-defence set out in subsections 34(1) and (2) were available only to persons who defended themselves from assaults that they did not provoke. In cases where the accused provoked the assault, the more onerous conditions set out in s.35 had to be met. The Crown explained that s.34(2) did not expressly refer to this non-provocation requirement because the reference was inadvertently omitted when the Criminal Code was overhauled in 1955. In short, the absence of the words “not having provoked the assault” in s.34(2) was due to a drafter’s error.

McLachlin J. (dissenting) accepted this argument. In her view, the legislative evolution of ss. 34 and 35 offered persuasive evidence that the drafter who prepared the 1953-54 version of the Code accidentally changed the meaning of the provisions when all he sought to do was improve the style. She pointed out that the former wording of s.34(2) in the 1892 and 1906 versions of the Code made it clear that non-provocation was a requirement of the offence:

(1) Criminal Code, S.C. 1892

Self-defence against unprovoked assault

45. Every one unlawfully assaulted, not having provoked such assault, is justified in repelling force by force, if [...] [same as now]; and every one so assaulted is justified, though he causes death or grievous bodily harm, if [...] [same as now] he causes it under reasonable apprehension of death or grievous bodily harm from the violence with which the assault was originally made or with which the assailant pursues his purpose, and if he believes, on reasonable grounds, that he cannot otherwise preserve himself from death or grievous bodily harm.

(2) Criminal Code, R.S.C. 1906

Self-defence against unprovoked assault

53.(1) Every one unlawfully assaulted, not having provoked such assault, is justified in repelling force by force, if [...] [same as now]
Extent justified

(2) Every one so assaulted is justified, though he causes death or grievous bodily harm, if [...] [same as now]

However, in the 1953-54 Criminal Code, the words “everyone so assaulted” became “everyone who is unlawfully assaulted”:

(3) Criminal Code, R.S.C. 1953-54

Self-defence against unprovoked assault

34.(1) Every one who is unlawfully assaulted, not having provoked such assault, is justified in repelling force by force, if [...] [same as now]

Extent justified

(2) Every one who is unlawfully assaulted and who causes death or grievous bodily harm, is justified if [...] [same as now]

It was obvious to McLachlin J. that the drafter incorrectly replaced “so assaulted” with “unlawfully assaulted”; he should have written “unlawfully assaulted, without having provoked the assault”.23

Lamer C.J., looking at the same evidence, drew a different inference. He thought the 1892 and 1906 versions were both ambiguous, but this ambiguity was deliberately resolved in the 1953-54 version. He wrote:

There is a clear ambiguity in this provision. Does the expression "every one so assaulted" refer to "[e]very one unlawfully assaulted", or to "[e]very one unlawfully assaulted, not having provoked such assault"? This question is academic, since Parliament appears to have

23 Supra, note 5.
resolved the ambiguity in its 1955 revision of the Criminal Code, S.C. 1953-54, c. 51.²⁴

Unlike McLachlin J., who blamed the drafter, Lamer C.J. attributed the change in the 1955 revision to Parliament. Since this was an intentional change, it was his duty to give effect to it even though it admittedly was anomalous and led to absurdity:

Even though I agree with the Crown that the interpretation of s. 34(2) which makes it available to initial aggressors may be somewhat illogical in light of s. 35, and may lead to some absurdity, I do not believe that such considerations should lead this Court to narrow a statutory defence. Parliament, after all, has the right to legislate illogically (assuming that this does not raise constitutional concerns). And if Parliament is not satisfied with the judicial application of its illogical enactments, then Parliament may amend them accordingly.²⁵

Once again, fidelity to the fixed intention of Parliament proves to be wonderfully malleable. It justifies McLachlin J. in effectively rewriting the provision and it equally justifies Lamer C.J. in refusing to make any change. Like the plain meaning rule, it constrains the rhetoric of the court but not its practice.

D. FLI Trick # 4 — Presume That the Legislature Wants What You Want

Intentionalism’s most dazzling trick is the doctrine of presumed intent. To appreciate how effectively this trick has worked, in English Canada at least, all we need do is look at the success of Elmer Driedger’s so-called “modern principle” which appeared in the first and second editions of his book Construction of Statutes.²⁶ Anyone who has read these early editions of Driedger will be struck by their unremitting

---

²⁴ Id.
²⁵ Id.
positivism and commitment to the doctrine of fidelity to legislative intent. In Dreidger’s garden, judges do not make law; they do not even interpret it; they “construct” it! — a term which suggested to Dreidger a workmanlike subordination to legislative plan. His approach is summarized in the modern principle:

_Today there is only one principle or approach, namely, the words of an Act are to be read in their entire context and in their grammatical and ordinary sense harmoniously with the scheme of the Act, the object of the Act, and the intention of Parliament._

Despite my effort to give more meaningful content to this principle in the 3\textsuperscript{rd} edition of Driedger, the Supreme Court of Canada continues to cite and rely on the principle from the 2\textsuperscript{nd} edition. I think it is important to notice what Driedger meant by the intention of Parliament. In the 2\textsuperscript{nd} edition he wrote:

_It may be convenient to regard “intention of Parliament” as composed of four elements, namely:

1. The expressed intention — the intention expressed by the enacted words;

2. The implied intention — the intention that may legitimately be implied from the enacted words;

3. The presumed intention — the intention that the courts will in the absence of an indication to the contrary impute to Parliament; and

4. The declared intention — the intention that Parliament itself has said may be or must be or must not be imputed to it._

---

27 _Id., at p.187._
28 _Id., at p.1106._
By virtue of this definition, the intention of Parliament is deemed to include everything the courts care to impute to Parliament, so long as it does not contradict what Parliament actually said. This is truly a “convenient” way to regard legislative intent.

As defined here, the notion of presumed intention covers not only the standard presumptions (for example, presume that the legislature does not intend to interfere with private property rights) but also evolving presumptions (for example, presume that the legislature does not intend to make women poorer). It also covers the broader and vaguer presumptions (for example, presume that the legislature is reasonable and that it wants to avoid absurdity and injustice). In fact, Driedger’s notion of presumed intent has no real limit. It covers whatever values or policies the court sees fit to attribute to the legislature. With this one little trick, Driedger has made the legislature—our only legitimate law-maker!—intend whatever judges want it to. Unsurprisingly, the courts find it easy to be faithful to this notion of legislative intent.

Why does Driedger, arch-positivist and a founding father of Canadian intentionalism, do this? Why does he let this snake into his intentionalist garden? The answer is that he has no choice. Actual legislative intent is not enough. In a significant number of cases, the statute does not tell the court what to do and there is no evidence of relevant intention in sources like Hansard. When it appears that the legislature never considered the problem facing the court, when there is no relevant evidence of legislative intent, the court has no choice but to make something up. However, any embarrassment that might flow from such unavoidable bouts of judicial law-making is avoided by presuming that the part made up by judges is in fact what the legislature intended all along.

Conclusion

In my view, the notion of legislative intent is meaningful and has an important role in statutory interpretation. It is sometimes possible to draw compelling inferences about the intended purpose or meaning of a legislative provision and when this happens the court must ordinarily follow the direction of the legislature. The problem with legislative intent is that it
only goes so far. It is an important consideration but it is not the only consideration. And it often stops short of providing the court with the answers it needs.

The doctrine of fidelity to legislative intent becomes a problem when the court refuses to acknowledge the limitations of intention as a plausible or desirable constraint on judicial law-making.

The courts are not free to go off on frolics of their own, but they have an active role in statutory interpretation which includes working out the implications of legislation, adapting legislative directives and policies to changing circumstances and ensuring that important legal values are respected!— both those entrenched in the written constitution and those that are part of the evolving common law.

**What Should Be done?**

In my view it is not possible to bridge the gap between judicial practice and judicial rhetoric by making the practice conform to the rhetoric. The plain meaning rule presupposes that particular meanings somehow inhere in texts, but in fact they don’t. The doctrine of fidelity to legislative intent presupposes that the legislature has solved every problem, but obviously it couldn’t and it didn’t. In my view, we should respond to such unyielding realities by acknowledging them and incorporating them into our mythology. We should acknowledge that legislative text and legislative intention are incomplete sources of law and that interpretation is needed to complete the job. We should then focus on the possibilities this creates for developing a better mythology.

Ultimately the gap between rhetoric and reality will not be closed unless we re-imagine some of the rather simplistic notions in our mythology. In particular, the idea that legitimate, democratically-made law is co-extensive with the output of the legislature has to be reconsidered. The elected representatives of the people are controlled by the government and most of the legislation which affects our everyday lives is made by the executive branch. All of it is subject to interpretation by non-
elected officials. To respond to all this, we need a more complex and expansive notion of democracy. We could, for example, think of the resolution of interpretation disputes as a form of law-making in which subjects get to participate in the law-making process by arguing about how particular legislation applies to them. We could try to devise procedures that would decentralize and democratize interpretation.

We must also re-examine our tendency to equate the rule of law with the application of pre-fixed rules by a-political automatons. This is not how it works in practice. Those who interpret and apply the law inevitably bring their own context to the job. Instead of denying the role of context in interpretation, we should try to understand how it works and what its implications are. If we want to avoid bias and partiality in the application of law, for example, it seems obvious to me that we need to appoint a more diverse group of persons to our courts and tribunals. Historically judges have been drawn from a very narrow and elite segment of society. Quite naturally, they have perceived their own assumptions to be unbiased truth and protection of their own interests to be impartial protection of the public interest. In this context, challenges by those with different assumptions or interests are readily dismissed as biased or partial or serving a “special” interest. While courts and tribunals have become somewhat more representative in recent years, significant change cannot occur until the legal profession is not only opened to members of traditionally excluded groups but also opened to the contexts occupied by those groups. This will not happen overnight.

In the meanwhile we must repeatedly point out to the emperor that illusions are not clothes.
The Questionable Empirical Basis of Article 2’s Incorporation Strategy: A Preliminary Study

Lisa Bernstein*

Introduction.............................................................................................................. 191

I. Trade Associations’ Attempts to Codify Custom.................................................. 200
   A. Hay ............................................................................................................... 202
   B. Grain and Feed............................................................................................ 210
   C. Textile Industry............................................................................................ 217
   D. Silk ............................................................................................................... 223

* Professor of Law, University of Chicago. I would like to thank my research assistants Marc Blitz (silk), Richard Welch (hay), Drew Porter (coffee), Andrea Adelman and Deborah Healy (grain and feed), Mary LaBrec (textiles), as well as Connie Fleischer, Bill Schwesig, Chuck Ten-Brink, David Barrett, and Lorrie Ragland. I would also like to thank Yoram Barzel, Omri Ben-Shahar, Lucian Bebchuk, Edward Bernstein, Brian Bix, David Charny, Dick Craswell, Richard Epstein, Clay Gillette, Victor Goldberg, Jack Goldsmith, Philip Hamburger, Alon Harel, Claire Hill, Ehud Kamar, Alan Kamp, Louis Kaplow, Avery Katz, Dan Klerman, Doug Lichtman, Ronald Mann, Ed Morrison, Dennis Patterson, Eric Posner, Richard Ross, Keith Sharfman, Steve Shavell, Alex Stein, Cass Sunstein, Bill Vukovitch, Jim Whitman, Eyal Zamir and participants at the Northwestern University Law School Symposium on Relational Contracting, Harvard Law and Economics Workshop, University of Montreal Law School Workshop, Columbia Law School Law and Economics Workshop, Berkeley Law and Economics Workshop, University of Virginia Faculty Workshop, American Law and Economics Association Annual Meeting (Contract Law Panel), Law and Society Association Annual Meeting, and The University of Chicago Law Review Symposium Formalism Revisited for helpful comments or discussions. This research is part of a larger project on private commercial law that is sponsored by NSF Grant SBR-9422795. Additional funding was provided by the University of Chicago Law School.
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. Methodological Problems</td>
<td>229</td>
</tr>
<tr>
<td>F. Conclusion</td>
<td>236</td>
</tr>
<tr>
<td>II. The Incorporation Strategy Revisited</td>
<td>237</td>
</tr>
<tr>
<td>III. The Relationship-Creating Role of Custom</td>
<td>254</td>
</tr>
<tr>
<td>Conclusion</td>
<td>274</td>
</tr>
</tbody>
</table>
The Uniform Commercial Code ("Code").1 the Convention on Contracts for the International Sale of Goods,2

---

1 This essay focuses solely on U.C.C. ("Code") Article 2 as supplemented by the relevant parts of Article 1. U.C.C. §§ 1, 2 (West 14th ed 1995). The incorporation principle is expressed in the Code sections dealing with course of dealing and usage of trade, id., at § 1-205, and course of performance, id., at § 2-208, as well as in the Code's definition of "Agreement," which includes "the bargain of the parties in fact as found in their language or by implication from other circumstances including course of dealing or usage of trade or course of performance." Id., at § 1-201(3). It is also at the heart of the Code's duty of good faith, which requires that merchants act in accordance with "the observance of reasonable commercial standards of fair dealing in the trade," id., at § 2-103(b), which is to be "further implemented by Section 1-205 on course of dealing and usage of trade," id., at § 1-203 cmt, as well as the Code's interpretive approach, which directs courts to determine "the meaning of the agreement of the parties [...] by the language used by them and by their action, read and interpreted in light of commercial practices and other surrounding circumstances. The measure and background for interpretation are set by the commercial context, which may explain and supplement even the language of formal or final writing." Id., at § 1-205 cmt 1. See also id., at § 2-301 cmt ("In order to determine what is 'in accordance with the contract' under this Article usage of trade, course of dealing and performance, and the general background of circumstances must be given due consideration."). The incorporation principle is also embodied in numerous Code provisions and Official Comments that direct courts to take into account "usages of trade," "commercial standards," "the law merchant," and other aspects of the contracting context in filling gaps. See generally the obligations in U.C.C. Art.2, Part 3 ("General Obligation and Construction of Contract") and its associated Official Comments. Less noticeably, and typically ignored by commentators, the incorporation strategy, in the form of explicit references to usages of trade, commercial standards, and commercial context, as well as references to what is commercially reasonable, reasonable, reasonable, reasonable, and commercial impracticable, also runs through numerous Code provisions (as explicated by their associated Official Comments) that might broadly be termed "traffic rules" and that unlike gap fillers are, as a practical matter, difficult, if not impossible, to negate. See generally U.C.C. Art. 2, Part 2 ("Form, Formation and Readjustment"); Part 5 ("Performance"); Part 6 ("Breach, Repudiation and Excuse"); Part 7 ("Remedies"). Finally, most generally, one of the stated "underlying purposes" of the Code is "to permit the continued expansion of commercial practices through custom, usage and agreement of the parties." U.C.C. at § 1-102.

2 See, for example, United Nations Convention on Contracts for the International Sale of Goods ("CISG") Art 9(2) (1980), reprinted in United Nations, Conference on Contracts for the International Sale of Goods, 19 Intl Materials 671, 674 (1980) ("The parties are considered, unless otherwise agreed, to have impliedly made applicable to their contract or its formation a usage of trade which the parties knew or ought to have known and which in
and the modern *Lex Mercatoria*,³ are based on the premise that unwritten customs and usages of trade exist and that in commercial disputes they can, and should, be discovered and applied by courts.⁴ The existence of commercial customs that can be discovered and codified by diligent observers is also at the heart of some proposals for creating commercial law in developing or formerly socialist countries.⁵ More broadly, the

³ The term *lex mercatoria* is generally understood to include!: international conventions on the sale of goods, written compilations of custom like the Incoterms (infra, note 137), as well as unwritten customs and practices that are either specific to a particular trade or are applicable to all commercial transactions. See, for example, Harold J. Berman and Felix J. Dassler, “The ‘New’ Law Merchant and the ‘Old!’ Sources, Content, and Legitimacy,” in Thomas E. Carbonneau (Ed.), *Lex Mercatoria and Arbitration: A Discussion of the New Law Merchant*, Transnational Juris, 1990, pp. 21-36 (discussing the customary basis of the *lex mercatoria*); and Keith Highet, “The Enigma of the Lex Mercatoria, in T.E. Carbonneau", op. cit., pp. 104-105.

⁴ This Article’s discussion of the U.C.C. focuses nearly exclusively on disputes between merchants. Early drafts of the Code provided for the use of either arbitrators or merchant juries to determine the content of trade usages and other commercial standards. See, for example, 1941 Revised Uniform Sales Act (Report and Second Draft), at § 59(I)(d) (setting out matters to be submitted to merchant experts, including the effects on obligations of “mercantile usage” or “usage of trade,” whether tender was conforming, whether an action was reasonable, or any other “issue which requires for its competent determination special merchants knowledge rather than general knowledge”). The failure to include a merchant jury provision in the final Code was due to a need to obtain political support for the proposed Code, not to Llewellyn’s abandonment of the concept. See Zipperah Batshaw WiseMAN, “The Limits of Vision!: Karl Llewellyn and the Merchant Rules”, (1987) 100 Harv. L. Rev. 465. Most of the objections to the incorporation strategy put forth in this essay, however, would not be weaker if a merchant jury provision had been included because they are not based primarily on the limits of courts’ institutional competence.

⁵ See, for example, Paul H. Rubin, “Growing a Legal System in the Post-Communist Economies”, (1994) 27 Cornell Int'l. L. J. 1 (suggesting that trade associations should be formed to interpret and enforce contracts according to commercial custom so that, over time, customs will become sufficiently well developed to supply the basic principles of a public commercial law); Anthony T. Kronman, “Contract Law and the State of Nature”, (1985) J. L. Econ. & Org. 5; and Robert D. Cooter, “Decentralized Law for a Complex Economy!: The Structural Approach to Adjudicating the New Law Merchant”, (1996) 144 U. Pa. L. Rev. 1643 (suggesting that as the economy increases in complexity it becomes increasingly important for courts to
idea that courts in deciding cases should look to immanent business norms, consisting of both the practices of contracting parties and unwritten customs, is a fundamental tenet of the legal realist approach to contract interpretation, an approach that was developed, championed, and ultimately codified by Karl Llewellyn, a leading legal realist and the principal drafter of Article 2 of the Uniform Commercial Code (“Code”).

Academic commentators have long debated the proper role of customary practices in commercial adjudication. They have explored the difficulty of defining the parameters of the custom to be sought, have debated the efficiency of custom, have explored the actual role played by custom in various

enforce certain types of business custom). Similarly, commentators have suggested looking to usages of the trade to set standards for electronic commerce. See, for example, Alejandro E. ALMAGUER and Roland W. BAGGOTT III, “Shaping New Legal Frontiers: Dispute Resolution for the Internet”, (1998) 13 Ohio St. J. on Disp. Res.711, 714 and 716 (advocating the use of “Self Regulatory Mechanisms [to] Promote a Cyber-Usage of Trade,” because, “[i]n the age of mature electronic commercial transactions, parties must be able to rely on custom and usage of trade”).

6 See generally William TWINING, Karl Llewellyn and the Realist Movement, chs. 11-12, University of Oklahoma Press, Norman, 1973; see also, Richard DANZIG, “A Comment on the Jurisprudence of the Uniform Commercial Code”, (1974-75) 27 Stanf. L. Rev. 621, 624 (“Llewellyn saw law as an articulation and regularization of unconsciously evolved mores—a crystallization of a generally recognized and almost indisputably right rule (a ‘singing reason’), inherent in, but very possibly obscured by, existing patterns of relationships.”).

7 There is no widely accepted definition of a custom. This essay defines it as an unwritten practice, which would be considered a usage of trade under the Code, see U.C.C. § 1-205, or the type of commercial standard that would be incorporated into a contract or taken into account in the contract interpretation process under the Code. See supra, note 1. This Article uses the terms trade practice, usage of trade, and custom interchangeably.


9 See, for example, Eric A. POSNER, “Law, Economics, and Inefficient Norms”, (1996) 144 U. Pa. L. Rev. 1697 (providing a host of reasons that norms might not be efficient); Jody S. KRAUS, “Legal Design and the Evolution of Commercial Norms”, (1997) 26 J. Legal Stud. 277 (drawing on a theory of cultural evolution to argue that customary practices are unlikely to be efficient but may be more nearly optimal than either individual decisions with respect to particular transactions or legislative enactments).
adjudicatory fora, have discussed the extent to which custom was absorbed into the common law, and have noted the problems of institutional competence that might inhibit the accurate determination of the content of customary practices. However, a more basic and naturally prior question has not been adequately addressed. Namely, to what extent do the types of industry-specific meanings of words and the types of unwritten, industry-wide “usages of trade” and “commercial standards”—that the Code directs courts to incorporate into commercial agreements through both gap-filling and the interpretive process—actually exist as to most aspects of contracting relationships in merchant communities?

---


11 See, for example, E. Karl MCGINNIS, “Present Legal and Practical Methods by which Business Custom is Enforced”, (1927) 5 N.C.L. Rev. 136 (discussing various ways of “enforcing” custom including trade association arbitration and incorporation into the common law).

12 See, for example, Randy E. BARNETT, “The Sounds of Silence: Default Rules and Contractual Consent”, 78 Va. L. Rev. 821, 908 note 231 (noting that while “judges may be good surrogates for the rationally ignorant consumer, they are often deficient interpreters of more specialized usages of trade.”).

13 The Code was assumed to be based on a solid empirical foundation. William A. Schnader, a primary mover behind the Code project, chose Karl Llewellyn as Chief Reporter because

“Not only was Professor Llewellyn a student of commercial law as it appeared in the law books, but he was the type of law professor who was never satisfied unless he knew exactly how commercial transactions were carried on in the market place. He insisted that the provisions of the Code should be drafted from the standpoint of what actually takes place from day to day in the commercial world rather than from the standpoint of what appeared in statutes and decisions.”

William A. SCHNADER, “A Short History of the Preparation and Enactment of the Uniform Commercial Code”, (1967) 22 U. Miami L. Rev. 1, 14 (U.C.C. Symposium). However, with the exception of seeking (and then ignoring) the opinions of merchants in hearings on the Code (see infra, note 144 and accompanying text), rigorous empirical research into what types of rules would actually be responsive to merchant concerns was never undertaken. While Llewellyn’s defenders recognize that the lack of an empirical basis for the Code was inconsistent with his realist and scientific approach to law as well as his often-expressed position that in drafting a commercial code attention should be paid to the “wide basis of established commercial experience,” W. TWINING, loc. cit., note 6, 524, they are quick to point out
Starting from the generally accepted premise that unwritten commercial customs are most likely to arise and endure in situations where transactors interact on a repeat basis, over a long period of time, in relatively similar transactions, this Article addresses the question from a largely empirical

that “critics who have been suspicious of Llewellyn’s alleged ‘unscientific,’ ‘impressionistic,’ or ‘anecdotal’ approach to facts have yet to point to any major factual assumptions of the Code that were misleading or inaccurate. Nor have suggestions been forthcoming as to specific empirical research that might have been worth doing.” Id., at 319. See also the comment of Robert Summers in id., at 467 (“I think the biggest and best reason for lack of empirical research is this: most of the law is ‘suppletive’ law—it applies only when the parties have not agreed as to the matter in hand. It says what the law is when the parties don’t say.”). This Article, together with earlier work on other industries, see Lisa BERNSTEIN, “Merchant Law in a Merchant Court: Rethinking the Code’s Search for Immanent Business Norms”, (1996) 144 U. Pa. L. Rev. 1765 and Lisa BERNSTEIN, Private Commercial Law in the Cotton Industry: Value Creation Through Rules, Norms, and Institutions (Sept. 8, 1998) (unpublished manuscript, on file with U. Chi. L. Rev.), suggests that the lack of solid empirical research led Code drafters to adopt provisions that are detrimental rather than accommodating to merchant concerns, and to a commercial law based on a deeply flawed understanding of merchant reality.

It should be noted that Llewellyn was aware of the sources discussed in this Article. In a memorandum he presented in defense of the proposed Code’s battle of the forms provision, U.C.C. § 2-207, he cited the existence of similar provisions in merchant trade rules in support of his position; yet when these sources did not support the position he was advocating, Llewellyn neglected to mention them. Karl N. LLEWELLYN, Memorandum by K.N. Llewellyn Replying to the Report and Memorandum of Task Group 1 of the Special Committee of the Commerce and Industry Association of New York, Inc., on the Uniform Commercial Code, in Study of Uniform Commercial Code Memoranda presented to the Commission and Stenographic Report of Public Hearing on Article 2 of the Code 42, 56-57 (1954), reprinted in State of New York, 1 Report of the Law Revision Commission for 1954 and Record of Hearings on the Uniform Commercial Code 106, 121 (Williams 1954). In addition, Llewellyn’s early writings show that he was aware of the existence of these private legal systems. See, for example, Karl N. LLEWELLYN, “The Effect of Legal Institutions Upon Economics”, (1925) 15 Am. Econ. Rev. 665, 672 and 673, note 24. (“Increasingly, associations are forming which adopt their own rules of action and even settle their own disputes [...]. And the rules which [...] such associations lay down and apply, are part of the body of our law [...]. [T]he association-made rules are like enough to law to deserve careful attention.”).

These preconditions for the emergence of custom are widely accepted. Several additional factors, such as that the transactors play reciprocal roles in the transaction (in the commercial law setting this means that they are buyers one day and sellers the next), and that most transactions are among members of an
perspective. In keeping with Llewellyn’s view that commercial law should reflect merchant reality, it identifies

ethically homogeneous or geographically concentrated small group, are said to make the emergence of custom more likely but are not strictly required. For endorsements of these criteria, see Richard A. EpSTEIN, “The Path to The T.J. Hooper: The Theory and History of Custom in the Law of Tort”, (1992) 21 J. Legal Stud. 1, 11-16. ("[T]he key variables on the emergence of custom seem to be the symmetry of results and the frequency of the dispute with the question of severity of the loss playing a secondary role."); Bruce L. BENSON, “Customary Law as a Social Contract!: International Commercial Law”, (1992) 3 Const. Pol. Econ. 1, 7 (endorsing these criteria as corresponding to the game theoretic conditions for the emergence of commercial custom and cooperation). See also Robert C. EL-liCKSON, Order Without Law: How Neighbors Settle Disputes, Cambridge, Harvard University Press, 1991. Although the merchant industries discussed in this essay do not perfectly fulfill the conditions theorists identify as being ideal for the emergence of custom, they come far closer to doing so than most contemporary industries whose disputes are adjudicated under the Code’s merchant rules.

Most commentators simply assert or assume the existence of custom. See, for example, H. J. BERMAN and F. J. DRASSLER, loc. cit., note 3, 28 and 32. ("Yet, that should not stop us from seeing what is right in front of our noses! It is the factual existence of international custom and its continuous use [...] that allows us to speak of international trade as a special type of international law. Nobody denies that there is a body of international rules, founded on the commercial understandings and contract practices of an international community principally composed of mercantile, shipping, insurance, and banking enterprises of all countries."); id., at 25 ("[C]ustomary commercial understandings enjoy almost total recognition [...] notwithstanding the fact that [] such rules may not be very conspicuous and may easily be overlooked by scholars."). The only notable exceptions are Jack L. GOLDSMITH and Eric A. POSNER, “A Theory of Customary International Law”, (1999) 66 U. Chi. L. Rev. (drawing on game theoretic models and fact-based case studies to suggest that customary international law does not exist); CRASWELL, loc. cit., note 8 (questioning whether trade customs can be given content by adjudicators independent of "the goals, beliefs and other normative premises of the person doing the identifying," and concluding that they cannot because the problems faced by a court "finding" the content of a custom are analytically similar to the problems faced by a court trying to find the content of the common law"). See also Leon E. TRAKMAN, The Law Merchant: The Evolution of Commercial Law, Littleton, Fred B. Rothman, 1983, pp. 45-60 (attempting to empirically establish the existence of custom in international oil contracts); but see Chris WILLIAMS, Book Review, “The Search for Bases of Decision in Commercial Law!: Llewellyn Redux”, (1984) 97 Harv. L. Rev. 1495, 1501-1504 (arguing that Trakman’s evidence does not establish the existence of custom).

See, for example, Walter D. MALCOLM, “The Proposed Commercial Code A Report on Developments During the Period from May 1950 through
several merchant industries—hay, grain and feed, textiles, and silk—that in an early stage of their development were roughly characterized by conditions favorable to the emergence of customs. It then explores the attempts of national trade associations in these industries to codify their industries’ customs into written trade rules.

The debates surrounding these codification efforts suggest that there was not widespread agreement among merchants as to either the meaning of common terms of trade or the content of many basic commercial practices. Rules committee debates sometimes went on for years, customs relating to important aspects of transactions were left uncodified because consensus could not be achieved, and in most industries drafting committees eventually engaged in only selective codification. In addition, over time, many associations came to explicitly concede that they were attempting to change rather than merely incorporate existing practices.

These findings, together with interview evidence and the testimony of merchant associations on the proposed commercial Code, suggest that “usages of trade” and “commercial standards,” as those terms are used by the Code, may not consistently exist, even in relatively close-knit merchant communities. While merchants in the industries examined here sometimes do and did act in ways amounting to loose behavioral regularities, most such regularities are either much more geographically local in nature or far more general in scope and conditional in form than is commonly assumed.17 These

---

February 1951”, (1951) 6 Bus. Law 113, 126 (quoting the Report of the Committee on the Proposed Commercial Code!: “the practices of businessmen and business houses are important factors in construing their contracts and actions and in determining their rights and liabilities [...] [M]any of the changes effected by the Code are designed to adapt rules of law to the way that business is actually carried on”) (emphasis in original). See also National Conference of Commissioners on Uniform State Laws, Report and Second Draft: The Revised Uniform Sales Act, § 59 at 253 (cited in note 4) (noting, in a comment thought to have been drafted by Llewellyn, that while “the law about the effect of ‘business custom’ is quite [...] uncertain [...] that has not been because any sane Court has for half-a-century doubted the wisdom of fully incorporating the relevant usage of trade into the agreement and into the decision on adequacy of performance.”).

17 Although commentators have long recognized local differences in customary practices, many have dismissed them as trivial without providing any
industries’ efforts at codification, and the subsequent operation of modern merchant-run private legal systems, also suggest that merchants differentiate between written and unwritten customs and that their understanding of customary practices is both different from and far more nuanced than Llewellyn’s.

The understanding of merchant reality gained by looking at these codification debates, together with other merchant-related sources, suggests that, given the Code’s flawed empirical basis, it may be time to reconceptualize the role played by custom in commercial transactions and to rethink the wisdom of the Code’s incorporation-based approach to gap-filling and contract interpretation,18 an approach that is endorsed and strengthened in current drafts of proposed revisions to the Code,19 drafts that represent the undeserved triumph of legal realism in commercial law.

---

18 Merchant tribunals’ rejection of the incorporation strategy does not necessarily undermine the “situation sense” component of Llewellyn’s jurisprudence. See generally James WHITMAN, “Commercial Law and the American Volk!: A Note on Llewellyn’s German Sources for the Uniform Commercial Code”, (1987) 97 Yale L. J. 156 (describing the origins and content of Llewellyn’s notion of “situation sense”). The opinions produced by merchant tribunals reveal that arbitrators’ background knowledge of the trade may enable them to better assess the credibility of testimony and may give them a better understanding of the types of evidence that ought to be submitted. Because most merchant legal systems authorize arbitrators to request additional information in lieu of giving the parties a right of discovery, arbitrators’ background understanding of transactional practices should enable them to intelligently exercise this authority.

19 See Permanent Editorial Board for the Uniform Commercial Code, PEB Study Group, Uniform Commercial Code Article 2: Preliminary Report 16-17 (1990) (after reviewing the present Code’s incorporation strategy and its use of open-ended terms like “reasonable” and “seasonable,” the Study Group “endorses the drafting style utilized in Article 2 and recommends that the general sales policies [...] be retained [...]. We recommend that the Drafting Committee consider ways beyond those recommended by the Study Group to articulate these policies and to improve their implementation. The objective is to achieve a more complete utilization of them by the parties and the courts in the resolution of commercial disputes.”). See also id., at 9, 32 (emphasizing that the Code’s provision on course of dealing and usage of trade “is a crucial component of the [Code’s] broad definition of agreement,”) and
To this end, this Article proposes an alternative conception of the types of customs and practices that do exist and the role they play in commercial relationships. It suggests that while the types of generally agreed upon practices that Llewellyn thought merchants viewed as supplying implicit contract provisions do not consistently exist, merchants do consider it valuable to have an understanding of the ways transactions are usually done, an understanding gleaned from a rough aggregation of practices in the market as a whole. It then argues that these types of understandings, which this Article refers to as weak-form customs, provide transactors with a pool of common knowledge that in the early stages of their contracting relationship enables them to better assess whether the other transactor is a cooperator or a defector, thereby facilitating the emergence and maintenance of repeat-dealing cooperative contracting relationships. On this view, weak-form customs, which are different from the strong-form Hayekian customs whose existence is assumed by the code, can be understood as providing transactors with a set of vaguely defined yet workable relationship-creating norms that initially add tremendous value to contracting relationships but that gradually diminish in importance as contracting relationships mature.

Part I of this Article looks at several associations’ attempts to codify custom and discusses the methodological limitations of drawing conclusions from the codification debates. Part II briefly revisits and critiques the justifications that have been developed for the incorporation strategy. Part III proposes a theory of the relationship-creating role played by weak-form customs in commercial transactions. It then provides suggestive, though not conclusive, evidence that it is both a plausible and analytically useful account of merchant interactions. Part IV concludes that while some industry-wide usages of trade do exist, and highly local customs might have existed, the pervasive existence of usages of trade and commercial standards, whose geographic reach is coextensive

---

recommending that the section on course of performance or practical construction be moved to Article 1, because “[t]his important principle of interpretation should not be limited to contracts for the sale of goods”).

20 See infra, notes 174-175 and accompanying text.
with the reach of the relevant trade, is a legal fiction rather than a merchant reality. It therefore suggests that new justifications for the Code’s interpretive approach and gap-filling methodology are needed.

I. Trade Associations’ Attempts to Codify Custom

Between 1860 and the mid-1900’s, a period during which many merchant industries had already become significantly national in scope, numerous national trade associations created private arbitration tribunals and soon thereafter began to codify industry practices into written trade rules. These rules typically state that they are either codifications or selective codifications of industry customs, or attempts either to clarify, or to achieve uniformity in, industry practices. The preambles to many sets of trade rules

---


22 See, for example, The American Yarn Spinners Association, Inc. (AYSA), The Yarn Rules of 1989 Introduction (AYSA 1989) (“For more than fifty years, members of the Yarn industry have utilized industry rules regarding contract terms and conditions and industry norms for the sale of yarns. The Yarn Rules have been used as a statement of trade practice and from time to time have been revised to reflect developments in the industry over the years. The Yarn Rules and the customary contract terms are recommended to serve as a reference to members of the yarn industry of trade practices commonly in use throughout the industry.”).

23 See, for example, Rubber Trade Association of New York, Inc., General Rules, Prefatory Note (revised 1978) (“The compilation is the result of a study of the usages and customs that have developed in the trade and, in the opinion of the Association, is representative of those terms and conditions of sale which will eliminate misunderstandings between buyer and seller.”).

explicitly state that their goal is to reduce the number of commercial misunderstandings that arise by clarifying the scope and content of contractual obligations. In most industries, these trade rules provide a detailed set of contract default rules covering most aspects of contract formation, performance, and breach.

In most associations, trade rules are drafted and subsequently amended by committees of experienced industry members who serve without compensation and are either appointed by the board of directors or directly elected by the membership. In many industries, both the original rules and subsequent amendments are subject to floor debate at the associations’ annual conventions and must be approved by a specified percentage of the membership as a whole.

---

25 See, for example, American Peanut Shellers Association, Official Trading Rules: Farmers Stock, Domestic and Export 2 (May 1994) (noting that the rules were adopted to “make more definite the terms of contracts of purchase and sale”). See also, infra, note 33.

26 See, for example, NHA, Report of the Twelfth Annual Convention (1905) (reporting verbatim debates). Other associations mail copies of proposed changes to their members or print proposed changes in trade publications.

27 Many by-laws give each member firm (regardless of its size) one vote on proposed trade rules. At the National Grain and Feed Association, trade rules may be temporarily amended by a vote of two-thirds of the Board of Directors, Bylaws of the NGFA, Art. XVII § 2, reprinted in NGFA, NGFA Trading and Trade Rules Seminar (1994) (no pagination in original), subject to “an affirmative vote of two-thirds of the voting power present at the next annual meeting of the members,” id., at § 2 (see also id., at § 3), with each “Active member and each Affiliated member [...] entitled to one vote,” id., at Art. IV, § 1 (see also § 2). In associations with this voting structure, even if large firms control the rules committees, it will be difficult for them to secure passage of rules that are greatly biased in their direction. On the other hand, if they do control most committees, small firms would have a hard time even
This section explores the efforts of trade associations in core merchant industries to codify their commercial customs into trade rules. It finds that, contrary to the assumptions of the Code drafters, even in close-knit communities, merchant transactors do not, except within very local spheres28 or in very general ways, have similar views about the meaning of common contractual terms or the content of precisely those types of commercial standards and usages of trade that the Code and its Official Comments direct courts to take into account in deciding cases.

A. Hay

The National Hay Association (“NHA”)29 was formed in 1895.30 It began arbitrating cases in 189931 and adopted its first getting the rules they might want onto the agenda. It is, however, important to note that in some associations, proposed rules and rules amendments must be approved by a sellers’ group and a buyers’ group. See, for example, American Textile Manufacturers Institute, Inc. (“ATMI”) and American Cotton Shippers Association (“ACSA”), The Southern Mill Rules for Buying and Selling of American Cotton 32 (ATMI & ACA 1995) (requiring the approval of members of the ACSA and the ATMI).

Most Bylaws provide for annual review of the trade rules, but the actual frequency of rule amendments varies widely. See NGFA, Trade Rules and Arbitration Rules (pamphlet) (1995) (noting that the Grain Rules were amended 59 times and the Feed Rules 33 times); see also, infra, note 84 (Worth Street Rules) and infra, notes 114, 118, 120 (silk rules).

28 In the context of the incorporation debates, there are also reasons to be skeptical about strong statements suggesting that local customs exist. If, for example, a transactor is arguing for adoption of a particular rule (especially one that is favorable to his locality rather than simply to a subset of firms in it), he might invoke the alleged universality of the practice in his locality to give his argument legitimacy and persuasive force.

29 The NHA is “made up of producers, dealers, brokers and representatives of related industry […] [It] is dedicated to the development and maintenance of better quality hay and improved marketing practices.” NHA, National Hay Association (Association brochure mailed to author in 1997).

30 At the time the NHA was formed, the interstate trade in hay was fairly well established, having begun in the 1840’s. See William J. REINKE, Arbitration in the National Hay Association 2 (1955) (unpublished Ph.D. dissertation, University of Chicago, on file with the University of Chicago Libraries) (“The shipment and marketing of hay is reported to have originated in the 1840’s.”).

31 In 1897, a bylaw amendment to provide for arbitration was proposed and apparently was adopted. NHA, Report of the Fourth Annual Meeting 46
set of trade rules in 1907, after two years of work on the project.\textsuperscript{32} The goal of the rules creation process was not to memorialize already uniform customary practices, but rather to prevent disputes by actively promoting uniformity.\textsuperscript{33} Prior to the adoption of the rules:

\begin{quote}
[P]acking, shipping and handling hay was an irregular business. There were no established customs to govern, and every transaction was typical of the parties engaged in it. Balers and shippers followed the bent of their own inclination in the details of baling, weighing, buying and shipping, and distributing markets also points of consumption, were under local influences and often dominated by whimsical notions, and at the same point of shipment, or in the same receiving market there was irregularity of method or consistency in business.\textsuperscript{34}
\end{quote}

\textsuperscript{32} NHA, Report of the Eleventh Annual Convention 34-35 (1905).
\textsuperscript{33} See NHA, Report of the Twelfth Annual Convention 170 (1906) (“[D]isputes are more apt to arise, owing to misunderstandings than to anything else. I think a majority of the people mean to do right, but misunderstandings will creep in. Therefore it seems very important that this Association should have trade rules.”). See also NHA, Constitution and By-Laws, “Preamble,” reprinted in NHA, Report of the Second Annual Meeting 29 (1895) (declaring that a goal of the association is to “use our best efforts to have established and maintained uniformity in commercial usages and in the grades of hay and straw in the different markets of the country”); NHA, Constitution, “Preamble,” in NHA, Report of the Eighth Annual Meeting 42 (1901) (reaffirming the Association’s commitment “to use our best efforts to have established and maintained uniformity in commercial usage”).
\textsuperscript{34} NHA, Report of the Eighteenth Annual Convention 3 (1911).
In fact, hay transactors were not even able to agree on the meaning of trade terms as basic as “bale” and “N° 1 Hay.” As one industry participant noted:

What is a bale of N° 1 hay? There is not a man in this room can tell you. Put twenty bales of different grades of hay along that room, and there will not be five men among you who will agree. You have decided that a ton of N° 1 hay may contain not over one fifth of tame grasses. A gentlemen spoke up and said as his opinion that tame grass is clover. If you know anything, you know that a bale of hay with one-fifth clover is clovery mixed hay. N° 1 hay should be pure timothy.

And, as another observed:

The large bales of New York and New England means a different bale from the large bale in the Western States, and the same is true of the small bales. In Chicago at present there is a lack of clear definition of small bales.

---

35 NHA, Report of the Fourth Annual Convention 40 (1897). See also NHA, Report of the Twelfth Annual Convention 37 (1905) (“Grades cannot be expected to suit the South, North, East and West with the same degree of satisfaction.”); NHA, Report of the Eighteenth Annual Convention 154 (1911) (“The hay trade of this country will not be on a sound basis, on an honest basis, until N° 1 hay is N° 1 hay East, West, North, and South.”); NHA, Report of the Twenty-Seventh Annual Convention 54 (1920) (“There is at present no uniform grade or at least no uniformly interpreted grade that can be used as a medium for making sales or purchases of hay in territories remote from each other. What is considered as N° 1 timothy, for example, in one producing section may be considered as N° 2 timothy in another producing section, and still of another grade in the consuming section to which it may be shipped.”). Disputes over grade were not due merely to subjective disagreements about quality assessments, but rather were due, in part, to the fact that different markets defined the same grade differently. NHA, Twenty-Seventh Annual Convention at 74-77 (comparing the grading requirements of eleven sets of association and exchange grading rules and finding wide differences); NHA, Report of the Twenty-Eighth Annual Convention 70-75 (1921) (recounting a long debate as to whether “Choice” hay or N° 1 hay is the highest grade, with some participants questioning whether “Choice” was even a recognized grade designation).

36 NHA, Report of the Tenth Annual Convention 76 (1903). See also id., at 80 (“[B]ales are not governed by size so much as by weight in the Northwest. In Chicago, I know, they like light bales, weighing from eighty-five to ninety-
The debates surrounding the adoption and amendment of the hay rules also suggest that there were no agreed upon usages in relation to some of the precise aspects of a standard transaction that the Code and its Official Comments explicitly direct courts to discern by reference to usage of trade or commercial standards.

For example, the Code provides that when the time for “delivery or any other action under a contract” is not specified, it “shall be a reasonable time,”\(^\text{37}\) determined by reference to “commercial standards,” “acceptable commercial conduct” and “usage of trade,”\(^\text{38}\) However, the debate over a proposed rule, which would determine when certain freight charges had to be requested, reveals that there was no agreement as to what a reasonable time might be.\(^\text{39}\) As one participant opined, “that

five pounds; and in the East they like heavier bales. In Wisconsin they will put in 125 to 135 pounds, and that makes a pretty heavy bale.”); NHA, Report of the Fourteenth Annual Convention 72 (1907) (“[T]here is no general definition of the terms as a Small Bale in one section may mean a Medium Bale in another, or a Medium Bale a Small Bale in another, a Medium Bale a Large Bale in another and so on.”). The causes of this disagreement were partly technological since different hay baling machines produced different size bales of hay. Similar disagreements existed in other industries. See, for example, ASTA, ASTA Yearbook 59 (1914) (“Seedsmen handle large quantities of [...] seeds [...] for few of which legal weights per bushel have been established. They have, therefore, to arrive at customary weights only, which vary in the different States.”) (emphasis added). The weight associated with the designation “bale” in the silk industry also varied from country to country. See Revision of Raw Silk Rules Completed, 3 Silkworm 76 (May 1921).

\(^{37}\) U.C.C. § 2-309.

\(^{38}\) Id., at § 2-309 cmt 1. See, for example, Superior Boiler Works v. R.J. Sanders, Inc, 711 A.2d 628 (1998) (“In the usual case the question of what constitutes a reasonable time under the U.C.C. is one for the finder of fact to determine from [...] several factors, including] usage of trade in the pertinent industry.”); James Town Terminal Elevator, 246 N.W.2d 736, 740 (N.D. 1976) (holding that “based on the previous ‘course of dealing’ of the parties and the industry’s ‘usage of trade,’ the jury could determine that Aug. 31, 1973, was a reasonable time for delivery”).

\(^{39}\) See also, NHA, Report of the Sixteenth Annual Convention 220 (1909) (observing that “[t]he word ‘ample’ [as used in a rule requiring “ample margin”] may not have the same meaning in the minds of different people”); NHA, Report of the Twenty-Eighth Annual Convention 68 (1921) (“You will see at the end of each paragraph ‘well baled.’ That term is so indefinite [...] [T]here should be something more definite brought into it.”); id. (“I think the
‘reasonable time’ business will not tell anything. You might as well leave it out.” 40 And, in response to a suggestion that a more definite rule be adopted, one transactor proposed “nine months,” another “fifteen days,” and still another, “within ten days after the freight bills have been paid.” 41

Prior to adoption of the rules, there was also widespread disagreement about the meaning of the type of common contractual language that the Code and its Official Comments direct courts to interpret “as meaning what it may fairly be expected to mean to parties,” which is determined, in part, by looking to “usage of trade as a factor in reaching the commercial meaning.” 42 For example, while hay contracts tended to include delivery-time provisions like “prompt” or “immediate,” provisions the Code would look to usage to define, 43 there was no consensus as to what those terms meant. 44 This lack of consensus led to many misunderstandings.


41 Id., at 222-223. Only one participant spoke in favor of the reasonable time proposal, saying “if it requires fifteen days in their market that is a reasonable time for them. If it can be done in three days in another place that is a reasonable time in their market. I think anyone can readily form an opinion of what reasonable time is.”

42 U.C.C. § 1-205 cmt 4. See also id., at § 1-205 cmt 1 (“This Act rejects both the ‘lay dictionary’ and the ‘coveyancer’s’ reading of a commercial agreement. Instead the meaning of the agreement of the parties is to be determined by the language used by them and by their action, read and interpreted in light of commercial practices and other surrounding circumstances.”).

43 See, for example, Kreglinger & Fernau v. Charles J. Webb Co., 162 F. Supp. 695, 697 (1957), aff’d, 255 F.2d 680 (3d Cir. 1958) (noting that the “whole controversy centers around the single question, What did the parties intend when they inserted the word ‘prompt’ in the contract?” the court looked to trade usage to define “prompt” but concluded that no usage existed, “because the plaintiff contends it meant ‘by first available ship’ and defendant that it meant thirty days”).

44 See NHA, Report of the Twelfth Annual Convention 155 (1905) (quoting a member as stating): “Suppose [...] I should purchase of a shipper a carload of hay for prompt shipment. Technically that means nothing. It does mean something if I purchase a car of oats on the [...] floor” of an official exchange with written trade rules (emphasis added). Similarly, the Code defines a “Commercial unit,” in terms of “commercial usage,” explicitly noting that it
The trade rules were conceived of as a way to avoid such misunderstandings by providing “a sort of dictionary to which all the members of this Association can go.”

The hay industry codification debates also suggest that some of the Code rules relating to the formation of contracts that rely on trade usage and commercial standards for their substance may be problematic. For example, while the Code directs courts to look to “commercial standards on the point of ‘indefiniteness’” in determining whether an agreement is sufficiently definite to be enforced. However, hay arbitrators, like merchant arbitrators in many industries, routinely point out that contracts ought to be far more detailed than they are and frequently chastise parties for entering into unduly vague agreements.

---

may be defined in terms of “a quantity [such as] a bale [...] or carload.” U.C.C. § 2-105(6). However, in the hay industry there was no accepted meaning of either “bale” or “carload.” In 1905, in a debate over whether trade rules should be adopted, one participant noted that “it is a very grave question of doubt in any state what is meant by the shipment of ten cars,” NHA, Report of the Twelfth Annual Convention 156 (1903). And, in 1910, the Arbitration Committee decided to “respectfully recommend that [the NHA] establish a rule to be added to our Trade Rules in which the number of tons of hay is specified that shall constitute a contract carload.” NHA, Report of the Seventeenth Annual Convention 67 (1910), but the issue was not fully debated until 1921. NHA, Report of the Twenty-Eighth Annual Convention 85 (1921) (recounting the debate over the meaning of the term “carload,” with some people saying it meant “10 tons” and others asserting “12 tons”). See, infra, note 69 and accompanying text.

NHA, Report of the Twelfth Annual Convention, 155 (1905).

U.C.C. § 2-204 cmt.

See, for example, Hinson v. Parker Grain Co. Arbitration Case No!1628 (Dec 23, 1986) (“[T]he arbitration panel was unanimous in wishing to caution the trade as to the necessity of using well-thought-out contracts that clearly encompass all of the obligations of both buyers and seller to each other.”); ASTA, ASTA Yearbook 164 (1927) (“Counsel has rarely, if ever seen a so-called business-man’s contract which was not shy of some desirable provision.”).

H.H. Driggs v. Walters Bros, in NHA, Report of the Eleventh Annual Convention 145-146 (1905) (“[W]e find many of the terms of a contract of this kind ambiguous and indefinite.”); NHA Committee on Arbitration, Case 676; in NHA, Report of the Twenty-Eighth Annual Convention 140, 141 (1921) (“This controversy arises principally because plaintiff’s confirmation in regard to time of shipment was indefinite, and further defendants also exercised gross carelessness in respect to confirmation in accordance with...
Despite these disagreements over the content of general usages and the meaning of common contractual language, a few general usages may have existed, and local pockets of customary practice were not uncommon. Many rule adoption and amendment debates, such as those relating to grading and inspection, took the form of debates among representatives of

custom and trade rules.”); NHA, *Report of the Twenty-Ninth Annual Convention* 108 (1922) (“We [the Arbitration Committee] are impressed that in most all cases trouble arises from the fact that contracts are carelessly made and too much is taken for granted by the contracting parties. It might be well to adopt a slogan, ‘Take nothing for granted.’ Both parties to a contract should file written confirmation, clearly defining all specifications and details governing said contract. All modifications should also be confirmed in writing and many cases for arbitration would be eliminated.”); NHA, *Report of the Thirteenth Annual Convention* 59, 61 (1923) (where the arbitration committee notes: “there is one suggestion I would like to leave with you and it is this—in making your sales and purchases pay more attention to trade rules No 1 [setting out what a contract should specify] and No 2 [dealing with sending confirmations] [...]. Make definite contracts and observe your trade rules, and when this is done membership in The National Hay Association will be a real asset to your individual business.”); NHA, *Report of the Twenty-Fifth Annual Convention* 131 (1918) (noting “the fact that 75% of the cases which come before this [Arbitration] Committee would be avoided if contracts were made definite as to the time of shipment, grade, weights, and proper confirmations mailed by both parties”); NHA, *Report of the Seventeenth Annual Convention* 48 (1910) (where the arbitration committee attributes the rising number of cases to the use of “poor and faulty contract[s].” and recommends the adoption of an association-drafted “uniform explicit and binding contract for the use of its members.”).

49 For example, in most cash commodities industries, price adjustments were often given for slightly nonconforming tender. See, for example, *Jones v. Henderson* (NGFA 1904), reprinted in NGFA, *Decisions of the Arbitration and Appeals Committees of the Grain Dealers National Association* 23 (1920) (In a dispute over rejection of an off grade delivery of corn, the arbitrators held that while “custom has held that where off grade grain is shipped, discounts may often permit contracts to be filled by applying such cars or quantities that are not equal to contract grade to be applied on contract at a difference. But there is no rule in the grain trade making it obligatory on the purchaser to accept lower grades of grain.”).
regions and localities, each with their own practices. In some localities these customs took the form of unwritten practices, while in others they were codified in the rules of local hay associations or commodities exchanges.

More generally, much of the impetus for both codification and rule amendments came from members of the NHA’s arbitration committee who sought clearer guidance about how to decide cases. Although hay arbitrators took unwritten custom into account in filling gaps, they often did so with some reluctance and the industry itself had a strong preference for clear and bright-line rules.

---

50 While some of the practices so debated had distributional consequences, others such as how to define the quality associated with each grade designation—see, for example, NHA, Report of the Twenty-Seventh Annual Convention 74 (1920) (comparing grade rules of 11 sets of exchange rules)—related to routine matters where only coordination was important but transactors nonetheless fought for their preferred definitions. See NHA, Report of the Twentieth Annual Convention 46-50 (1913) (debating grade changes). Similar obstinacy was exhibited in the grain industry. See 17 Who is Who in the Grain Trade (“WWGT”) 31, 33 (Jan. 5, 1927-28) (after noting that many local rules relating to shipping time contradicted the Grain Dealers National Rules, the chairman of the Trade Rules Committee stated that: “in taking the matter up with the markets that differ we find on a whole that they are very conservative and tenacious in maintaining their rules but we are working on the question and gradually obtaining results”).

51 See for example, NHA, Report of the Fourth Annual Convention 24-25 (1897) (“[W]e are old fashioned folks at Boston, and this Association must not forget one thing, that what is applicable to one section of the country is not applicable to another”); NHA, Report of the Twenty-Eighth Annual Convention 68-72 (1921) (containing a debate over grades that emphasizes the existence of regional differences).

52 NHA, Report of the Twelfth Annual Convention 154 (1905) (“We wished to recommend the adoption of trade rules. We think it necessary for this Association to have trade rules printed. It would simplify the work of the Arbitration Committee and would make less work for that committee to do.”).

53 NHA, Report of the Twelfth Annual Convention 74 (1905) (“We should always keep in mind the benefits which associations like the Chicago Board of Trade and the New York Stock Exchange confer on their members, are due to their rigid rules, the basis of which is business integrity and mutual interest.”)
B. Grain and Feed

The National Grain and Feed Association ("NGFA") was formed in 1896 and began arbitrating disputes shortly thereafter. It adopted the first Grain Trade Rules in 1902 and the first Feed Trade Rules in 1921. Achieving uniformity of rules and practices was an important goal of the rule adoption and amendment process.

Prior to the national rules, the written trade rules of different local markets varied widely, and unwritten customs

---

54 The NGFA, formerly the Grain Dealers National Association, is a trade association "for the North American grain and feed industry." Today, it has "more than 1,100 [member] companies [...] that store, handle, merchandize and process more than two-thirds of all U.S. grains." David C. BARRETT, Jr., "Arbitrating Agricultural Disputes!: The National Grain and Feed Association's Experience" 1-2, reprinted in NGFA, NGFA Trading Seminar (no pagination in original), supra, note 27. For a more detailed description of the industry’s contemporary private legal system, see L. BERNSTEIN, loc. cit., note 13, 1771-1787.

55 Cross country trade in grain was already common by the 1840’s and accelerated quickly after the growth of railroads in 1898. William CRONON, Nature’s Metropolis: Chicago and the Great West, ch. 3 (1991). Many grain traders were buyers one day and sellers the next, whereas in feed markets many transactions were between merchant sellers and end users.

56 However, the first written arbitration opinion was not issued until 1902. Randall C. GORDON, A Century of Agricultural Abundance Through Free Enterprise: A Centennial Observance of the National Grain and Feed Association 53 (NGFA 1996). Prior to this time, numerous local associations had their own arbitration tribunals. Today, however, in the cash trade, only the Pacific Northwest Grain Exchange, the Los Angeles Grain Exchange, the Colorado Grain and Feed Association, and the Texas Grain and Feed Association, continue to conduct their own arbitrations.

57 President Clement makes Committee Appointments, 10 WWGT 25 (Nov. 20, 1920-21) ("After your committee has drafted trade rules and regulations which have been adopted by your Association, we recommend that your Association suggest to the various exchanges of this country that they in turn adopt these same rules and regulations so as to insure uniformity in transactions in mill feed.").

58 R. C. GORDON, loc. cit., note 56, 49. ("[W]hen the Grain Dealers National Association was founded [...] each terminal market and grain exchange operated under its own set of trading rules and dispute-resolution procedures. Very little consistency existed between the trading rules of different markets."). In addition, the rules of local associations and exchanges were not regarded as necessarily establishing general trade customs. See, for example,
were non-uniform even within particular localities. As a grain dealer observed at the 1901 convention, many grain men “differ in opinion. Their ideas concerning what should and should not be done under certain conditions varies as much as the markets themselves.” The divergence in opinion over the content of

---

Ewart Grain Co. v. Wells-Abbot-Nieman Co. (Apr. 28, 1920), reprinted in NGFA, Decisions of the Arbitrators and Appeals Committees of the Grain Dealers National Association 316, 317 (1920) (“[T]he rules of various exchanges which permit second inspection and discount based thereon, are not an expression of general trade customs and are applicable only at markets which carry them in their rules.”). The differences in the rules of local markets persisted even after the promulgation of the national rules, despite numerous efforts of the national association to promote uniformity. Effort to Unify Exchange Trade Rules Revived, 13 WWGT 27 (Dec. 20, 1924). Some of these differences were attributed mostly to local dealers attempting to gain advantage. Are the Association’s Trade Rules in Danger?, 13 WWGT 28, 29 (Feb. 20, 1923-24) (“It is safe to say that in the local rules governing the different boards of trade there is unnecessary protection for the local members.”). However, differences also existed over unifying practices with no distributional impact. See, supra, note 23. Conflicts over whether a particular region’s rules and customs should govern a particular transaction were not uncommon sources of disagreement between transactors in different localities. See, for example, Trade Rules: Chairman Replies, 13 WWGT 39, 41 (July 5, 1922-23) (“While it is true as you state, that the general custom in your market is to buy on arrival drafts, it is also customary in the Western markets to make all sales subject to demand drafts unless otherwise specified. Therefore if the seller in this case was not familiar with the Richmond market and made the trade in good faith presuming that the Kansas City custom of demand drafts was understood it might raise a question as to whether there was any contract at all as there would be no agreement or meeting of the minds on this term of the contract.”). In addition, the differences in local rules, and the unwillingness of local markets, associations, and exchanges to change their practices even as to aspects of trade that had no clear distributional impact, were a frequent subject of trade press columns. See, for example, 17 WWGT 31, 33 (Jan. 5, 1927-28) (“[U]niformity in the rules is a matter of slow growth as the different markets are very conservative in making changes in rules of years standing.”) See also, supra, note 50 and accompanying text.

---

See, for example, Minutes of Meetings: Secretary’s Book 111 (Nov 9, 1896) (“Secretary’s Book”) (unpublished book of clippings; copies on file with author) (reporting that the Illinois Grain Dealers’ Association created trade rules, to, among other things, “establish and maintain uniformity in commercial usages as far as the grain trade is concerned”).

---

R. C. Gordon, loc. cit., note 56, 51. See also The President’s Address, in Secretary’s Book, op. cit., note 59, at pp. 117-118. (Peavy Address) (“Trade rules and custom will be an interesting subject and will help to enlighten us in regard to some of the ambiguous rules and customs now in vogue.”).
industry-wide trade practices was so wide that “a high probability existed that a dispute would arise whenever grain was traded outside local markets.” More generally, as a transactor noted in an article published before the rules were drafted:

As the situation now is, there are customs in the grain trade that are supposed to be established, but the trouble in respect to them is, they are not fixed, are not understood alike, some understand them in one way and others in another way, and for that reason, if for no other, [they] cause difficulty.

In response, in part, to this uncertainty, a Grain Trade Rules Committee was appointed. In 1901 it submitted fourteen recommendations relating to trade practices that it hoped would “do much to bring about a more uniform custom, eliminate friction and foster a better understanding and closer relations between the interior dealer at primary and intermediate markets.” After submitting these recommendations, the committee sought additional time to draft a complete set of rules, explaining that the “wide area to be covered and the diverse interests to be equally represented” made its task difficult. It asked the state associations and local exchanges to “formulate their trade customs into a set of rules governing the transactions,” so that the national committee would have a basis from which to work.

---

61 R. C. GORDON, loc. cit., note 56, 49. In addition, as an early trade publication explained, “the farther the buyer is from the seller, the more the need of a careful and distinct understanding, for reasons well known.” Secretary's Book, op. cit., note 59, at 111. Even today, in the Texas Feed Trade, transactors pay “much more attention to contract when you deal with those in other areas.” Statement of Presenter, TGFA Conference (1999).

62 Secretary's Book (no pagination in original), op. cit., note 59.

63 R. C. GORDON, loc. cit., note 56, 50. These recommendations dealt with confirmations, time for shipment, billing instructions, shipment, demurrage, sample sale, loading, terms, telegram, acceptances, surplus shipments, regular market terms, interior shipments and invoices. Secretary's Book (no pagination in original) (1901), op. cit., note 59.

64 R. C. GORDON, loc. cit., note 56, 51.
The first set of grain rules was quite detailed. However, some aspects of trade were not covered because customs differed significantly across the country and consensus could not be achieved. As one trade journal column noted:

*I regret that it is not possible to give you a definite rule covering the case. The Grain and Feed Dealers National Association for two or three years endeavored to frame a rule which would cover a condition of this nature, but we found that the customs vary so largely in different sections of the country and so many technical questions arose that it was impossible to reach an agreement that would be national in scope.65*

When the first set of rules was finally introduced, merchants had many questions about how to interpret them. In

---

65 *Chairman Replies*, 20 WWGT 31 (Oct. 5, 1930-31). See also, *Chairman Replies*, 21 WWGT 28 (Dec. 5, 1931-32) ("[T]he rules of the Grain and Feed Dealers National Association cover the broad principles of trading and details might differ as to the rate of interest, or the custom of charging to a certain date, in different sections of the country, so we have not attempted to formulate a rule on this point."); *Hankerson Proposes a Change in the Trade Rules*, 14 WWGT 27 (July 5, 1927-28) ("[W]hile a rule [regarding diversion dates] has been suggested several times, none has been adopted, because it has been impossible to frame a rule that would meet all conditions or would state the general principle that would govern all transactions. It follows that each case would be settled on its merits with due consideration to customs of the trade in specific territory.").
response, the Association solicited “Ask the Chairman” letters from members and published detailed answers in the Association’s trade journal. These columns reveal that an inability to agree on the content of custom was not only a reason for certain gaps in the trade rules, but was also a common cause of disagreement among transactors.

Even after the NGFA adopted Trade Rules, large variations in the written rules of local associations continued to exist. As the Chairman of the Trade Rules Committee explained in 1924, achieving “uniformity in the rules is a matter of slow growth as the different markets are very conservative in making changes in rules of years standing.”

The grain trade press also confirms that prior to codification there was no agreement concerning the meaning of many basic contractual terms. For example, the meaning of the term “carload” led to numerous controversies and questions

---

66 These types of columns were common in the merchant trade press and typically contained many inquiries about the content of custom. See, for example, National-American Wholesale Lumber Association Inc. (“NAWLA”), Questions and Answers, North Coast Weekly Letter No. 140 (Sept. 16, 1924) (“Question!: ‘A wholesaler has taken a cash discount on a Final Settlement sent to us four months after the date of shipment. Is he entitled to this discount and what is the custom in regard to this?’”); “Question!: ‘Where order is accepted for ‘85% 10’ and longer; balance shorter; usual lengths, well proportioned’ what does ‘well proportioned’ mean according to present customs?’”); NAWLA, Questions and Answers, North Coast Weekly Letter No. 150 (Nov 25, 1924) (“What is the custom in making final settlements with mills when the wholesaler has purchased ‘less 5% commission, underweights to mill to final destination?’”). See also note 141 (discussing the question and answer column in the silk industry journal, “The Silkworm”).

67 *Are the Association’s Trade Rules in Danger?,* 13 WWGT 28-29 (Feb. 20, 1923-24) (extensively discussing differences in practices and rules from locality to locality).


69 See, for example, *Dispute Over Size of Cars: Smith Bros Grain Co. v. Security Mill and Feed Co.,* 10 WWGT 28-29 (Oct. 5, 1920-21) (The “claim in this case arises over the question of the size of cars made applicable on a contract.”).
from industry members.\textsuperscript{70} Finally, it was suggested that grain should simply be sold by weight and the designation “carload” abandoned.\textsuperscript{71} Eventually a rule was adopted defining carload in terms of weight, but even after its adoption disputes continued to arise and the rule was amended several times.\textsuperscript{72} There was also no consensus as to the meaning of particular grades.\textsuperscript{73}

Unlike the Grain Trade Rules, the Feed Trade Rules did not take long to draft once the codification effort began. However, the need “for uniform rules” had been obvious for years. Prior to adoption of the rules, trade practices varied widely: “The feed men [...] never had uniform rules to govern their transactions [...] there were many sets of rules in different localities but there was no uniformity.”\textsuperscript{74} Although “sporadic

---

\textsuperscript{70} See, for example, \textit{Replies of Chairman Watkins}, 10 WWGT 38, 40 (Feb. 20, 1920-21) (answering what is meant by the term “capacity” car); \textit{Replies of Chairman Watkins}, 10 WWGT 42 (Feb. 5, 1920-21) (addressing the question what is meant by the “contents of one 80 capacity car”); \textit{Replies of Chairman Watkins}, 11 WWGT 43 (Dec. 5, 1921-22) (answering the question “what constitutes a car load when a sale of grain is made without any reference being made as to the size of the cars.”).

\textsuperscript{71} In a particular controversy there is reason to fabricate a dispute relating to the definition of a carload since the answer may have a key distributional impact. As one trade article noted, if the market “has declined [...] you are liable to find 60,000 capacity cars. Should the market advance the cars furnished will be 30,000 capacity.” \textit{Secretary’s Book, op. cit.}, note 59, at p. 50.

\textsuperscript{72} See \textit{Trade Rules of the Grain Dealers National Association} (1920), reprinted in 10 WWGT 39-40 (Nov. 20, 1920-21) (A carload shall consist of bushels as follows!: “Wheat 1,1000; shelled corn, milo maize, kaffir corn, and feterita, 1,100; ear corn 700 [...]. Provided, that where the rules of carriers lawfully on file with the Interstate Commerce Commission or State Railway Commission provide for minimum carload weights in excess of the above, such minimum weights shall constitute a carload within the meaning of this rule”); \textit{Trade Rules}, 17 WWGT 33-34 (Jan. 5, 1927-28) (discussing a letter on the meaning of carloads). See also \textit{Trade Rules Changes}, 13 WWGT 38 (Oct. 5, 1923-24) (discussing the amended Rule 32 dealing with the definition of carloads).

\textsuperscript{73} See, for example, \textit{Secretary’s Book, op. cit.}, note 59, at p.54. (“[T]he grading of the different cereals in our markets in the United States vary so widely that it is almost impossible to tell by the inspection at one market on any kind of grain, what the same class of grain will grade in some of the other markets.”); \textit{id.}, at 63 (“We should discuss plans and adopt measures which would lead to uniform grading.”).

\textsuperscript{74} \textit{Trade Rules to Govern Transactions in Feedstuffs}, 10 WWGT 25 (Apr. 5, 1920-21). When the feed rules committee was first appointed, it was directed
efforts were made from time to time to achieve this most desirable object,"75 consensus could not be achieved and the efforts to adopt written trade rules had always failed. Finally, due to the efforts of a strong-willed man, a Feed Trade Rules Committee was appointed. It met in intense session for two days and came up with a draft set of rules. In reporting the work of his committee to the association as a whole, the committee chairman did not attempt to give the rules legitimacy by claiming that they codified custom. Rather, he explicitly acknowledged that they were the result of compromise, noting that “[w]e succeeded in formulating rules that all of us agreed upon as being the most equitable rules that it was possible to draw up. We were in session two full days. Every point was threshed out.”76 The fact that the feed rules were not based strictly on custom, and that they were similar in many respects to the grain rules, might account, in part, for the relatively short amount of time it took to draft them.

Although the NGFA Rules Committee has met annually since 1902 “to review the rules [...] [in an effort] to ensure that the Trade Rules reflect—but do not set—industry trade practices,”77 most rule changes have in fact been attempts to clarify vagueness,78 respond to technological changes,79 change customs,80 or to simply adopt more desirable practices.81

---

75 Id., at 25.
76 Id. (statement of E.C. Dreyer, Chairman of the Feed Committee).
77 R. C. GORDON, loc. cit., note 56, 52. Since their inception, the Grain Trade Rules have been amended fifty-eight times and the Feed Trade Rules thirty-two times. Id.
78 For example, the rules defining “business day/holiday” became more specific over time.
79 See Trade Rules and Arbitration Rules, 14 Grain Trade Rules, Rule 45 (“Electronic Data Interchange”) (1995) (making the trade rules applicable to “trades that include electronic transmission and receipt of data in agreed formats in substitution for conventional paper-based documents”).
80 See, for example, Grain Trade Rules, Rule 29 (1922), reprinted in 12 WWGT 37 (Nov. 20, 1921-22) (“The specifications of a contract cannot be altered or amended without the expressed consent of both the buyer and the seller. This abolishes the custom of ‘silence confirms.’”). See also Hankerson Proposes a Change, 17 WWGT 27-28 (July 5, 1927-28) (“[W]e [the Trade Rules
Today, the Trade Rules Committee is comfortable adopting and amending rules based solely on their desirability and likely effect.\textsuperscript{82}

C. Textile Industry

The debates surrounding the 1936 adoption of the Worth Street Rules ("WSR"),\textsuperscript{83} which in modified form continue to

\begin{flushright}
\textsuperscript{81}See, for example, Proposed Changes in Association’s Trade Rules, 11 WWGT 23 (Sept. 20, 1921-22) (proposing “add[ing] to the present rule” a provision that “[a]ny loss resulting from irregular or incorrect invoices shall be paid for by the seller,” and justifying it on the grounds that “[t]he idea of the proposer of this change is that the invoice shall be given more nearly the same dignity as the Bill of Lading to prevent losses due to careless invoicing’’); NGFA, Trade Rules and Contracts 7 (pamphlet, no date) (“The general objectives of the Trade Rules Committee are […] [among other things] to formulate and recommend to the membership, trading rules that will bring about improvement of marketing procedures in the industry […] and enact new rules as needed to impartially govern transactions.”).

\textsuperscript{82}This conclusion is based on my attendance at a meeting of the committee (Washington, D.C. 1995).

\textsuperscript{83}These rules were jointly adopted by the Textile Fabrics Association, Cotton-Textile Institute, Fine Goods Committee of the Cotton-Textile Institute, International Association of Garment Manufacturers, Union-Made Garment Manufacturers’ Association, American Cotton Manufacturers Association, National Association of Cotton Manufacturers, New Bedford Cotton Manufacturers Association, Wholesale Dry Goods Institute, National Association of Purchasing Agents, Textile Brokers’ Association, and Association of Cotton Textile Merchants of New York. WSR 2 (1936). Before and during the rules-creation process, the textile industry was geographically concentrated “in a compact area of some six blocks, located just north of city hall square in lower Manhattan,” which contained the “market for virtually all of the nation’s output of cotton cloth.” The Association of Cotton Textile Merchants of New York, 25 Years 1 (Parker-Allston 1944) (“Textile History”). Although industry members were drawn from many walks of life all over the country, they developed a close-knit culture of their own. “[T]he first Worth Streeters were incurably addicted to the stovepipe hat and frock coat which continued to be a symbol of this calling, and virtually their
govern most textile transactions today, together with other evidence about the textile trade, suggest that that prior to codification, both trade practices and the industry meaning of particular terms varied widely.

The idea of codifying textile industry customs, in an effort to “improve trade practices,” was first raised at a 1918 meeting at the Union League Club, where the idea “of devising a uniform sales contract to simplify all market transactions” received strong industry support. Shortly thereafter, the Association of Cotton Textile Merchants of New York (“ACTM”) was formed, and it made the adoption of trade rules a high priority, explaining that:

[C]ontracts, to be sacred, must be sound. They must be based upon fairness to buyer and seller alike; and this implies acceptance on all sides of standards of fair trade

uniform, until the presidency of the first Roosevelt.” The Worth Street Story, 22 Am Fabrics 47, 48 (1952). Most merchants ate lunch at one of three eating clubs, and the social ties they formed were so strong that most transactions were repeat and “every year millions of dollars worth of goods were sold without a paper that could be taken to court in the case of a dispute. This unique and really wonderful method of handling even the largest transactions still represents the philosophy behind the Worth Street way of doing business.” The Worth Street Story, 22 Am. Fabrics 48-49, (1952).

Jean E. PALMIERI, “Lawyers Group Publishes Revision of Worth Street Textile Market Rules”, (1986) 16 Daily News Record 9. (“The rules [WSRs] are commonly recognized as the standard code of procedure and trade customs for the purchase, sale and use of textiles and allied products.”). The WSRs were revised in 1941, 1947, 1964, 1971, and 1986. The 1986 revision was undertaken by a group of textile lawyers whose goal was to simplify the rules and make them more accessible. Id. The rules are incorporated into most textile contracts, which also provide for arbitration under the Rules of the General Arbitration Council of the Textile and Allied Trades, which are today administered by the American Arbitration Association. Id.

See Note, “Enforceable Arbitration of Commercial Disputes in the Textile Industries”, (1952) 61 Yale L. J. 686, 711, note 147 (“Before the rules [WSRs], major disagreement was over the inclusion of second quality cloth in each shipment, grace periods after delivery, and the meaning of trade terms. In rayon, where there is virtually no agreement, conflict on the meaning of trade custom is rife [...] [In addition,] the definitions of imperfections reflect the manufacturer’s interest that ‘normal irregularities natural to the fibers used’ exist and that ‘a reasonable number of manufacturing defects must be expected.’ The standards leave ‘normal irregularities’ undefined [...] There is basic disagreement on what non-conformity is”).
practice. Such standards usually evolve from practical conditions as customs or unwritten laws, and men have tried perennially to write them down, to codify them, so that disagreement and misunderstanding might be reduced to a reasonable minimum.\textsuperscript{86}

The process of drafting the WSRs was fraught with conflict, involved negotiations among numerous trade associations, and proceeded slowly. The effort to draft a uniform sales note, which was to become a key section of the WSRs, began in 1910. In 1920, six proposed standard notes were recommended to ACTM members, but as late as 1932 uniformity of practice had not been achieved.\textsuperscript{87} In 1932, after two years of committee work devoted solely to that subject, a standard set of eight Salesnote Clauses was finally published.\textsuperscript{88} In 1934 the first standard Cotton Textile Salesnote was introduced. It was subsequently incorporated into the 1936 WSRs after an additional year of committee work.\textsuperscript{89}

Work on other aspects of the rules, such as the compilations of usages in different branches of the trade, proceeded similarly slowly. In 1928, for example, a committee was appointed “to begin negotiations with representatives of the converters ‘for the purpose of defining some customs prevailing in the purchase and sale of grey goods,’”\textsuperscript{90} an effort that was not completed until 1931.

The 1936 WSRs, a culmination of eighteen years of concerted effort on the part of industry participants, were an extraordinarily detailed set of contract default rules. They covered many stages of the manufacture and distribution of a variety of textile products and defined numerous quality specifications, trade terms, and trade customs for transactions in

\textsuperscript{86} Textile History, loc. cit., note 83, 37-38.

\textsuperscript{87} Id., at 40-41.

\textsuperscript{88} Id., at 40; see also Uniform Salesnote Clauses Recommended to Cotton Textile Trade, Textile World 33 (Apr. 6, 1932) (reprinting the Uniform Salesnote Clauses and a brief statement in their support).

\textsuperscript{89} Textile History, loc. cit., note 83, 41-42.

\textsuperscript{90} Id., at 39.
different types of goods.\textsuperscript{91} Nevertheless, even after their adoption, substantial disagreement over the trade meaning of even words that were themselves defined in the Rules persisted. For example, although the 1936 WSRs included definitions of the widely used designations “first quality” and “seconds,”\textsuperscript{92} precisely the types of terms the Code directs courts to look to usage to define,\textsuperscript{93} a 1949 National Federation of Textiles Report of association activities contains a discussion of attempts to compile “A Dictionary of Trade Expressions!”

\textit{What is meant by a ‘piece of goods’? What does r.o.m. mean? What is a ‘second’? Does ‘as are’ include

\textsuperscript{91} The 1936 WSRs were thirty-nine pages long. They included a Standard Cotton Textile Salesnote, codified customs and definitions dealing with allowances for deficiency, arbitration, “as are,” cancellation, rejections and claims, deliveries, general strike or lockout and normal production, goods not sold by description, latent defect, methods of testing materials in dispute, packing, patent defect, quantities run of the loom or mill, seconds, selection of representative pieces, selvage count, standard seconds and tailing clauses, storage and insurance, strike or casualty, tailings, tape selvage and feeler motions, tensile strength, and use—as well as special customs relating to the converting trade and what is termed “pertinent data concerning the bag trade.” WSR “Table of Contents” (1936). The 1941 and 1947 WSRs were seventy-one pages long. By 1964, the WSRs had expanded to one hundred twenty-eight pages, and even after the 1986 revision, which was aimed at simplification of the rules, they were still fifty-eight pages. Despite the extraordinary detail of the WSRs, traders did not view them as providing the terms of a near complete contingent-state-contract. See Note, loc. cit., note 85, 701, note 88. (“[I]t is quite evident from interviews and questionnaires that the \textit{Worth Street Rules} have many loopholes [...] Many of the existing definitions are inadequate. For example, definitions of first quality or run of the mill depend on a specific mill’s records. These two terms are extremely variable [...] Merchants also point to the failure of the rules to cover many issues raised in disputes [...]. Specifically, textile men report that the rules overlook finished goods and finishing qualities [...]. Correcting these deficiencies, however, is extremely difficult because textile disputes present an infinite number of variables.”).

\textsuperscript{92} \textit{Definitions and Trade Customs}, WSR 32 (1936) (“The word ‘Seconds’ is applied to cloth inferior to that which the subject mill grades as ‘Standard’ or ‘first quality’ [...] Unfortunately, the grading of cloth does not lend itself to specific definition. Practice varies with different mills and for different uses.”).

\textsuperscript{93} See, for example, \textit{Foxco Industries, Ltd. v. Fabric World Inc.}, 595 F.2d 976, 984 (5th Cir. 1979) (holding that in determining the meaning of “first quality” it is “proper to look to trade usages,” since parties are “presumed to have intended the incorporation of trade usage in striking their bargain”).
remnants? These were questions asked so repeatedly during 1949 that at the close of the year plans were under way to compile an official list of trade expressions of this type.94

Similarly, a 1952 academic study of textile arbitration found substantial disagreement among transactors as to the meaning of these and other terms,95 concluding that in the industry as a whole, “trade custom [...] is often amorphous and unsettled.”96 Indeed, the study found that lack of consensus on the meaning of customs was one of the main reasons that transactors strongly preferred three arbitrator panels to single arbitrator adjudication in which the single arbitrator’s idiosyncratic view of trade custom would most likely be applied.97

94 National Federation of Textiles, Inc., The National Federation of Textiles Reviews [...] Activities, Officers, Members (1948, 1949, 1950). See also National Federation of Textiles, Annual Report of the National Federation of Textiles, Inc., 75th Annual Report 19 (1947) (“Coincident with suggestions for revision of the recommended contract [for Rayon] came repeated questions from buyers of grey goods as to the meaning of such phrases as ‘run of the mill,’ ‘as are,’ ‘seconds,’ etc. The cotton market had incorporated definitions of such terms in their Worth Street Rules, but nothing comparable had been prepared for rayon fabrics, despite the apparent general use of similar terms in that branch of the industry. It was thought this might be a subject of future determination through the federation.”).

95 See Note, loc. cit., note 85, 691 and notes 26-27.

Trade practice in classifying goods as first or second quality may facilitate buyer’s efforts to escape. These terms are not absolute; they are based on the particular plant’s past performance, a vague and shifting standard at best [...]. The difference between first and second quality is not clear. By definition second quality goods are merchantable and reasonably free from major defects; yet they are inferior to and contain more imperfections than first quality goods. The definition says no more [...]. In cottons the standard [for first and second quality] not only varies from plant to plant but also from fabric to fabric [...]. Rayon finished goods are classified on an industry-wide standard [...]. But rayon merchants assert that the standard is just as vague as that used for cottons.

96 Id., at 700.

97 Id., at 712, note 152. (“[S]ome converters state that the possibility of trade bias is so great in the absence of agreement on trade custom that they will not go to arbitration.”).
Finally, the WSRs themselves posed a challenge to the Code drafters’ decision to accord written usages (in the form of trade codes) and unwritten usages the same weight in adjudication. The Forwards to various editions of the WSRs reveal that textile merchants conceived of a hierarchy of trade customs. As the 1936 WSR Forward stated, contingencies not covered by the standard salesnote “shall be interpreted in accordance with the established rules and customs of the trade, particularly with those rules and customs which have been formally approved by authorized bodies, representing both buyer and seller.” 98 The WSRs themselves separately listed customs that applied to the trade as a whole and those that applied only to particular subdivisions. Moreover, the 1936 Forward cautioned that constant vigilance was needed to ensure that rules and customs were reviewed periodically so that “those [customs] that are determined to be entitled to recognition,” could be “list[ed] in proper form.” 99 Over time, the WSRs themselves came to be recognized as custom 100 and the rules themselves came to discourage recourse to unwritten custom. The Standard Salesnote in the 1986 WSRs introduced a term providing that arbitrators “shall have no power to alter or in rendering their award to depart from any express provision of

98 Forward, WSR 1 (1936). Although the quoted material seems to give unwritten custom some legitimacy, the Forward subsequently discusses the proper sources of authority for arbitral decisionmaking, noting that “the salesnote and specifications nearly always are considered together, but not everyone realizes [...] that Trade Customs [as defined in these rules] are the interpretations of them which should govern sellers and buyers in their trading and which, in turn, should guide arbitrators in their decisions.” Id. The rules give the Salesnote precedence over even written custom. WSR 23 (1960) (“In the case of conflict, express or implied, between the Salesnote and the section on Definitions and Trade Customs, the Salesnote governs.”). Later WSRs forbid the arbitrators from looking to anything other than the contract and written custom. See infra, text accompanying note 101. Moreover, even when merchants agree that something is indeed a customary practice, absent a contract clause requiring that the practice be followed, they do not necessarily infer from the fact that something is done that it should be done. See infra, note 215.

99 Forward, WSR 1 (1936).

100 See WSR 3 (1941) (“By common consent and almost universal usage, [the WSRs] have come to be recognized as the standard code of procedure and trade custom applicable to the purchase and sale of cotton textiles and allied lines.”).
this contract, and their failure to observe this limitation shall constitute grounds for vacating their award.”

More generally, the existence of comprehensive codified customs was considered an essential reason for the success of textile arbitration. As the 1952 study noted, “Lack of mutually acceptable codes of trade custom in rayon has in large part been responsible for any dislike of rayon arbitration. Agreement on codified trade custom in cotton, on the other hand, has been a major factor in the success of its arbitration.” And, as a 1919 editorial in Textile World Journal noted, “in the past, sellers, as well as buyers […] have viewed with suspicion such [an arbitration] plan for settling disputes. They have felt that competitors in their line of business might not be in a position to give an unbiased decision on matters involving usage in the trade with which they are connected.”

D. Silk

The Silk Association of America (SAA) had one of the earliest and most well-regarded arbitration systems. Like

101 WSR 3 (1986).
102 Note, loc. cit., note 85, 713.
103 Advantages of Arbitration, Textile World Journal 30 (May 17, 1919). See also Legal Cancellations, Textile World Journal 38 (Dec. 18, 1920) (noting that “[a]t various times when the subject of a standard sales note has been considered, the need of such standards and tolerances has been recognized, but with very few exceptions, none have been developed that could not be punched full of holes by clever lawyers. The usual procedure when a case of this kind comes before the court is for both the plaintiff and defendant to call as witnesses a number of so-called authorities to testify regarding trade custom.”).
104 American Arbitration Association, Year Book on Commercial Arbitration in the United States, 1927 772 (Oxford 1927) (“The Silk Association of America, Inc., is a national organization, established ‘to promote the advancement and prosperity of the silk interests. [Its members include] raw silk importers, dealers and brokers, commission throwsters, manufacturers of swing silks and twists, broad silks, ribbons and hatbands, laces, nets and veilings, knit goods and glove silks; skein dyers, piece dyers; printers and
the textile system, it governed transactions among merchants who played fixed roles in the chain of production and distribution of silk. The arbitration system was created in 1898, but trade rules were not adopted until several years later. The SAA By-Laws noted that the Association aimed, among other things, to “establish and maintain uniformity and certainty in the customs and commercial usages of the silk trade,” and early reports noted that “committees of the Association are intelligently and constantly [...] seeking to improve trade usages, and to substitute better methods as between buyers and sellers.”

The rules creation process at the SAA was even more contentious and drawn out than in other associations. For example, work on the Trade Rules for Raw Silk began in 1901, but after “two years of fruitless discussion of the rules proposed, the committee was discharged at the request of its own chairman.” As an Association report explained, “[t]he get-together spirit was not sufficiently pronounced to override the differences that arose when the rules in detail were considered.” Work on the rules resumed in 1907, and new

---

105 Irving S. Paul, J.W. Millard, and James S. Taylor, Trade Association Activities 112 (Department of Commerce 1927) (describing and lauding the operation of the SAA arbitration tribunal); I.L. Blunt, “American Commercial Arbitration”, (1939) 3 Arbitration J. 299 (describing the SAA as an arbitration “pioneer”).

106 I.L. Blunt, loc. cit., note 105, 300. The SAA also sponsored an Examination Bureau to make quality determinations and an Adjustment Bureau that offered informal mediation services.


109 One reason for the unusual level of tension may have been that buyers associations negotiated with sellers associations so that the distributive impact of the outcomes reached might, along many dimensions, have been quite serious.

110 Efficient Distribution an Aid to Industry, 6 Silkworm 367 (Feb. 1925).

111 SAA, Thirty-fifth Annual Report 23 (1907).

112 Id.
rules were adopted in 1908, after fourteen separate committee meetings, “several conferences between the Raw Silk Division [...] and the Board of Managers” to address “difficulties,” and the “approval of some amendments.” In approving the rules, the Board of Managers emphasized, as did the preambles to the rules themselves, that they were merely default rules.

The creation of the Silk Throwsters Rules in 1907 after a year of “active group work,” was also contentious.

114 Id., at 36. These Rules were amended in 1912, SAA, Fortyeth Annual Report 31 (1912), in 1921 after “a year of careful study on the part of the committee,” Revision, 3 Silkworm 73 (May, 1921), and again in 1924 after the “culmination of many months of intensive effort on the part of a committee of representative buyers and sellers to revise the rules in such a way as to cover changes in trade customs which had developed since [...] 1921,” including changes in credit terms and “provisions for claim, rejections and replacements which were rephrased so as to conform with customs which had developed in the trade and to cover those points which had not, perhaps, been clearly understood.” SAA, Fifty-third Annual Report 26 (1925).
115 SAA, Thirty-Sixth Annual Report 83 (1908); Raw Silk Rules and Regulations, id., at 77 (“[N]othing in the following rules shall be construed as waiving the right in individual transactions to make any special or distinct contrary agreement, but that the rules shall govern only in cases where no special or specific contract exists.”).
116 These rules were quite detailed. They defined acceptable tolerances, manufacturing techniques, various allocations of liability, payment, and transport. They also noted that “the amount of loss to be allowed in the actual working of a given silk (to be arrived at as above stated) is universally a matter of agreement between the manufacturer and the throwster.” SAA, Thirty-Sixth Annual Report 32 (1908).
117 Trade Practices, 2 Silkworm 28 (Feb. 1921).
118 The creation of other sets of rules was somewhat less contentious. The Rules Governing Broad Silks, “which embody the best sentiments of prominent manufacturers in this leading branch of the industry,” were drafted in 1912-13. SAA, Forty-First Annual Report 45 (1913). The effort to adopt them, however, began in 1907 and “was continued intermittently for a number of years until, by virtue of the most dogged efforts and persistency, a set of trade practices was adopted and approved.” Trade Practices, 2 Silkworm 28 (Feb. 1921). The Rules Governing Spun Silk were drafted during 1921-22 and adopted in 1923 with little fanfare. SAA, 51st Annual Report 56 (1923). They were amended in 1928 in a process described as “translating customary trade practice into uniform trade rules.” SAA, Fifty-Sixth Annual Report 16 (1928). In addition, “Rules Governing the Commission Throwing of Silk” were adopted in 1927 as part of a renewed attempt to achieve “advancement in the
but proceeded relatively quickly, perhaps because the group adopted, “practically the same rules which obtain in Europe.”\footnote{SAA, 55th Annual Report 17 (1927).} The goal of the codification movement was to “[get] business conditions on a more uniform basis and [formulate] a standard set of trade rules, which would be acceptable to both manufacturer and throwster.”\footnote{SAA, Thirty-Fifth Annual Report 55 (1907). See also SAA, Thirty-Sixth Annual Report 30 (1908) (amending the Throwsters’ Rules to include a “standard weight” provision to eliminate “the former uncertainty and divergence in the kind of weight charged for by throwsters.”).}

The creation of Rules for Thrown Silk was also a controversial process. After several years of committee work, the first draft of the rules was proposed in 1923, but rejected by the Association. This raised the ire of their author, who stormed:

*The Thrown Silk Division of the Association has been at work with a committee appointed by the association for a number of months past on rules to regulate the sale of thrown silk in all its branches, and after a great deal of time and labor, their recommendations were submitted to the Board of Managers of the Association. These recommendations were based on current market practice, and it is to be regretted that the Board did not see fit to accept them as submitted, and thus place an important branch of the silk industry on a sound economic basis.*\footnote{SAA, Fifty-First Annual Report 55 (1923).}

The rules were eventually adopted in 1924, “the result of four years’ endeavor to compile the list of best trade practices [...] based on suggestions and experiences from both buyers and sellers,”\footnote{SAA, Fifty-Second Annual Report 23 (1924).} and were said to “fill a long felt need.”\footnote{Id.} In the late 1920’s the Association also began to promulgate standard form

standardization of trade practices through the adoption of rules and forms.”

\footnote{Id.}
contracts in an effort to bring about even greater uniformity in trading terms. The various rules adopted by the SAA were amended numerous times during the 1920’s and 1930’s, ostensibly to respond to changes in trade practices.\footnote{See supra, notes 118 and 120; SAA, 57th Annual Report 22 (1929) (“Both revisions of finished goods rules and compilation of raw goods rules have been based on market practices and customs, generally accepted in the silk industry.”). Some of the changes introduced in the 1930s might have been responses to the National Industrial Recovery Act and government initiatives related to it.}

The contentiousness surrounding the adoption of silk trade rules is in some respects surprising. There was much more agreement about the content of existing practices than in other industries, perhaps because of the somewhat smaller number of individuals engaged in the trade. However, unlike merchants in many industries who often venerated existing practices, silk transactors frequently criticized existing practices as being out of date, inefficient, or just plain stupid.\footnote{SAA, Thirty-Fifth Annual Report 57 (1907) (“[W]hy this stupid custom [of \textquoteleft another bill and the old price\textquoteright] should continue, no satisfactory explanation can be given. There is no more reason for this, than for the equally foolish custom that prevailed for many years in giving the jobber \textquoteleft protection in case of a decline in price.\textquoteright”); SAA, XLIII Annual Report 41 (1915) (noting that tie silk manufacturers would be better off if it were \textquoteleft possible to induce the trade to break away from their fixed traditions.\textquoteright). See also SAA, Third Annual Report 64-65 (1875) (“It is strange how far custom will lead us, when our better judgment would not go a step. This may be especially the case, in regard to the custom of charging single ounces on a bale of silk. Not a merchant, not a manufacturer, not a throwster, but knows that silk is very sensitive of atmospheric changes, that bale will vary by weight a few ounces if weighted twice in one day, and yet they all sanction this custom by tacitly conforming to it without objection.”); the Ribbons Report, in SAA, Forty-Fourth Annual Report 65 (1916) (harshly condemning the practice of the \textquoteleft the open order, with no definite date of assortment,” attributing the practice to \textquoteleft the custom of the trade, as constituted at present [which] did not compel [buyers] to assort the goods until they were needed, which might be at any future time, if at all”); SAA, XXXI Annual Report 38 (1903) (“It has often been noted that it is the custom with most manufacturers to continue their output for the Fall consumption on the same basis as the Spring, when it is known, without contradiction, that the demand for the last part of the year is much less than for the first half.”).} Like the grain dealers, they lamented the difficulty of changing existing practices. As one SAA banquet speaker put it, \textquoteleft[\textquoteleft uge question marks are boldly scribed on the walls of custom, hoary with
Modifications of trade customs are exceedingly difficult to introduce. Trade practices are handed on from generation to generation of office administrators and when a business routine is established in an organization it often requires a reorganization, a so-called ’shake-up,’ to introduce innovation. Considering this conservatism the [success of the SAA in modifying practices] may be looked upon justly as a real accomplishment of the Association.

The hostility toward custom in this industry might have been due, at least in part, to the fact that many customs were perceived to be, and might well have been, “customs of adhesion,” or attempts of particular branches of the industry to secure competitive advantage.

Although customs in the silk industry were better established than customs in other industries, there was still much disagreement about the meaning of basic trade terms. For example, at the 1915 SAA Convention, one speaker pointed out that there was no common understanding of the term “Double Extra” silk:

There is not a man here who could give a definition. Some of you men, however, can tell us what N° I Buckwheat coal is, and for the simple reason you have a standard to go by, and certain tests that N° I Buckwheat must live up to. When we have a standard for Double Extra silk, and certain tests that a Double Extra must

---

126 SAA, XLII Annual Report 145 (1914). See also SAA, Thirty-Sixth Annual Report 21-22 (1908) ("The silk interests have come to one absolute conclusion, and that is that all past methods that have not proven sound and reliable need to be revised.").

127 SAA, 46th Annual Report 82 (1918). See also SAA, Forty-Fourth Annual Report 57-58 (1916). (Where the Tie Silk Association boasted that progress had been made in achieving change in the face of powerful customs; noting that “[w]hat had been hitherto considered impossible for the neckware trade, that is to pay more than 80 to 82 1/2 cents for their 50 cent line, was accomplished. Old traditions were brushed aside.").
meet, then will you be able to define a Double Extra. The present Double Extra silk is any one man’s opinion.128

In addition, numerous other practices, particularly those relating to grade designations, were quite varied in different geographic areas.129

E. Methodological Problems

A serious difficulty raised by drawing on the trade association codification debates to explore whether industry-wide, unwritten customs existed is that the data are also consistent with the possibility that industry-wide customs did exist, but that the codification debates provided an opportunity for rent-seeking subgroups to fabricate disputes about the content of custom130 and to lobby for the recognition of the

128 See Proceedings of the First National Silk Convention 136 (1915) (“First Silk Convention”). More notably, in 1923 a Committee of the British Silk Association wrote a report attempting to define the word silk. See Proper Use of the Word “Silk” Defined by Committee of British Association, 5 Silkworm 176 (July 1923) (“Silk means the natural product of the silkworm, whether Net Silk or Spun Silk” and should not, with the exception of “[c]ertain smallwares containing Silk in combination with other fibres [...] [that] have been by long established custom known as Silks,” be “used for combined fabric products.”).

129 See, for example, First Silk Convention, loc. cit., note 128, art. 119. (“There exists in raw silks a Japan, Italian, China, French, Turkish and Canton classifications, which are all unlike [...] Taking everything into consideration, the six classifications have no resemblance to each other at all.”); James CHITTICK, Silk Manufacturing and its Problems 25 (James Chittick 1913) (“Yokohama classifications do not correspond with those of New York, the N° 1 of Japan being called here Best N° 1”). Moreover, the way quality designations were defined tended to change from year to year as crop conditions changed. See, for example, First Silk Convention at 119 (attributing the “fluctuation of classifications,” to the fact that “after a rainy season, when all the qualities of silk are poorer, the same classifications are retained so that an Extra in 1915 is different quality from an Extra in 1914”). This type of variation would create significant problems for a court applying the Code because cases typically go to trial several years after a dispute arises, so reconstructing the content of time-dependant industry customs of this type would be quite difficult.

130 Conversely, statements that custom exists must also be interpreted cautiously. The Report of the 1909 convention of the NHA, for example contains a discussion in which one transactor asserted the existence of a custom relating
trade practices most favorable to them.\footnote{Indeed, in some industries, such as silk, this might have been a significant cause of disagreement. However, even if rent-seeking did account for much of the disagreement over the content of customs, the existence of such disagreements would have serious consequences for the incorporation strategy. If rent-seeking attempts could not be distinguished from genuine differences in practice at association hearings, it is highly unlikely that an arbitration panel or court could make such a distinction in an actual case where the parties also have strong financial incentives to lie.} It is difficult to negate this possibility entirely.\footnote{The actual opportunity to rent-seek likely varied widely across industries due to differences in their voting rules, committee structures, and committee compositions (elected or appointed), as well as the order of their annual meeting agendas.} However, even if rent-seeking does

to a shipment term and another member nonetheless suggested that they “write it out.” NHA, Report of the Sixteenth Annual Convention 216 (1909). As they attempted to formulate a rule on the matter they tried to answer a series of hypotheticals until it became clear that the proponent of the custom could not defend it as providing the relevant answers. \textit{Id.}, at 216-217. Similarly, at the 1999 Texas Grain and Feed Association conference, a speaker on the arbitration process noted, “It turns out [with respect to customs of the trade that] people who think they agreed have very different ideas when it is applied to specific cases.”

It is also possible that, to the extent that participants in the codification debates received prestige or improved business contacts from actively participating in the committees of the national trade association in their industry, they may have had an incentive to exaggerate the lack of uniformity in trade practices in order to enhance the perceived importance of the Association and its codification efforts. Although this explanation might account for some of the statements made by high-ranking association founders in the early years, it is unlikely to be a complete explanation. The codification of trade rules was considered an important association function. However, the formation of national associations was largely motivated by, and derived its prestige from, associations’ efforts to lobby the Federal government for railroad tariff reductions, agricultural subsidies, and a uniform bill of lading law. In addition, many codification debate participants had very different incentives from association executives. For example, much of the impetus for uniform written rules came from merchant arbitrators confronted with the need to decide cases. The prestige of the title “arbitrator” and the perceived legitimacy of these tribunals depended on the arbitrators resting their decisions on reasons that industry members would regard as proper, yet the arbitrators consistently lamented in their opinions and annual committee reports the lack of uniform customs that could be used to resolve disputes. Although it might be argued that this gave them an incentive to lobby for the adoption of rules, the position of arbitrator generally lasted only for a year or two, so it is unlikely that they would have an incentive to undermine the legitimacy of the tribunal today for a benefit that would accrue, if at all, only to those appointed to decide cases in the future.

131

132
account for the lack of consensus as to the content of some trade practices in certain industries, it is unlikely to account for all or even most incidences of disagreement.\textsuperscript{133} In addition, because all industry trade rules are default provisions that can be varied by contract, and because in many industries proposed rules had to be approved by the membership at large, with each firm getting one vote regardless of its size, the amount of rent that a group could capture through the rules-creation process, relative to the rent it could capture through the contracting process and/or the custom-creation process, is not so large.\textsuperscript{134}

There are a number of additional reasons why rent seeking is unlikely to account for most disagreement about the content of customary practices. First, in industries that were composed of numerous local associations and exchanges, before the formation of a national association, the written rules adopted by these small associations and exchanges differed widely. This gives some assurance that the debates relating to differences in customary practices reflected at least some real differences in the way that business was done in different parts of the country, rather than mere fabricated differences between rent-seeking industry subgroups. Second, the 1952 textile study found widespread disagreements over the meaning of usages of trade in a context where neither amending the rules nor deciding an actual case was at stake.\textsuperscript{135} The existence of similar disagreements in other industries such as grain, lumber, and silk are confirmed by articles in their trade press and/or similar surveys.\textsuperscript{136} Third, many of the codification efforts were

\textsuperscript{133} In addition, if rent-seeking was part of the process of codifying national rules, it might also have been part of the local rules-creation process and the process of custom-capture. It is therefore yet another force that works against rather than for the emergence of strong-form, pure Hayekian customs. See notes 174-175 and accompanying text.

\textsuperscript{134} Some trade rules had to be separately approved by buyers and sellers associations. See, for example, ATMI and ACSA, \textit{Southern Mill Rules}, \textit{loc. cit.}, note 27, 32. The WSRs were also separately approved by groups of buyers and sellers. See WSR 2 (1936) (listing sponsoring organizations).

\textsuperscript{135} See Note, \textit{loc. cit.}, note 85, 700.

\textsuperscript{136} For example, lumber associations conducted and published a number of surveys of trade practices that were not in any way related to the rules-creation or amendment processes. These surveys found that industry members did not generally agree on the content of business practices. See, for example, NAWLA, \textit{Adjusting Salesmen’s Commissions}, North Coast Weekly Letter N°
motivated by a perceived need to make customs more uniform if misunderstandings were to be avoided and arbitration was to become an effective way to resolve disputes. If, however, the geographical scope of customs was co-extensive with the geographical scope of efficient trade, and if these customs were, in fact, generally known and implicitly assented to, codification should not have been necessary to achieve predictable arbitral outcomes and avoid commercial misunderstanding.137 Fourth,

192 (Sept. 29, 1925) (Reporting the results of a fifty person survey—asking “Where a commission salesman is paid so much per M on his sales and the customer claims a shortage of say, a few thousand feet, is it common practice to deduct the amount of the commission to be paid the salesman where an allowance is made covering such a shortage? Where a commission man is paid on a percentage basis, say five percent, and a claim for any reason allowed to customer, is a deduction, based on the allowance, made from commission paid?”— and noting that “[t]he majority (35) state they do not think it is customary or good business policy to revise the commission where shortages or grade complaints develop [...]. Fifteen members stated they believed the salesman should stand his pro-rata of the deduction in both cases [...]. The majority who answered the second question stated that where salesman is working on a percentage of the profits, the salesman should stand his share of any losses”; also reporting six statements of respondents setting out further nuances:); NAWLA, Who Collects Rate Overcharges?, North Coast Weekly Letter No. 193 (Oct. 6, 1925) (noting similar divergence of opinion, correlated with the size of the firm, to the question: “When stock is bought f.o.b. mill basis on guaranteed weight, and the transportation company makes an overcharge in rate, is it up to the buyer to collect from the railroad, or should he charge the mill and let the mill collect claim?”). See also, infra, note 221 (discussing silk surveys).

137 If the type of consensus about and widespread knowledge of the content of trade usages that is assumed by the Code is thought to consistently emerge in mercantile contexts, it is necessary to explain why, from at least the middle ages to the present, there have been so many attempts to codify mercantile customs. Bruce L. BENSON, “The Spontaneous Evolution of Commercial Law”, (1989) 55 S. Econ. J. 644, 649. (“[A]s the norms of [medieval merchant] commercial law became more precisely specified they were increasingly recorded [...] not [in] statutory codes [...] [but in] written commercial instruments and contracts.”); Gerard MALYNES, I Consuetudo, vel Lex Mercatoria : or the Ancient Law-Merchant a, a2 (“To the Courteous Reader”) (T.:Baffet 1685, Professional Books reprint 1981) (compiling, in meticulous detail, trade practices and laws “[f]or the maintenance of Traffick and Commerce [which] is so pleasant, amiable and acceptable unto all Princes and Potentates, that Kings have been and at this day are of the Society of Merchants!: And many times, notwithstanding their particular differences and quarrels, they do nevertheless agree in this course of Trade [...]. Wherupon I have been moved, by long observation, to put the worthiness of the Customary Law of Merchants, in plain and compendious writing, by undoubted principles, familiar examples and demonstrative reasons”); A
THE QUESTIONABLE EMPIRICAL BASIS OF ARTICLE 2’S

Member of the Massachusetts Bar, The Business Guide and Legal Companion 59, 60 (no publisher listed 1845) (defining, like many commercial dictionaries, words like “hogshead,” “barley corns,” and “pennyweights”); John R. MCCULLOCH, A Dictionary, Practical, Theoretical, and Historical, of Commerce and Commercial Navigation, Philadelphia, Thomas Wardle, 1841, pp. 720 and 724 (providing tables of weights and measures, and defining words, and recording customs such as the custom “of allowing more than 16 ounces to the pound of butter [which] used to be very general in several parts of the country”).

Today, the International Chamber of Commerce (“ICC”), a major provider of international arbitration services, has promulgated several quasi-official codifications of customs. Most important are the Incoterms, “a set of international rules for the interpretation of the chief terms used in foreign trade contracts, for the optimal use of businessmen who prefer the certainty of uniform international rules.” ICC, Incoterms: International rules for the interpretation of trade terms 6 (ICC 1980). These rules acknowledge the undesirability of interpreting agreements by reference to custom, explaining that “[e]very endeavor has been made to limit such references to custom to the absolute minimum.” Id., at 8. However, noting that it has been “impossible to avoid [references to custom] altogether,” the rules advise the “seller and buyer [...] to keep such general and particular customs in mind when negotiating their contract.” Id. The Incoterms implicitly recognize that customs may be highly local in nature by including a choice-of-custom provision, which provides that customary matters are to be “decided by the custom of the particular trade or port.” Id. To further reduce uncertainty, the ICC has published several more particularized complications of customs. See, for example, ICC, Uniform Customs and Practice for Documentary Credits, ICC Pub N° 500 (1993); ICC, Uniform Rules for Collections, ICC Pub N° 522 (1995); ICC, Uniform Rules for Contract Guarantees, ICC Pub N° 325 (1978).

However, it is important not to attach too much weight to the possibility that the perceived need to write customs down suggests that customs are not sufficiently definite or uniform. There are a variety of other reasons, particularly in the trade association context, that the codification of custom might have been undertaken. First, when the relatively new national associations decided to adopt rules, they may have enhanced the legitimacy of the rules and reduced suspicion about the drafter’s motivations by claiming that these rules were mere codifications of existing practices. Second, because most trade association arbitrators were industry participants, codification might have been viewed as useful for constraining some types of arbitral bias, or making such bias easier to detect, thereby reducing arbitral discretion and enhancing the legitimacy and perceived fairness of the tribunals. Maintaining the tribunals’ appearance of fairness was especially important prior to 1920, when both ex ante agreements to arbitrate and arbitral awards were legally unenforceable. Third, because most arbitrators were not lawyers, codified rules enabled them to decide cases without having to familiarize themselves with the intricacies of sales law. Fourth, codification might have been viewed as necessary because, as the pace of technological and market changes increased, practices needed to change more quickly and in more coordinated
contemporary interview evidence from the feed trade in Texas suggests that even within the state, trade practices regarding core contractual terms vary widely.\textsuperscript{138} Fifth, at associations like the NHA, where proposed rules were subject to floor debate at the Association’s annual convention,\textsuperscript{139} and debates were reported verbatim in the Association’s Annual Report, there were publicity-based reputational constraints on debate participants. A transactor who suggested something was a custom when it was not risked reputational harm. If, for example, there was a custom that it was reasonable to take ten days to inspect goods before rejecting them, and a buyer stood up at the convention and said that it was reasonable to take twenty days, sellers would be wary of dealing with him in the future.\textsuperscript{140} If, however, the buyer could be sure that all other buyers would get up immediately and say twenty days, he might well engage in this type of distributional bargaining. In industries like silk, where the rules were the outcome of bargaining among groups of buyers and sellers who each deliberated in secret and then held a meeting between designated representatives of each group to negotiate the final rules, reputational constraints would have been much weaker, and the silk debates do indeed contain many instances of what might have been rent-seeking behavior. Sixth, some of the terms about whose meaning transactors disagreed have no distributive impact. For example, outside of the context of a specific dispute, a decision on how many days are meant by the terms

\begin{footnotesize}
\begin{enumerate}
\item[	extsuperscript{138}] Interview with TGFA executive (1998) (on file with author); Interview with TGFA Board Member (1998) (on file with author); Discussant at TGFA Meeting (Feb. 18, 1999) (explaining that the TGFA “Rules Committee didn’t add certain truck rules because there were many different customs,” but noting that this was “okay because trucks tend to travel in a 150 mile radius and people know how things are done in their area.”).
\item[	extsuperscript{139}] In the early days of national trade associations, one of the main purposes of the annual convention was to form relationships and learn about the reputations of transactors in other localities. As a consequence, transactors attending these meetings were quite conscious of guarding their reputation.
\item[	extsuperscript{140}] Similarly, in the grain industry where many merchants were buyers one day and sellers the next, they would have had no reason to skew the rules toward buyers or sellers.
\end{enumerate}
\end{footnotesize}
“prompt” or “quick” does not have distributional implications, because most trade rules contained a menu of time designations to choose from. Seventh, in industries whose trade journals had columns where members could pose hypothetical questions to association executives, members often asked “what is the custom with regard to X.”

Finally, the testimony of representatives of merchant associations in the hearings on the proposed Code provides support for the notion that customs did not exist and reveals that merchants were consciously aware of, and opposed to, the Code’s incorporation strategy. The opinions they expressed were remarkably consistent with the arguments made during the intra-association codification debates, most of which had taken place many years earlier. For example, the Commerce and Industry Association, a trade organization of merchants, objected to including “observance of reasonable commercial standards” in the Code’s good faith provision. The Association explained that “the usages, customs and practices of business are far from being uniform, and the determination of whether a merchant has conformed to reasonable commercial standards would be difficult and would produce excessive litigation.” These views were also echoed by the New York

---

141 See, for example, Questions and Answers, 8 Silkworm 240 (Oct. 1926) (“A piece goods manufacturer inquires if it is customary in the trade for the manufacturer to return the caps to the throwster.”); Questions and Answers, 8 Silkworm 300 (Dec. 1926) (“Is it customary to include in a contract with commission weavers a clause to the effect that the house must use only such quantities of yarns as are in accordance with the standards and rules of the Association.”).

142 Formerly the Merchants Association of New York.

143 U.C.C. § 1-203 & cmt.

144 Commerce and Industry Association of New York, Inc. (“CIANY”), Memorandum of Task Group of the Special Committee of the Commerce and Industry Association of New York, on the Uniform Commercial Code on Article 2, Sales and Article 6, Bulk Transfers, in Study of U.C.C. Memoranda 24, 29, reprinted in State of New York, 1 Report of the Law Revision Commissioner for 1954, loc. cit., note 13, 88-93. In addition, during the hearings, Professor John O. Honnold posed some questions to representatives of the Merchants Association in an attempt to see if proposed Code provisions conformed to trade understanding of particular terms. In most of their responses, association representatives stated that practice was not “exactly uniform.” For example, when asked about the meaning of “F.O.B. vessel,” a representative noted, “while it is trade understanding in certain fields that the
Thus, while Llewellyn justified the incorporation strategy on the grounds

buyer must arrange for steamer space [...] in practical operation buyers customarily depend upon the sellers to make arrangements and ship via the most suitable and expeditious vessel available at the time the goods are ready for export.” CIANY, Memorandum Requiring to Questions Propounded by Professor Honnold at Law Revision Commission Hearing 1-D, in U.C.C. Memoranda, at 67, reprinted in id., at 131. In response to questions about the meaning of other proposed terms, the merchants said,

“there appears to be a diversity of opinion as to inspection right where the contract terms are ‘F.O.B. vessel’ or ‘F.A.S.’ While the view has been expressed in a particular industry that the purchaser has the right to inspect goods before payment under such terms, the general export practice apparently is that the buyer has no right of inspection before payment unless there is specific prior agreement that such inspection or examination is required by the buyer. It is recommended practice that such prior agreement also specify whose account is to be charged for the cost of such inspection since such cost may be appreciable when the charges for inspection engineers, opening and closing constraints, and making goods available for examination are considered.”

Id. The Association also objected to part of the proposed Code’s cure provision that gives the seller additional time to cure if he “had reasonable grounds to believe [the non-conforming tender] would be acceptable,” U.C.C. § 2-508(2), where the reasonableness of grounds is determined, in part, by reference to “prior course of dealing, course of performance, or usage of trade,” id., at cmt 2. In its criticism of the provision, the Association explained, “There is no reason why a seller should ever believe that a non-conforming tender would be acceptable to the buyer.” CIANY, Memorandum of Task Group, in U.C.C. Memoranda at 37, reprinted in id., at 101.

The New York Law Revision Commission

“was a little dubious about the widespread use of business terminology throughout the Code. Not that it has anything against business terminology as such—it is often very convenient—but it often lacks precision. It often means something in one part of the country, something else in another part, or as between different industries or lines of business[...]. [T]he Commission is anxious to see a statute that would have an ascertained or ascertainable meaning.”

Panel Discussion on the Uniform Commercial Code, (1956) 12 Bus. Law 49, 57. The Commission also expressed doubts about the anticipatory repudiation rules, explaining that “reasonable grounds for insecurity, adequate assurance of due performance [and] commercial standards,” were terms that “should be more concrete, more specific. It did not like this resort to rather vague, uncertain, indefinite language.” Id., at 56-57. Even greater skepticism about the incorporation of trade usage was expressed in connection with the United Nations Convention on Contracts for the International Sale of Goods. See Cesere M. BIANCA and Michael J. BONELL, Commentary on the
that it would make merchant law more accommodating to merchant concerns, the approach was not viewed in a positive light by the very merchants whose transactions it was designed to govern.

F. Conclusion

Although the evidence presented here has not conclusively demonstrated that the types of usages of trade and commercial standards, and industry-specific meanings of terms, referenced in the Code do not ever exist, it has suggested that the empirical foundation on which the Code in general, and its incorporation strategy in particular, is built, may be weak. In addition, as the next Section suggests, the lack of consensus on the content of customary practices documented in the case studies, is in no way surprising since there are a number of reasons to suspect that uniform customs relating to many aspects of a transaction would not exist. As a consequence, it is useful to reconsider the justifications for the Code’s incorporation strategy and to look briefly at the ways that it has been invoked by parties and used by courts in decided cases.

II. The Incorporation Strategy Revisited

Code drafters and later commentators justified the pervasive incorporation strategy by noting that contracts are incomplete because rationality is bounded; because the transaction and strategic costs of including provisions dealing with all contingencies, particularly remote or unforeseen contingencies, are high; and because it is difficult to reduce the parameters of some obligations to a writing. They further maintained that customs are intended and understood by merchants to be an integral part of their agreement\textsuperscript{146} and that these customs, along with the merchants’ course of actual performance under their contracts, provide the best indication of

\textsuperscript{146} See U.C.C. § 2-202 cmt 2.
what they intend their writing to mean. In addition, at least with respect to using custom to fill gaps, the incorporation strategy has also been defended on the grounds that the provisions provided by custom are likely to be more efficient than those that a court could provide. A closer look at these justifications, however, suggests that they are based on inaccurate empirical assumptions and are theoretically weak.

A review of a subset of Code cases, from 1970 to the present, invoking the usage of trade provision, reveals that very few alleged customs relate to remote or unforeseen contingencies. Most of the alleged customs relate to core aspects of any merchant transaction such as: the permitted

147 See id., at § 2-208 cmt 1 (“[P]arties themselves know best what they have meant by their words of agreement and their action under that agreement is the best indication of what that meaning was.”).

148 See id., at § 1-205. The sample includes cases in Uniform Commercial Code Case Digest Para 1205.2(1) (West 1998), under the heading “Usage of Trade—In General” (excluding supplement). Eliminated were cases dealing with the secured financing aspect of the transaction; one case where the concept of a usage was simply contrasted with the concept of course of dealing, see Capitol Converting Equipment Inc v. Lep Transport, Inc, 750 F. Supp. 862, (ND Ill 1990), affd, 965 F2d 391 (7th Cir 1992) (where the course of dealing or usage dealt with terms of payment); and one case dealing with whether a contract existed that held the existence or nonexistence of any usages was irrelevant, see Wichita Sheet Metal Supply, Inc v. Dhalstrom and Ferrell Construction Co, 246 Kan 557, 792 P2d 1043, (1990). Although the rather small number of cases in the digest category “Usage of Trade—In General” has led some to suggest that the Code’s incorporation strategy is, as a practical matter, unimportant, this ignores the numerous other Code provisions and comments that incorporate usage of trade. See supra, note 1.

149 But see Cosden Oil & Chemical Co. v. Karl O. Helm Aktiengesellschaft, 736 F.2d 1064, 1076 (5th Cir. 1984) (considering whether “a custom or trade usage relating to force majeure existed in the polystyrene industry.”).

150 Most industry trade codes either require these aspects of a transaction to be covered by contract and/or provide bright-line gap-filling default rules relating to them. See, for example, Trade Rules of the National Hay Association, Rule 1, in NHA, Report of the Twenty-Eighth Annual Convention 83 (1921) (“It shall be the duty of both buyer and seller to include in their original articles of trade, however conducted, the following specifications: Number of cars, tons, or bales. Size of bales. Grade of hay or straw. Price. Terms of payment. Rate basing point. Time of shipment. Route.”); Grain Trade Rules (as amended Mar 13, 1995), Rule 1, in NGFA, Trading Rules and Arbitration Rules 3 (NGFA 1995) (noting that it is the duty of the buyer and seller to include in their contract or confirmation, the “(a) Date of contract; (b) Quantity; (c) Kind and grade of grain; (d) Price; (e) Type of
tolerance in quality; the definition of the good to be exchanged; the time required for notice; the time for inspection and rejection; the time, place, and manner of delivery; price and/or quantity adjustments; payment terms; and limitations on remedies. Nevertheless, even if

inspection[;] (f) Type of weights[;] (g) Applicable Trade Rules to apply[;] (h) Transportation Specifications:
(1) Type of conveyance; (2) Type of billing[;]
(3) Port of origin or delivery; or rate basing point[;]
(4) Loading weight requirements[;] (5) Time of shipment or delivery[;]
(6) Route[;] (7) Responsibility for freight increases or decreases[;]
(8) Buyer’s and Seller’s conveyance[;] (9) Type of bill of lading[;] (i) Payment terms; (j) Other terms”.

(where the alleged custom related to quality verification usually undertaken before a purchase is final).

152 Williams v. Curtin, 807 F.2d 1046, 1049 (D.C. Cir. 1986) (where the alleged usage related to the meaning of the term “slaw cabbage” in a situation where one party contended it only meant “large cabbage” while the other contended that it meant “all cabbage suitable for making coleslaw”); Latex Glove Co. v. Gruen, 146 Ill App. 3d 868, 497 N.E.2d 470 (1986) (where the alleged usage went to the definition of good to be sold, specifically whether the sale of printed material included the sale of related by-products of the printing process).

153 Lackawanna Leather Co. v. Martin & Stewart, Ltd, 730 F.2d 1197 (8th 1984)
(noting that the appellant sought to introduce usage defining what constitutes a reasonable time for inspection and revocation).


155 See Southern Concrete Services, Inc. v. Mableton Contractors Inc., 407 F Supp 581 (N.D. Ga. 1975), affd. 569 F.2d 1154 (5th Cir. 1978) (although contract included price and quantity provisions, one party claimed a trade usage that these were to be treated as “mere estimates.”); Nanakuli Paving and Rock Co. v. Shell Oil Co., 664 F2d 772 (9th Cir. 1981) (alleged custom related to the price).

156 Ore & Chemical Corp., 455 F. Supp. at 1152 (where a battle of the forms issue turned on whether or not provisions in the plaintiff’s confirmation dealing with “payment, loading Incoterms [Paris plus supplement]” were usages of trade); Union Building Materials Corp. v. Heus and Haynie Corp., 577 F.2d 568, 571 (9th Cir. 1978) (where the alleged usage related to the terms of payment, specifically “whether the payment made to the
courts looked to custom primarily to fill gaps when remote contingencies arose, this is precisely the type of situation in which incorporation is likely to yield the least desirable results. The more remote or unforeseeable a contingency, the less likely it is that there is an established customary practice covering it. Moreover, even if such a custom did exist, one would want to subject it to particularly exacting scrutiny. Assume for a moment, as the Code does, that customs are Hayekian in nature, gradually emerging over time from transactors’ independent choices. Now consider how a pair of transactors will decide to act when a remote contingency arises. In such a situation, each transactor will reason that the likelihood that this situation will arise again is, by definition, remote, so that even if she is a buyer one day and a seller the next, she should fight for the outcome that gives her the greatest share of the pie today, rather than the outcome that will, over time, maximize the size of the pie. As a consequence, to the extent that patterns of practices dealing with remote contingencies arise, they may favor stronger or particular types of firms.

157 Western Ind., Inc. v. Newcor Canada Ltd, 739 F.2d 1198, 1201 (7th Cir. 1984) (discussing an alleged “custom of the specialty welding machine trade not to give a disappointed buyer his consequential damages but just to allow him either to return the machines and get his money back or (for example if the breach consists in delivering them late) keep the machines and get the purchase price reduced to compensate for the costs of delay”); Posttape Associates v. Eastman Kodak Co., 450 F. Supp. 407, 410 (E. D. Penn 1978) (discussing “a trade usage limiting a commercial buyers remedy to replacement of the negligently manufactured film”).

158 See notes 174-175 and accompanying text (discussing Hayek’s views on the evolution of custom).

159 If, however, the happening of the contingency and the way it is dealt with by these transactors are observable to members of the relevant market, a transactor concerned about her reputation might act in the most advantageous way that will be regarded as “fair.” Alternatively, a single very general custom such as “share the loss” might evolve to deal with large groups of unforeseen contingencies, but again there is no reason to think that this allocation of losses will create desirable incentives. Nevertheless, unforeseen contingencies, at least in domestic merchant industries, are quite rare since most trade rules cover adverse weather, labor, transportation, and financial events, making most remaining unforeseen contingencies quite remote.
The review of these Code cases also suggests that the limits of language were unlikely to have been a significant barrier to transactors’ memorializing the asserted usages in written provisions. In none of the cases surveyed did either party have any difficulty expressing the content of the usage in clear terms. Moreover, the codification debates contain no indication that there were rules that drafters would have liked to include, but that could not be adequately expressed in words. In addition, when the lumber industry newsletter reported

160 It is, however, important to note that the cost of capturing a usage in writing *ex ante* is likely to be higher than the cost of giving it content *ex post,* and the cost of giving it content *ex post* need only be incurred if the relationship breaks down and the parties disagree about its content. In addition, an *ex ante* provision memorializing a usage might have to be more complex than an articulation of a custom *ex post* since it might have to cover numerous preconditions and situations that would be irrelevant *ex post.* If, however, complexity and nuance are to be invoked as reasons not to memorialize the usage in a contract, but to nevertheless incorporate it as an enforceable part of an agreement, it is important to note that as the complexity of and number of contextual preconditions to a usage increase, the likelihood that there is reasonably widespread consensus about its content, and that its content can be accurately determined by a court, decreases. In addition, there are unrecognized costs of not capturing the usage in writing *ex ante,* such as a greater risk of misunderstanding and subsequent transaction breakdown, as well as the costs of the overly-high or overly-low precautions that the transactors might take if they do not clearly understand their obligations.

161 The only possible exceptions are the debates relating to grades. While the limits of language were not invoked as reasons for the inability to arrive at a consensus over how to memorialize the description of a particular grade in a rule, and most associations either promulgated standardized written descriptions of quality—see, for example, *Grades of Hay and Straw Established by the National Hay Association, Inc.,* in NHA, *Report of the Twenty-Eighth Annual Convention* 161 (1921) (“N° 1 Timothy hay—Shall be timothy containing not more than one-eighth clover or other tame grasses, bright color, sweet, sound and well baled.”)—or relied on descriptions provided by the government—see *Official United States Standards for Grain,* 7 CFR Part 810 (1998)—the fact that rules expressed in words could not ensure uniform grading given the subjectivity of the determinations that had to be made was an implicit theme of the debates. Nevertheless, this subjectivity did not lead the associations to direct arbitrators to incorporate custom on the question of quality. Rather, most trade rules provided for quality disputes to be resolved by wiseman-like intermediaries, some run by
arbitration decisions in cases arising from a disagreement about “custom of the trade,” the case report would frequently be accompanied by a suggested contract clause that could be used to avoid misunderstandings in the future. These clauses did not typically include linguistically complex formulations.162

the industry, as in the green coffee industry—see for example, Green Coffee Association of New York City, Inc., Rules of Arbitration (1989); ATMI and ACSA, Southern Mill Rules, Rules 43-44, supra, note 27, 18—and some run by the government, as in the grain industry by the Federal Grain Inspection Service. For an overview of the program, see Tuttle v. Missouri Dept. of Agriculture, 1999 U.S. App. LEXIS 5445.

It is, however, important to note that, wholly apart from the inability to fully capture grade categories in words, there were a number of other possible reasons for using separate quality intermediaries. First, to the extent that quality determinations are more subjective than rulings on other aspects of contractual performance, the use of a separate quality intermediary might be designed to guard against arbitral bias. Most rules governing industry-run quality arbitration tribunals contain procedural safeguards against bias that are not used in regular arbitrations for breach of contract, such as not telling the arbitrators the names of the parties. See, for example, Green Coffee Association of New York City, Inc., Rules of Arbitration R VII(1), VIII (1989). Second, because disagreements about quality are likely to arise even if everyone is acting in good faith, the availability of a tribunal that can objectively make this determination even before an action for breach is instituted should help promote cooperation. In practice, the use of a quality intermediary, particularly in those industries where the assessment of quality is subjective, is not particularly damaging to commercial relationships as long as recourse to it is not too frequent, perhaps because transactors understand that two people acting in good faith might have different assessments of quality. Interview with Cotton Merchant (July 1996) (on file with author).

Finally, the use of quality certification intermediaries may promote settlements in disputes where breach of contract for delivering improper quality is alleged. Having a wiseman determine quality effectively bifurcates the arbitration. With the issue of liability clear, and damages under these rules being determined using objective measures that do not require the revelation of firm-specific information—see L. BERNSTEIN, Private Commercial Law, op. cit., note 13 (noting the damage rules of the cotton industry and many other industries do not require the revelation of firm-specific information).

162 See, for example, NAWLA, Make this Text!, North Coast Weekly Letter No. 217 (April 6, 1926) (“Where lumber is bought at certain prices ‘less 5%’ the usual clauses used are found to be ambiguous and lead to unnecessary controversies at times. In order to comply with the usual custom in such cases and avoid misunderstanding, the following clause is suggested: [...] ‘Prices are f.o.b. mill; underweights to mill; less 5% after deducting actual freight.’”); NAWLA, Questions and Answers, North Coast Weekly Letter No. 225 (June 22, 1926) (“[T]he term ‘a carload’ is broad indeed and its interpretation by different persons and under different conditions leads to much trouble. The following qualifying clauses are suggested to avoid misunderstanding:”) Clause
Finally, even if it is linguistically difficult to capture particular aspects of the desired performance in words, the incorporation of trade usage may not be a desirable response. The Code requires the content of a usage to be established as “fact,” something that it is difficult to do if the custom cannot be adequately expressed in words. As a consequence, it might be better to encourage transactors to include a wise-man provision in their contracts, designating an industry expert who will make the necessary determinations in the event of a dispute. If the wise-man’s determinations were made binding in any litigated disputes, the wiseman would, in effect, function as a privately contracted-for merchant jury.

As regards the Code’s premise that merchants in fact consider usages to be fully part of their agreement, this the case studies of the incorporation debates, together with the text of trade rules, articles from the trade press, merchant testimony on the proposed Code, and the modern adjudicative approaches of many merchant tribunals, suggest that merchants did not view written and unwritten custom, much less unwritten custom and written contract provisions, as being on par with one another.

First, not all customs were included in the trade rules. This suggests that some were viewed as suitable or desirable for third-party enforcement while others were not. Second, some

---

N° 50: ‘Load between 25,000’ and 30,000’ on this order.’ Clause N° 51: ‘Not over ___ M ft.’).

163 U.C.C. § 1-205(2) (“The existence and scope of [...] a usage are to be proved as facts” by the litigant, who must then apply it to the circumstances of the case.).

164 Llewellyn’s early drafts of the Code contained a merchant jury provision that would have permitted the submission of many types of determinations to merchant experts whose decisions would be finding at trial.

165 In addition, the records of inter-association debates relating to amendments to the Southern Mill Rules, a set of rules governing cotton transactions between merchants and mills that are jointly drafted by the American Cotton Shippers Association (“ACSA”), a trade association of merchants, and the American Textile Manufacturers Association (“ATMI”), a trade association of mills, suggest that there were practices each association was willing to encourage its members to abide by, but that they refused to include in the rules as legally enforceable obligations. See, for example, Minutes of the 1990 ACSA/ATMI Joint Meeting 3 (June 19, 1990) (while the ATMI rejected a proposed ACSA rule dealing with sample approvals, the “ATMI did agree to notify its
rules, like the WSRs, explicitly distinguished between written and unwritten custom in their hierarchy of authority for interpreting contracts.\textsuperscript{166} Third, the trade press in various industries is filled with statements that a practice may be customary but that those who want it to be followed should include a provision to that effect in their contract.\textsuperscript{167} Fourth, in the rare instances that early trade codes contained provisions directing arbitrators to take custom into account, a choice of custom provision was often included in the rule.\textsuperscript{168} The inclusion of such provisions suggests that merchants recognized that customs varied from locality to locality. Fifth, there is a great deal of evidence that the practices that some industry participants described as customs were viewed by others as undesirable practices that ought to be changed, not as ideal, implicit contract provisions.\textsuperscript{169} Sixth, when testifying on the

members regarding shippers’ concerns with sample approvals"); \textit{Minutes of Joint ACSA-ATMI Meeting} 4 (June 8-9, 1981) (unpublished document, on file with author) (noting that “ATMI rejected any change in the rules [relating to rejections], but advised that they would ask ATMI members to handle rejections in a more expeditious manner.”).

\textsuperscript{166} See, supra, note 98 and accompanying text.

\textsuperscript{167} See, for example, Pacific Coast Shippers Association (“PCS\textsuperscript{A}”), \textit{The Secretary’s Weekly Letter} N° 82 (July 24, 1923) (“When a wholesaler orders lumber from an inland mill for delivery at a port for shipment by water, while it is customary for such shipments to be loaded on open equipment, it is not obligatory on the part of the mill who is allowed to use any available equipment unless it is specifically stated on the order that stock must be loaded on open equipment.”).

\textsuperscript{168} See, for example, Rule 5, 11 WWGT 35 (Nov. 5, 1921-22) (“[B]illing instructions must be furnished the Railroad Company in accordance with the custom then in vogue at the shipping point.”). See also ASTA, \textit{1955 Yearbook} 193 (1955) (“In the absence of any specific stipulation in the contract of sale or purchase applying to the type of package or packaging, it will be presumed that the custom prevailing in the area of production will apply.”). But see \textit{Trade Rules of the Grain Dealers National Association} Rule 36(e), 10 WWGT 40 (Nov. 20, 1920-21) (requiring brokers to negotiate “in accordance with the rules and customs governing such transactions”). Similarly, when answering questions involving custom, the NGFA was also careful to specify which locality’s custom governs. See \textit{Secretary’s Book} (1903) (no pagination in original), \textit{op. cit.}, note 59 (when the Trade Rules committee was asked, “[s]hould a receiver charge seller commission on grain failing to grade to contract when shipper orders elsewhere,” it “recommend[ed] that the usage of the market to which the grain may have been consigned shall govern”).

\textsuperscript{169} See for example, supra, notes 125 and 127 and accompanying text (silk); ASTA, \textit{ASTA Yearbook} 30-31 (1910) (“I wish to call your attention to a
proposed Code, merchants took the position that even practices that were almost universally followed in day-to-day relationships should not necessarily be written into law. For example, merchants acknowledged the practice of giving a price adjustment for nonconforming tender, but argued strongly for retention of the perfect tender rule, explaining that “the price adjustments that merchants made when goods ‘are not entirely up to standard’—the give and take of ordinary mercantile life’—should not be made obligatory in the law.”

Finally, modern merchant tribunals are, for the most part, very wary of taking unwritten custom into account in deciding cases. Although they will sometimes look to it when

---

Z. B. WISEMAN, loc. cit., note 4, 526 (emphasis in original). In arguing for the perfect tender rule, merchants explained that the type of buyer opportunism made possible by the rule, could “take care of [itself] ‘mighty quick’ through other merchant practices.” Id. In sum, “in the merchants view, the combination of the perfect tender rule and nonlegal sanctions was more advantageous than a rule of substantial performance with judicial discretion.” L. BERNSTEIN, “Merchant Law in a Merchant Court”, loc. cit., note 13, 1801.


In the early days of their operation, however, these merchant tribunals looked to custom more often. There are a number of possible explanations for the initial use of and gradual abandonment of incorporating custom for any purpose other than filling a pure contractual gap. First, even in the past, customs were not typically looked to for the purposes of varying terms or defining the meaning of written provisions, but rather to fill gaps. Because many early sets of rules were less detailed than modern trade rules, and written contracts tended to be less complete, perhaps because most firms were
faced with a true contractual gap, they have a much narrower conception of a gap than courts applying the Code, and they will almost never look to custom to interpret or vary a provision in a written contract or memoranda. In those instances when industry expert arbitrators do look to customs, they sometimes signal their discomfort by criticizing the contracting practices of the parties. More generally, merchants appear to have a clear sense that practices will often vary from contract provisions, sometimes in systematic and sometimes in relationship-specific ways, but that these variations should be left to the extralegal

small and more trade was local and personal, there were more gaps to be filled. See No Change in Rules 20 WWGT 34 (Nov. 30, 1930-31) ("I have been chairman of the Rules Committee for a number of years and it is very noticeable that suggestions for changes and inquiries for interpretation are becoming fewer each year. It is possible that this may be because my opinions are not of value, but I prefer to believe that it is because of a constantly increasing understanding of the mutual obligations and customs over different sections of the country, and through the Grain and Feed Dealers National Association a uniform and fair interpretation of rules and custom."). Second, in the early years more trade was local in scope, so those local customs that did exist may have been a sound basis for deciding cases. It is, however, important to exercise caution in drawing a straight inference that because merchant tribunals rejected incorporation, the public commercial law should as well. Merchant associations have institutional alternatives available to help solve the problems of contractual incompleteness that are unavailable in the public legal system. They can promulgate multiple sets of trade codes, carefully tailored to different types of transactions, as well as standard-form contracts that are even more closely tailored to particular contexts. In addition, they can draw on pre-existing forces of shaming and social suasion to encourage and discourage particular types of behavior, forces that are unavailable to either courts or to many of the industries whose contractual relations are governed by the Code. See, for example, Lisa BERNSTEIN, "Opting Out of the Legal System!: Extralegal Contractual Relations in the Diamond Industry", (1992) 21 J. Legal Stud. 115; L. BERNSTEIN, Private Commercial Law, op. cit., note 13, pp.16-25. However, because merchants testifying on the proposed Code sharply criticized the incorporation strategy—see notes 142 and 143 and accompanying text—the fact that most of their private legal systems reject this strategy is nonetheless quite telling.

172 L. BERNSTEIN, “Merchant Law in a Merchant Court”, loc. cit., note 13, 1775-1781. The NGFA arbitrators also do not permit custom to trump trade rules, even though the rules are said to be based on customary practices. See, for example, Texas Farm Products v. Topeka Mill and Elevator Co., NGFA Case N° 1507 (Mar. 31, 1970) ("[W]hile trade practices in the Kansas City area may differ in accepted meaning of terminology" from the practices in Texas, "this does not relieve the Seller from complying with Rule 38 unless so specified in writing"). See also infra, note 215.
realm for their enforcement and are essentially irrelevant to any dispute requiring third party adjudication.\textsuperscript{173}

The Code’s incorporation strategy has also been defended on the grounds that, at least as regards the gap-filling use of custom, customs are likely to be more efficient than any term the court can construct. The strength of this argument, however, depends strongly on acceptance of the strong-form Hayekian view of the evolution of custom that implicitly animates the Code, a view that suggests that efficient custom should evolve through the natural selection of rules and practices.\textsuperscript{174} This is, however, a view of custom that Hayek himself doubted in its purest form,\textsuperscript{175} and that seems increasingly untenable in light of the insights of modern game theory and cognitive psychology. Game theory makes clear that regularities in behavior can emerge from sets of circumstances vastly different from those envisioned by Hayek. For example, customs may reflect the unique or non-unique equilibria that are produced by interactions characterized by any of a number of games. These equilibria may or may not be efficient.

\begin{footnotesize}
\begin{enumerate}
\item\textsuperscript{173} For an extensive discussion of this point and supporting evidence, see L. BERNSTEIN, “Merchant Law in a Merchant Court”, loc. cit., note 13, 1777, n 43.
\item\textsuperscript{174} See Friedrich A. HAYEK, Law, Legislation, and Liberty, Vol. 1: Rules and Order, London, Routledge & Kegan Paul, 1973, pp. 35-54 and 74-91. However, Hayek himself recognizes that rules arising from an evolutionary process may “develop in very undesirable directions, and [...] when this happens correction by deliberate legislation,” may be “the only practicable way out.” Id., at p. 88.
\item\textsuperscript{175} Id., pp. 100-101 (where Hayek notes that for the best rules to emerge both the evolutionary forces, of spontaneous order, and the actions of courts and legislature are all necessary, explaining that “the system of rules as a whole [...] is the outcome of evolution in the course of which spontaneous growth of customs and deliberate improvements of the particulars of an existing system have consistently interacted. It is the outcome of a process of evolution in the course of which spontaneous growth of customs and deliberate improvements of the particulars of an existing system have constantly interacted.”).
\end{enumerate}
\end{footnotesize}
Their evolution may also be highly path-dependant and strongly influenced by information cascades or any of a number of heuristic biases. Although game theory also suggests that the usages reflected in some of these types of equilibria are unstable and can therefore be changed easily if they are thought to be undesirable, there are a number of forces that operate on commercial practices, including both the heuristic biases that helped create them as well as coordination problems and interpretive, educational, and enforcement-related network externalities, that make it less likely than game theory

176 See David HIRSHLEIFER, “The Blind Leading the Blind!: Social Influence, Fads, and Informational Cascades”, in Mariano TOMMASI and Kathryn I. IERULLI (eds.), The New Economics of Human Behavior, Cambridge, Cambridge University Press, 1995, p. 188, at 191. (“[A]n informational cascade occurs when the information implicit in predecessors’ actions (or resulting payoffs) is so conclusive that a rational follower will unconditionally imitate them, without regard to information from other sources [...] [C]ascades often spontaneously develop on the basis of very little information. People converge on one action quite rapidly, and their decisions are idiosyncratic.”). Because information cascades arise when people give relatively large weight to the behavior they observe around them, information cascades may well explain the existence of highly localized customary practices that differ widely from locality-to-locality in ways that cannot be explained purely in terms of transactional differences or other efficiency related criteria. See President Clement Makes Committee Appointments, 10 WWGT 25 (Nov. 20, 1921–22) (outside of the context of a particular dispute, how the “terms” “prompt,” “quick,” and “immediate” are defined when transactors are given a menu of time provisions to choose from has no distributional consequences). Although the theory of informational cascades shows that the results of the cascades are vulnerable to change, commercial customs established by cascades are subject to the same lock-in forces as other commercial customs, and in practice are therefore less likely to change in response to the introduction of new information than the models would suggest.

177 See, for example, Marcel KAHAN and Michael KLAUSNER, “Path Dependence in Corporate Contracting!: Increasing Returns, Herd Behavior and Cognitive Biases”, (1996) 74 Wash. U. L. Q. 347 (discussing how cognitive biases, including the status quo bias, the endowment effect, anchoring bias, conformity bias, and herd behavior, influence contracting decisions and reinforce recourse to standard written terms). Transactors may be particularly wary of suggesting departures from custom in markets where such suggestions are rare, since the relational costs of doing so are likely to be particularly high.

178 Interpretative network externalities are the benefits created by the fact that, as compared to the meaning of a specially-drafted provision, the meaning of a widely used contractual provision is likely both to be clearer (since it has already been interpreted by courts) and to become clearer over time (because it will be more likely to be interpreted by courts in future disputes).
standing alone would predict that trade practices, whether written or unwritten, that survive for a certain period of time are going to change.\footnote{179}

Moreover, even if customs did arise through Hayekian-type natural selection, there is no reason to think that they would provide desirable, legally enforceable contract provisions.\footnote{180} In the absence of an authority who stands ready to promulgate and enforce it, a custom will develop only if it was first a practice that was followed by a pair of transactors who found it in their best interest to do so.\footnote{181} The first time the

\footnote{179} Indeed, the trade press and association records in many industries suggest that it was extraordinarily difficult to change custom, see supra, note 127 and accompanying text, and that customs that no longer made even a modicum of business sense were often followed. \textit{Misuse of Order Bill of Lading}, 10 WWGT 30 (Feb. 20, 1920-21) ("A great many shippers are following a custom which has prevailed for years without knowing why or having any good reason for doing so. In many cases, the shipping clerk is merely following precedent without knowing why.").

\footnote{180} See E. A. POSNER, \textit{loc. cit.}, note 9.

\footnote{181} However, it is possible that under the Code a practice that came into being by conditioning only on observable information, might, as it evolved first into a course of performance, then into a course of dealing, and finally into a custom, come to condition on information that was a verifiable proxy for the observable information that the practice originally conditioned on. This process of transformation would likely be accelerated by parties' attempts to enforce these practices in court, since in each dispute the parties would be required to adduce verifiable information in support of their claim. As a custom comes increasingly to condition on verifiable information, unless the
practice was followed, it would have been legally unenforceable since it would be neither a course of performance nor a course of dealing. As a consequence, the practice could only condition on information that was observable to the transactors. Without the threat of legal compulsion, a practice that conditioned on information that was only verifiable, that is, a practice whose applicability could only be determined by a third-party who could compel the parties to reveal information, would not be workable. So even if customs were perfect Hayekian customs that arose from the force of natural selection operating on practices, there is no reason to expect them to be the optimal legally enforceable contract provisions or necessarily better provisions than a court could construct using any of a number of interpretive methods.

---

observability conditions concurrently survive the evolution, the cooperation-promoting aspect of the custom will begin to disintegrate.

182 Observable information is information that the transactors themselves can obtain at the relevant time at a cost they regard as reasonable ex ante. The observability of a condition may change over the life of a contracting relationship as the transactors learn about one anothers’ operations, thereby expanding, over time, the extra-legaly contractible aspects of the deal.

183 Verifiable information is information that transactors would, ex ante, view as both possible and financially desirable to prove to a third-party in the event a dispute arose. Alan SCHWARTZ, “Relational Contracts in the Courts: An Analysis of Incomplete Agreements and Judicial Strategies”, (1992) 21 J. Legal Stud. 271, 279. Some observable information is also verifiable, but not all verifiable information is observable.

184 The preference for contract over unwritten custom in trade associations is, however, influenced by considerations that may play out differently in the legal arena. While unwritten customs are subject to the lock-in forces and other cognitive distortions discussed in the text, most of these forces also operate in largely the same way on written contractual provisions in industries without trade rules or centrally drafted contractual forms. See M. KLAUSNER, loc. cit., note 178. The only difference is that written provisions containing customary understandings can initially condition on verifiable information. If one thought that lock-in forces caused more inefficiency in written terms than unwritten terms, it might be desirable to interpret written provisions by reference to unwritten custom as the Code directs. If, however, one thought that unwritten customs were more subject to these forces, it would be undesirable to look to them unless faced by an unambiguous gap. In trade associations, however, the trade rules-creation and amendment processes, and
More generally, both the Code’s incorporation strategy and the dominant evolutionary views of the emergence of custom are based on the same curious assumption. Namely, if most transactors in a market could costlessly arrive at and draft a term governing an issue, they would likely arrive at the same term, the customary term. Nevertheless, in many, if not most, contractual settings, this assumption is likely to be false. Transactors are likely to have different perceived trade-offs between price and other terms, different abilities to use slightly imperfect goods, different abilities to finance their cash gap, and different risk preferences. There is no reason to suppose that, given these differences, all transactors would choose the same contractual provisions. Indeed, even in cash-commodity sales that are relatively standard as compared to many of the

the widespread use of association-drafted standard forms, means that written contractual provisions can be more readily changed and are less subject to lock-in forces than are both unwritten custom and perhaps written contracts in non-association dominated markets. Indeed, the decision of trade associations to create written sets of trade rules might be understood as an efficiency enhancing institutional response to problems inherent in the evolution of both custom and individually drafted contracts. See L. BERNSTEIN, Private Commercial Law, op. cit., note 13, pp. 28-30.

185 In addition, it is likely that transactors in a given market will have more defined transaction types in their mind than a court, and how exactly the court defines the type of transaction perceived to be at issue may, as Craswell has argued, be outcome determinative. R. CRASWELL, loc. cit., note 8. Yet it is quite likely that this is the type of determination a court will get wrong. Consider a contract between a merchant-seller of cotton and a mill-buyer. There are many mills that produce a variety of goods, depending on the demand of their customers. Suppose the mill sometimes makes sheets, and that when it does it is terribly important that a particular quality of cotton is used. Other times it makes denim, a production process where quality is much less important. A court attempting to find a trade custom relating to the permissible quality variation might well aggregate over both types of transactions, yet one would not expect the optimal amount of quality variation to be the same in contracts for goods to produce sheets as in contracts for goods to produce denim. Or, consider a contract for the sale of silk. A court might not distinguish between a contract for silk to be used to make gloves and other types of silk, yet in the silk trade, when it comes to payment time, the difference is outcome determinative. Compare Standard Rules of the Silk Glove Group, Rule 4, in Organization of Glove Silk Group, 2 Silkworm 17 (March 1921) (“Shipment made within one week after specified date of delivery shall constitute good delivery.”) (emphasis added) with Broad Silk Rules Adopted, 2 Silkworm 10 (January 1921) (“Rule 5 [...] shipment made within two weeks after specified date of delivery shall constitute a good delivery.”) (emphasis added).
transactions governed by the Code, trade rules typically offer transactors a menu of possible terms for each core aspect of their trade.  

Moreover, the Code’s incorporation strategy reflexively incorporates customs into all contracts in a market, unless the customs are clearly and specifically negated, and looks to customs to interpret even explicit and facially unambiguous contract provisions. It thereby moves the meaning of explicit provisions as close as possible to the meaning of customary terms, and in so doing transforms many customary practices

---

186 See, for example, NGFA, Trade Rules and Arbitration Rules, Grain Trade R 8 (“Time of Shipment or Delivery”) (1995) (defining a selection of time designations including “immediate,” “quick,” “prompt,” and “first half of the month shipment”); id., R 4A (defining several weight classifications to choose from); id., R 4B (defining several types of inspection).

187 U.C.C. § 2-202 cmt 2 (“Unless carefully negated [usages of trade] have become an element of the meaning of the words used.”). In addition, because the duty of good faith cannot be “disclaimed by agreement,” id., at § 1-102(3), and because the Comment notes that the good faith requirement is to be “implemented by Section 1-205 on course of dealing and usage of trade,” contracting out of trade usage is quite difficult. Id., at § 1-203 cmt.

188 See Nanakuli Paving and Rock Co. v. Shell Oil Co., 664 F.2d 794-805 (9th Cir 1981) (binding seller to trade usage where a contract for the sale of asphalt contained an explicit provision that stated that the seller’s posted price at the time of delivery would govern, but the buyer’s claimed that it was entitled to “price protection”—a usage of trade in the asphalt-paying trade in Hawaii that required the seller to sell at the original price all the asphalt that the buyer had committed to use in jobs on which it had already bid in reliance on the sellers price, and the court found that seller was bound by this usage). See also Columbia Nitrogen Corp. v. Royster Co., 451 F.2d 3, 7 (4th Cir. 1971) (holding that, despite express price and quantity terms and a standard integration clause, evidence to show that it was a custom and usage of the fertilizer industry that ‘express price and quantity terms in contracts for material [...] are mere projections,” was admissible to establish a consistent, additional enforceable term in the parties’ agreement). See, for example, Sherrock v. Commercial Credit Corp., 277 A.2d 708, 711 (Del. Super 1971), revd on other grounds, 290 A.2d 648 (Del. Super 1972) (finding buyer to be a merchant buyer and therefore obligated to follow reasonable commercial standards in the trade, and holding that the buyer’s decision to pay early and let seller retain possession for several days thereafter was not in accordance with the custom of the automobile trade and was unreasonable, explaining that “departures from customary usages and commercial practices should be viewed as strong indicia that the practice is not reasonable”).

189 Llewellyn’s preference for moving the meanings of terms and obligations as close as practicable to their customary meaning is amply reflected in the
into quasi-mandatory standardized provisions in all contracts in the relevant market.\textsuperscript{190} It also prevents transactors from actually

legislative history of the Code. In a comment on a proposed Draft of the Code, Llewellyn wrote:\textsuperscript{169}

“No inconsistency of language and background exists merely because the words used mean something different to an outsider than they do to the merchants who used that language in the light of the commercial background against which they contracted. This is the necessary result of applying commercial standards and principles of good faith to the agreement [...]. Moreover, where the commercial background normally gives to a term some breadth of meaning so that it describes a range of acceptable tolerances rather than razor sharp-edged single line of action, any attempted narrowing of this meaning by one party is so unusual as not likely to be expected or perceived by the other. Therefore, attention must be called to a desire to contract at material variance from the accepted commercial pattern of contract or use of language. Thus, this Act rejects any “surprise variation from fair and normal meaning of agreement.”


For a careful discussion of the jurisprudential inseparability of the Code’s duty of good faith, its reliance on usage, and its definition of agreement, which traces the evolution of these ideas in Llewellyn’s thought and through drafts of the Code, see id., at 199-202. See also L. BERNSTEIN, “Merchant Law in a Merchant Court”, loc. cit., note 13.

\textsuperscript{190} Although the incorporation strategy implicitly treats many customs as terms in standard-form contracts, often referred to as contracts of adhesion, customs do not, for the most part engender similar distrust. But see R. DANZIG, loc. cit., note 6, 626-627. Although many of the objections to standard-form contracts are untenable from an economic point of view, these objections are widely accepted, so it is important to note that they provide an even stronger ground for objecting to implicit, customary, standard-form contract provisions.

The two main grounds for distrust of standardized contracts, are “[f]irst, [that] most persons presented with standardized forms do not bother to familiarize themselves with the specific content,” and “second, [that] because the advantages of forms would be lost if bargains were open to routine renegotiation, their users are often unwilling to do business on other than standard terms.” Avery WEINER-KATZ, “Standard Form Contracts”, in P.\textsuperscript{!}NEWMAN, \textit{op. cit.}, note 21, p. 502. Both of these concerns are even more problematic in relation to custom. First, the per transaction cost to a transactor of reading a standard-form contract drafted by someone he does business with frequently is small in comparison to learning all of the practices in vogue in the other transactor’s market. Indeed the cost, difficulty, and bother of learning the practices of numerous localities was a major impetus behind the drafting of national grain rules. \textit{Trade Rules in Danger?}, loc. cit., note 67, 28. (“No buyer can possibly keep up with all the rules and regulations of all the shipping markets of the country. That’s why the National rules were
following a custom that they perceive as a desirable, legally unenforceable practice, but undesirable legally enforceable contract provision, because if they follow the custom on more than one occasion, it would likely be transformed into a legally binding course of dealing or course of performance. In addition, if there are, in fact, customs that cannot be linguistically captured in a contract provision *ex ante*, the incorporation strategy transforms them into mandatory terms because in order to be excluded, the contract must negate them with specificity, something that by assumption is not possible.

**Conclusion.** In sum, the incorporation strategy is based on an overly broad conception of the types and geographical scope of customs that consistently exist in merchant communities. However, even if one believes that some customs do exist, support for the incorporation strategy in merchant industries would not necessarily follow. In most merchant industries, transactions are consummated orally and confirmed by sending standard-form memoranda with long, back-of-the-

---

191 See *supra*, note 187.
page printed recitals. In such contexts, those customary understandings that truly do exist could easily and inexpensively be memorialized in standard-form boiler-plate. In other words, the number of customs that are clearly enough defined and widely enough known to be true implicit contract provisions, but that are nonetheless not worth including in contractual boiler-plate, is likely to be quite small.

Nevertheless, because the concept of custom seems to retain some salience to merchants—who use it variously to describe their own past contracting practices or the loose distribution of contractual behavior in the market as a whole—it is useful to explore in more detail the role played by the notion of commercial custom in merchant relationships.

III. The Relationship-Creating Role of Custom

The empirical evidence presented in Part I casts doubt on the systematic existence of industry-wide unwritten customs that are generally known, geographically co-extensive with the scope of trade, and implicitly assented to by market transactors. It also substantiates the likely existence of some local customs as well as some relatively common industry-wide practices. More generally, however, the evidence strongly suggests that the types of customs that exist, even in these rather well-defined merchant communities, do not amount to anything close to the all-pervasive sets of implicit gap-filling provisions and dictionary-type interpretive guides assumed by the Code.

Nevertheless, while merchants do not, for the most part, conceive of customs as providing them with legally enforceable contract terms, most merchants evidence a keen interest in learning and talking about the way business is usually done in their industry. One explanation for this interest is that transactors approach transactions, particularly those with a stranger, with a rough sense, derived from aggregating behavior over the market as a whole, of the way such deals are usually concluded and performed.\footnote{Alternatively, the customs and commercial standards incorporated by the Code may actually be important sources of contractual understanding for smaller merchants who engage in mostly local trade.} These aggregations (or distri-
butions) of behavior, which might be termed weak-form customs, play an important role in the development of commercial relationships. This role can best be understood by drawing on the insights produced by basic signaling models and by recognizing that many merchant-to-merchant commercial transactions have features that can best be captured by the intuitions underlying repeat-play prisoners’ dilemma models, in which cooperation is best established and maintained when transactors follow any of a number of tit-for-tat strategies. These strategies dictate responding to cooperation with cooperation and defection (or a series of defections) with defection for a defined period of time followed by a return to cooperation.

One of the main barriers to both establishing and maintaining commercial cooperation is that it may be difficult for a transactor to determine, by simply observing actual outcomes, whether the other transactor has defected. Making this determination is a difficult and error-prone process, particularly in merchant industries where the commodity being sold passes through many hands and the quality of the good is so strongly affected by climactic, storage, and transportation conditions that even after optimal precautions have been taken, a significant probability remains that performance will not be exactly as promised. Each time a transactor mistakenly

---

193 For an overview of these signaling models that explores their application to numerous legal issues, see Eric A. POSNER, Law, Cooperation and Rational Choice (DRAFT) (1998) (on file with U. Chi. L. Rev.). The discussion here differs slightly from Posner’s in that the costly action that constitutes the signal sent by the seller is not a deadweight loss (in the way that burning money, which would also send the requisite signal, would be a deadweight loss) because the high quality delivery will benefit the buyer.


195 For example, a bale of cotton may pass though the hands of a farmer, a grain elevator operator, a country merchant, a city merchant, a warehouseman, another city merchant or broker, a transport company (railroad, truck, or barge), and a mill. See generally L. BERNSTEIN, Private Commercial Law, op. cit., note 13.

196 See, for example, R. C. GORDON, op. cit., note 56, p.49. (“Each step in the marketing chain [of grain] involves commitments and contracts between two or more parties. The vagaries of weather and transportation congestion and
classifies an outcome as defection, the likelihood that she will respond by defecting now or in the future increases, making relationship breakdown more likely.\footnote{197}

In order to understand the role played by these weak-form customs, consider a seller (S) who is dealing with a buyer (B) for the first time in a market where repeat-dealing relationships, once established, have an economic advantage over discrete transactions, perhaps because they reduce cooperation and performance costs as the transactors learn about one another’s expectations and business operations.\footnote{198} Suppose delays can add further disruptions. From the simplest grain purchase contract with farmers to complex, multi-layered string trades, there is an inherent opportunity for misinterpretation of a contract term or condition or nonperformance.”); NHA, Report of the Fourth Annual Meeting 23 (1897) (“[A] shipper may in perfect honesty consign a car of hay to the receiver under the impression that it is N° 1 hay, and if it proves to be N° 2 hay, that is not conclusive evidence that the shipper is dishonest.”).

\footnote{197} Where two transactors are playing a “pure” tit-for-tat strategy, a single act that is misinterpreted by Transactor 2 as defection will lead him to defect in response. This will, in turn, lead Transactor 1 to defect, and cooperation will disintegrate. In real-world interactions, however, because a “pure” tit-for-tat strategy is so sensitive to breakdown, game theory suggests that a more successful strategy would be a modified, more “forgiving” tit-for-tat strategy that does not dictate responding to every bad outcome (that is every real or suspected defection) by inflicting a punishment. See A. DIXIT and B. NALEYBUFF, op. cit., note 195, p. 113. These relatively forgiving strategies, such as tit-for-two tarts, are commonly followed by merchants. See L. BERNSTEIN, Private Commercial Law, op. cit., note 13, pp.57-59 (cited in note 13) (noting that a strategy of negotiating forgiving adjustments until a relationship is terminated is common in transactions between cotton merchants and cotton mills; and noting that a strategy of ignoring defections or making forgiving adjustments in response to a certain number of defections and responding to defections thereafter with punishment of a limited variety, such as refusing to deal for a specified period, and then returning to cooperation, is commonly followed in transactions among merchant members of the Memphis Cotton Exchange).

\footnote{198} For example, as S learns about B’s business, he can more accurately assess when slightly nonconforming tender will or will not disrupt the buyer’s business (if tendered with a price adjustment) and can adjust his precaution investment accordingly. As one Texas Feed Dealer explained, “I prefer to deal with old customers since when I ship, I know that the goods will be suitable and will therefore be accepted even if they are not exactly to contract.” Interview with Texas Feed Dealer (1999) (on file with author). Similarly, in the cotton industry, where mills and merchants have been unable to agree on rules regarding the unloading of truck shipments—particularly, the hours receiving stations must operate, the order in which arriving trucks will be
further that at the outset of the relationship B knows nothing about S. In such a situation, B will assume that if S is trustworthy (that is, if he is a cooperator) he will perform within customary bounds on all aspects of the deal, and that if he is not trustworthy (that is, if he is a defector) his performance will fall outside these bounds. The contract takes this into account through the price term. B is willing to pay a price corresponding to the customary ranges of performance on all aspects of the contract, weighted by the probability of their occurrence. These customary ranges roughly, though imperfectly, reflect the aggregate performances of all transactors in the market and are largely common knowledge to both buyers and sellers.  

Suppose, for example, that quality is the key term in a wheat contract. If the contract calls for N° 3 Wheat, the weak-form customary range would correspond to how often in the market, when a contract calls for N° 3 Wheat, N° 3 exactly is delivered, how often something better than N° 3 is delivered, and how often something worse than N° 3 is delivered.

Returning again to B and S. If S’s performance is within the customary range, B will conclude that S is a cooperator and

---

199 Although these aggregations of practices are unlikely to supply the optimal performance ranges for any particular pair of transactors, aggregating as they do over transactions that likely differed greatly in both purpose and scope, they may nonetheless be valuable in these early rounds precisely because they are common knowledge. The element of value created by the fact that they are common knowledge may outweigh their lack of a tailored fit with the specifications and error ranges the transactors really desire. B and S should be equally able to observe aggregate market behavior so weak-form customs should be a fairly good cooperation initiator. But see infra, note 221.

200 Transactors seem to instinctively use these types of ranges to assess whether the other transactor intended to cooperate or defect. See, for example, Secretary’s Board Book, supra, note 59, at 80-81 (“Most dealers state on their card bids that grain missing grade will be taken at market difference on day of arrival, and I believe that a fair interpretation of this term means that the seller who accepts must deliver grain of the grade he agrees to, or some so near it in quality that he may really expect it to grade. If a person sets up a target, turns his back to it and shoots in the opposite direction, it cannot reasonably said that he misses the target, and it is equally true that if a person sells N° 2 yellow corn and delivers corn which inspects N° 4 or no grade, it cannot meaningfully be said that the corn missed grade. This may seem an exaggerated illustration, but it explains the point.”).
will enter into another deal. This will continue as long as S’s performance is within the customary range and, over time, a pattern of S’s performance being closer than customary to the desired performance emerges. In these early stages of the emergent contractual relationship, S may be taking a loss on each contract. He is taking the level of precaution associated with delivering close to the promised performance, but is only receiving the price associated with customary ranges of performance. However, S’s willingness to take this initial loss adds value to the relationship. It is a credible signal to B that S has a low discount rate—that is, that he plans to continue to cooperate with B in the future. Over time, if S’s performance is closer than average to the desired performance, here the promised quality, B will see this course of dealing (“CD”) emerge. Gradually, she will come to pay S a higher price to reflect this better performance.

201 Alternatively, S may take the customary level of precaution and deliver only within the customary range, thereby not incurring any loss in the initial rounds. In these early rounds, the transactors will still learn about one another’s operations, giving rise to switching costs, which may facilitate the creation of a repeat-dealing relationship. However, this method of establishing a relationship is less likely to succeed than the method discussed into the text, since S’s behavior in earlier rounds conveys less information about his likely future behavior to B.

202 During these early rounds of the transactional relationship, the weak-form custom also provides benefits to S. In its absence, and in the absence of any other agreed range, S would take even more precautions to ensure that performance was exactly as promised, fearing that B would interpret any deviation as breach that would trigger a retaliatory response. However, in addition to giving B a way to ascertain that the deviation was unlikely to be a defection, the custom also constrains his behavior. When B rejects goods that are within customarily acceptable ranges, S can threaten to impose multilateral reputational sanctions—publicizing B’s action to market transactors who would likely agree that B’s action was improper. Indeed in the cotton industry, mills that acquire a reputation for being very inflexible on quality deviations find that they have to pay more for goods than do mills with a reputation for flexibility. See, for example, Interview with cotton merchant (date unavailable) (on file with author) (“Mills get a reputation about how flexible they are when certain circumstances arise [and] a merchant considers this very valuable business knowledge, this information about how flexible a mill will be. Some mills are more stringent about adjustments, and are very strict about demanding exact conformity to the contract. These mills have to pay more for cotton than mills who are willing to be more flexible. The market knows.”). Conversely, when B accepts goods in the customary range even though they are not exactly what he wants, S is not going to interpret this as acquiescence that lower quality delivery is “acceptable.” However, the
Despite the loss he suffers in early rounds, S has confidence that, once a repeat-dealing relationship is established, B will want to remain in it and will therefore let S slowly recoup his initial investment.\(^{203}\) S's confidence that B will behave this way is rational because the early rounds of any contracting relationship are costly to B even though she is paying only the customary price and may, in fact, be receiving higher than customary quality. Because quality is important to B, during these initial transactions she may have to deal with multiple suppliers to ensure she will have the requisite amount of a particular quality of the commodity on hand. Alternatively, she may have to keep sufficient inventory on hand to cover her needs in case the new S delivers towards the low end of the customary range.

The information transactors learn about one another’s businesses in these early dealings gradually reduces the coordination and performance costs of dealing with one another,\(^{204}\) giving rise to switching costs for both B and S. Once the relationship is established, both transactors therefore have a reason to want to maintain it.\(^{205}\)

---

\(^{203}\) This can be thought of simply as a relationship-specific investment that will be amortized over time.

\(^{204}\) See note 198 (cited in note 13).

\(^{205}\) Alternatively, even in the absence of either weak-form customs or good reputation information, cooperation could emerge if a buyer and seller had a common view of the distribution of outcomes (where an outcome is an actual completed aspect of the promised performance) after an agreed-upon level of precaution had been taken. In such a situation, outcomes within the agreed zone of error, or perhaps within two standard deviations of its mean, would be considered cooperation. Outcomes outside the zone, or repeated outcomes outside the zone, would be considered defection. More specifically, if transactors had a common understanding of the error ranges associated with each aspect of performance after optimal precautions had been taken, if the Code were to define customary ranges in terms of these random errors, and if courts too could make these determinations with reasonable accuracy at reasonable cost, enforcing this “custom” as a term of the contract would be value creating. It might remove the incentive for S to engage in too high a level of precaution early in the relationship where a random but large deviation is most costly to the relationship (either because the relationship-specific background to assess it has not yet developed, or because it is one of a fewer number of data points on behavior and so will have larger influence
Once B and S establish a relationship-specific CD zone, and the contract price adjusts both to reflect it and to permit S to recoup his initial investment, if S’s performance is within this zone it will be considered cooperation, and if it is outside this zone, or outside of the zone on several occasions, it will be considered defection. At this stage of B and S’s contracting relationship there are two kinds of defection. Defection that is outside the CD zone but within customary range and defection that is outside even the customary range.

Defection that is outside the private CD zone but within the customary range, may, if it is a non-regular occurrence, call for a price adjustment on the next deal, or some other measured response such as not dealing for a few rounds. Although the CD zone is narrower than the customary range, making more outcomes defection, a CD relationship is actually less vulnerable to breakdown than a customary one. Because transactors in a CD relationship are relatively well informed about one another’s operations, they can more accurately infer whether the other cooperated or defected than can transactors in the early and customary stage of their relationship. As a consequence, transactors’ responses to outcomes just outside of a well-established CD zone are more tempered. This intuition is reflected in the grain poem “If I Knew You and You Knew Me”!

If I knew you and you knew me
’Tis seldom we would disagree;
But never having yet clasped hands,
Both often fail to understand
That each intends to do what’s right
And treat each other “honor bright.”
How little to complain there’d be
If I knew you and you knew me.
When’er we ship you by mistake,
Or in your bill some error make,
From irritation you’d be free
If I knew you and you knew me.

---

on the transactors’ assessments of one another’s behavior than would a later deviation) and will effectively constrain opportunism by both B and S. In such a regime, S might try to be opportunistic by consistently delivering to the low end of the range, but if he did so B could terminate the relationship.
Or when the checks don’t come on time
And customers send nary a line,
We’d wait without anxiety,
If I knew you and you knew me.\textsuperscript{206}

In most industries, merchant transactors’ responses to particular breaches do depend strongly on the identity of and their prior relationship with the breaching party.\textsuperscript{207}

In contrast, defection outside the customary range, such as either a delivery or a series of deliveries outside the customary range, is likely to signal defection and to trigger an end-game round—that is, termination or suspension of the relationship and a resort to arbitration to recover money damages.\textsuperscript{208} Moreover, because all market participants know the

\textsuperscript{206} 12 WWGT 35 (June 20, 1922-23).

\textsuperscript{207} NHA, Report of the Twenty-Seventh Annual Convention 55 (1920) (“We have noted instances of where a shipper would invoice a car as being N° 2 timothy, when as a matter of fact there might be some ‘off’ grade hay. However, the receiver knowing the shipper and of the conditions under which the hay was bought and loaded, knew that there would be no trouble over the invoicing as he knew the shipper was not trying to slip something over on him. Hence the value of personal relationship between buyer and seller. In other instances this practice causes a lot of trouble.”). See also E-mail from participant in the Cotton-L List Serve (One cotton transactor explained that if a transactor asks to be released from a contract “you have to look at the circumstances for not providing the cotton (no prices going from 70 to 80 is not acceptable). A disaster is a disaster. Who you are dealing with is also important, for both sides. Have you been selling your cotton to this guy for 10 years.”).

\textsuperscript{208} Although it is difficult to establish whether or not a dispute is an end-game dispute, in part because even if two agents never deal again, their companies might still choose to do so, there is nonetheless evidence that in some merchant tribunals most cases are end-game disputes. For example, between 1975 and 1996, 54 percent of the cases heard by the Board of Appeals, the tribunal that resolves disputes between cotton merchants and cotton mills, were absolute end-game disputes. The most common triggers of the end-game were the insolvency or financial distress of a party or the closing of a cotton office, and the second most common were a change of control of one of the entities involved, or the retirement of a person directly involved in the transaction, both events that upset settled expectations. Another 18 percent of the cases involved disputes that might fairly be classified as end-game. They involved the effects of a government subsidy program that made large sums turn on who had possession of cotton on a particular day. These cases were ones in which an event took the contract out of the self-enforcing range and made it worthwhile for one of the parties to end their contracting relationship. The remaining 27 percent of the cases involved primarily factual disputes about
loose contours of the customary range, defections outside of it can also be sanctioned through multilateral reputation sanctions. These sanctions are only available when a significant number of market transactors agree that an outcome, or a series of outcomes, are defection and when the defection or defections are either observable to many market participants or can be credibly revealed to them at a reasonable cost by the breached-against party.209

As each transactor comes to know the other’s business and switching costs increase, making defection more costly, the CD zone becomes increasingly stable and the weak-form customs that initially provided useful relationship-creating norms diminish in importance. Once this occurs, the transactors can begin to develop value-creating relationship-specific, relationship-preserving norms (RPNs) that may (or may not) depart from industry practice, the transactors’ written contracts, and, in the trade association context, the default provisions supplied by trade rules.

Sometimes these RPNs may be explicit extralegal agreements that contradict the contract’s terms. An example is an agreement not to demand the federally supervised weights required by most feed contracts, but instead to rely on one another’s in-house weights, typically splitting any differences that may arise until opportunism is suspected and one transactor

late payment and late delivery. Transactors explained that if they had to take a case to the Board of Appeals they would be quite unlikely to deal with the other party again. L. BERNSTEIN, Private Commercial Law, op. cit., note 13, pp. 46–47. See also NHA, Report of the Twenty-Fifth Annual Convention 126 (1918) (where the arbitration report notes that “once a case is brought to the attention of the Arbitration Committee, the parties concerned will no longer have the confidence in each other. Their difference begets contradiction, contradiction begets heat, and heat rises into rage and ill-will. Then all human judgment is laid aside and men no longer consider their business transactions as sacred.”).

209 Although the sanction for a breach that falls outside of the customary zone is larger than it would be in a world with no rough consensus as to the parameters of the zone, it is not clear that deterrence will necessarily be too high. In the repeat-play context, this hybrid legal and reputational sanction is not usually imposed for any one deviation but rather for several. It is therefore likely to be imposed less often and more accurately than the standard monetary remedy for breach.
reverts to demanding compliance with the contract.\textsuperscript{210} In just this vein, one lumber dealer explained:

\textit{Of course, we all know that many small adjustment are made every day, where allowances asked are not unreasonable and where the salesman can verify the customers contention, but when it looks as the ‘something was being slipped over’ or where the mill will not agree to any allowance, THEN IS THE TIME TO DEMAND OFFICIAL INSPECTION ACCORDING TO THE TERMS under which THE LUMBER WAS SOLD.\textsuperscript{211}}

Other times, RPNs may be variants of the contract’s explicit provisions that add flexibility by defining acceptable ranges of performance. Alternatively, RPNs may recast the contract’s explicit provisions, which may condition on information that is verifiable but not observable\textsuperscript{212} into extralegal understandings that impose approximately the same obligations but condition only on information that is observable to both transactors.

In repeat-play contracting relationships in particular, provisions that condition only on information that is either observable or both observable and verifiable are more likely to successfully maintain cooperation than provisions that condition on information that is verifiable but not observable. When transactors bargain in the shadow of information that is only

\textsuperscript{210} L. BERNSTEIN, “Merchant Law in a Merchant Court”, \textit{loc. cit.}, note 13, 1793, n96.

\textsuperscript{211} See PCSA, The Secretary’s Weekly Letter N\textdegree{} 43 (Oct. 4, 1911) (emphasis in original); See also \textit{Fagg & Taylor v. S. F. Scattergood & Co.} (NGFA 1912), reprinted in \textit{Decisions of the Arbitration and Appeals Committees, loc. cit.}, note 49, 74-75 (in a case where the acceptability of particular routing was at issue, the defendant “acknowledges that he has used [that is, accepted] this routing on similar contracts, but states that he did so because it did not cause him loss or inconvenience, and further that it is their custom to ignore breaches of the contract where they are not inconvenienced,” explaining that they “seldom pay any attention to technicalities [...] [that] do not cost us anything [...] but where a shipper does not fulfill his contract and the amount involved is sufficient to warrant our insisting upon a shipper fulfilling a contract we do not think that we should be called technical”).

\textsuperscript{212} See \textit{supra}, notes 182 (defining observable) and 183 (defining verifiable).
THE QUESTIONABLE EMPIRICAL BASIS OF ARTICLE 2’S

verifiable, in the absence of the requirement of full disclosure required by the discovery rules, it is often difficult—if not either impossible or prohibitively costly—for transactors to credibly reveal to one another the information necessary for a cooperative adjustment to be reached. It may therefore be difficult for transactors to recognize the parameters of the zone of agreement, thereby increasing the risk of cooperative breakdown. Although most of the economics literature discusses information that is observable but not verifiable, or information that is both observable and verifiable, in legal settings one of the most important types of information is information that is verifiable but not observable. The rules of civil discovery would make no sense if this latter category were not of primary importance in defining legal relations and deciding cases.

Consider, for example, a contract provision calling for the payment of fully compensatory expectation damages in the event of breach. Suppose that S inadvertently breaches the contract and, in a cooperative vein, immediately offers to pay B full expectation damages. In many contexts, to make B truly whole this measure would have to include lost profits, which must be calculated by reference to both expected revenue and expected costs. While B might be able to credibly reveal his expected revenues, perhaps by showing S written orders he will be unable to fill, B cannot credibly reveal his costs. The lower B’s alleged costs, the greater his recoverable damages. S therefore has no reason to believe that B has accurately revealed all of his costs. In the absence of even the weak threat of discovery sanctions for incomplete information revelation, B’s revelations will not be credible. This uncertainty may give rise to suspicion, and may lead the parties to have different views of the settlement zone. In contrast, had the contract called for damages in the amount of the difference between the contract and the market price, the zone of agreement would be clear and, because the breach was inadvertent, cooperation would more likely be restored.

213 In addition, these costs may also be very complicated to prorate to a specific contract. Moreover, B may hesitate to reveal them since the next time B and S negotiate a deal, S will be able to more accurately assess B’s reservation price.
More generally, because most efficient customary norms and the relationship-specific RPNs that give rise to them will not condition on information that is only verifiable, and may condition on information that is observable but not verifiable, customary norms are unlikely to be the optimal rules for a tribunal to apply in the event of a dispute. And, if transactors knew, as they would if their contract were governed by the Code, that any relationship-specific RPN followed more than once may be transformed into a legally enforceable obligation in the event of a dispute, they would be far less willing to diverge from the terms of their written agreement, even in situations where doing so is highly beneficial to one transactor and costless to the other.

Understanding these and other reasons why the optimal norms to govern a relationship are likely to differ at different stages of a contracting relationship, and recognizing that relationship-creating and relationship-preserving norms are likely to differ in content and structure from the optimal end-

---

214 For example, in markets where most transactions are between transactors with long-term, repeat-dealing relationships with one another, if those customary RPNs that do exist emerge from pair-wise sets of transactors choosing independently to follow them as parts of established courses of dealing, it is likely that these RPNs will be relatively well-suited to mature, long-standing transactional relationships but relatively ill-suited to new contractual relationships where trust has not yet been established and the transactors have minimal information about one another’s business operations. As a consequence, courts looking to these customary norms to interpret contracts made in early stages of relationships will be imposing obligations the transactors did not, and would not, have voluntarily assumed. In addition, because it will be easier to prove to a court those prior instances where a practice was followed than it will be to prove that a precondition failed to occur so a practice was not followed, court interpretations of custom are likely to leave out important preconditions. Perhaps the most important such precondition is the degree of trust (defined here to mean a transactor’s perception of the likelihood that the other transactor will act opportunistically if unconstrained by a legal or nonlegal sanction) the transactors have for one another. The types of norms one would agree to be governed by if dealing with an angel are quite different from the types of norms one would agree to be governed by if dealing with a scoundrel; yet the ability of courts to distinguish between angels and scoundrels is, on the margin, quite limited.

For a discussion of additional reasons why one would expect both actual and optimal relationship-preserving customary norms to differ from optimal legally enforceable contract provisions, see L. BERNSTEIN, “Merchant Law in a Merchant Court”, loc. cit., note 13, 1796-1815.
game norms for a tribunal to apply in the event of a dispute, makes it easier to understand why merchants do not want either their relationship-specific courses of performance and courses of dealing, or their every-day customary practices—like granting price adjustments for non-conforming tender—incorporated or written into the law.

It is, however, difficult to empirically verify that weak-form customs are in fact useful relationship-creating norms and are being used to initiate repeat-dealing relationships. Nevertheless, as demonstrated earlier, transactors attach a great deal of importance to their prior dealings with a transactor in deciding how to respond to different situations that may arise, and there is some evidence that transactors view the early stages of a relationship as a time to learn about one

---

215 For example, the Trading Rules of the National Cottonseed Products Association, Inc. (“NCPA”) prohibit arbitrators from enforcing “terms or conditions not expressly provided for in these Rules,” except where such variant terms are explicitly permitted by the rules. NCPA Rules Rule 1 at 34 (NCPA 1994-95). See also Interview with Association Executive (July 1996) (on file with author). This approach is also adopted by the American Spice Trade Association whose arbitration board excludes contextual considerations and only hears cases based on breach of unmodified association-drafted standard form contracts. Interview with Association Executive (June 1996) (on file with author). Similarly, preliminary research suggests that arbitrators at the Green Coffee Association of New York are strictly unwilling to look to course of performance or unwritten custom of trade. Interview with Arbitrator (April 16, 1998) (conducted by Drew Porter, on file with author) (noting they do not look to custom); Interview with Arbitrator (April 15, 1998) (conducted by Drew Porter, on file with author) (noting that arbitrators don’t look to conduct in interpreting a contract and stating that any variation from the contract would be subject to approval by both parties, and any decision to depart from the contract would require some sort of written proof for arbitrators to take it seriously”); Interview with Arbitrators (April 14, 1998) (conducted by Drew Porter, on file with author) (telling the interviewer there are no unwritten customs). The Association of Food Industries, however, has adopted a slightly less formalistic approach. If a contract is silent, arbitrators look to custom and will give contractual language its ordinary meaning in the trade. However, they do not permit custom to override explicit contractual provisions and will not take courses of performance or courses of dealing at odds with contractual language into account in deciding cases. Interview with Arbitrator (February 1997) (on file with author).

216 See supra, note 170 and accompanying text.

217 See supra, notes 206-207 and accompanying text
another’s reputation, reliability, and business practices. In the cotton industry, for example, where mill-buyers consider the reputation of merchant-sellers to be very important, most mills consider their own past experience dealing with a merchant to be one of the most important sources of reputation information. Mill-buyers explain that to learn about reputation they will typically start with a “lower quantity deal with an intermediate price,” and then see what happens; “if [the merchants] are bad,” or “fuss at them a lot,” then they won’t deal with the person again.

---

218 See Heggblade-Marguleas-Tenneco, Inc. v. Sunshine Biscuit, Inc., 59 Cal. App. 3d 948, 951-52, 131 Cal. Rptr. 183 (1976). (recounting seller’s statements that it would “overplant ‘a little bit’ since this would be their first contract and they wanted to be sure they could produce the quantity needed,” and the sellers “have picked our best land [...]. For we want to be sure to effect the best delivery possible;” also quoting from a letter from the buyer stating that “we would like to start with you on the basis of obtaining 100,000 sacks of Kennebec potatoes from your operation [...] this arrangement would balance our needs with the opportunity for you to perform directly for us and evaluate our relationship. At the same time, it would enable us to do what we have always done in the past—maintain our loyalty to those how have served us well in the past.”).

219 When ten mills were asked how they obtained information about the reputation of merchant sellers, nine mentioned their own past experience with the merchant as a significant source of information, though, as one noted, this method of obtaining information taught “hard lessons.” Mill Questionnaire N° 2 (no date available); id., (“my own experience”); Mill Questionnaire N° 3 (Jul. 28, 1997) (“trade”); Mill Questionnaire N° 4 (Aug. 6, 1997) (“experience mostly”); Mill Questionnaire N° 5 (Aug. 8, 1997) (“past experience”); Mill Questionnaire N° 6 (Jul. 30, 1997) (“experience”); Mill Questionnaire N° 7 (Jul. 31, 1997) (“from prior dealing with him”); Mill Questionnaire N° 8 (Jul. 29, 1997) (“personal experience”); Mill Questionnaire N° 9 (Jul. 29, 1997) (“By doing business with him”); Mill Questionnaire N° 10 (Jul. 28, 1997) (by looking at “our past experience with him”). All Mill Questionnaires are on file with author. Mills explain that they also obtain information about reputation from other sources including, gossip, the trade press, and their bankers.

220 Mill Questionnaire N° 1 (Oct. 7, 1997) (on file with author). Merchants also consider past dealings with a particular mill important to evaluating the mill’s reputation, but because there are many fewer mills than merchants, they tend to rely more on the “word on the street,” specifically Front Street in Memphis, for their information. Yet they too rely on their own experience to assess reputation, but as one warned, “if you have never dealt with the other party before, deal in small quantities.” Interview with merchant (no date available) (on file with author).
Caveat. This account of the emergence of repeat-dealing relationships, though plausible, is also subject to empirical challenge. In markets where buyers and sellers play fixed roles in the chain of production and distribution, buyers as a group and sellers as a group sometimes have divergent views about the content of these aggregate market practices.\(^2\) These divergences may, in fact, be wide enough to impede the emergence of repeat-dealing relationships. However, even in the absence of consensus on the customs relating to all or most aspects of a transaction, cooperation may still arise if there is a common view of the practices relating to a few, preferably core, aspects of trade. If the signal sent with respect to these issues is strong enough, transactors will likely be more flexible in dealing with other aspects of trade that contradict their expectations because they will have greater confidence that their transacting partner is not a defector.

Perhaps more importantly, however, many merchant industries, particularly those with private legal systems, are characterized by institutional rules and features,\(^2\) as well as a

---

\(^2\) See, for example, Bureau of Trade Relations, National Retail Dry Goods Association (“NRDGA”) and SAA, Survey of Trade Practices in the Silk Trade 10 (NRDGA & SAA 1944) (“Retailers report that the average length of time usually consumed in making returns is approximately four days although a substantial number of replies indicate that from seven to ten days is the amount of time usually required [...]. From the manufacturer’s viewpoint, [however], replies indicate that the usual length of time consumed by the retailer in making returns is seventeen days [...]. Both the manufacturers’ and retailers’ replies to this question [about return times] vary from less than one day to thirty days.”). See also Note, loc. cit., note 85, 702 (noting that in the early rayon industry in which “trade customs, which are understood and in part written, have not been accepted by both buyers and sellers [...] there is no uniformity.”).

It is also important to note that in some industries, such as textiles, certain definitions of quality, like “Seconds” and “first quality,” are not even defined in market, regional, or local terms. Rather they are defined in terms of the past output of a particular mill. See Definitions and Trade Customs, WSR 32 (1936).

\(^2\) In some industries, social conditions or past social conditions may also have supported the emergence of commercial relationships, not only by making the fact of defection more observable to market participants through gossip networks that were well-established to meet other needs, such as detecting religious deviance, but also through strong interpersonal relationships that transcended business dealings. These types of forces served to increase the reputation bond posted in each transaction and to promote common
variety of contracting practices, that can best be understood as promoting the emergence and endurance of the type of common knowledge that is important to the emergence of repeat-dealing relationships. The adoption of trade rules might, for example, be understood as an effort to explicitly create this type of common knowledge. Some of these rules explicitly set out acceptable ranges of contractual performance, while others provide focal points around which acceptable ranges of outcomes could arise. As one drafter of the 1986 revisions to the WSRs explained, “The rules provide two businessmen a way to settle an argument by getting insight into what other businessmen think is standard.” Today, some trade rules have themselves come to be considered the customs of the trade.

Moreover, association-sponsored educational efforts to promote knowledge of both trade rules and the content of work-a-day business practices were far reaching, and even today the

understanding. In the cotton industry, this function was served by the culture of honor in the Old South and encouraged by groups like the Cotton Wives Club of the Memphis Cotton Exchange. In the diamond industry, it was served by the close-knit nature of Orthodox Jewish Society. See L. BERNSTEIN, loc. cit., note 171, 130. In the textile industry, closeness was fostered by geographical proximity, see supra, note 83, and social events, some of which, like the textile follies, a musical lampoon, continue to the present day. See Textile Distributors, 55 Women’s Wear Daily, N° 101 at S 36 (May 24, 1988).

Interestingly, codified local customs can also promote cooperation even between transactors in different locations with conflicting codified customs. Suppose that S promises to deliver lumber “promptly” and does so within 10 days. If “prompt” in the B’s locality means five days, he might classify the delivery as deflection if the term is not defined either orally or in the contract. If, however, B makes his objection and the reason for it known to S, and S shows him Trade Rules from her locality defining prompt as 12 days, the B will likely classify the S’s action as cooperation and simply negotiate a more specific clause in the future, rather than treat the delay as deflection and terminate the relationship. If, however, the 12 day practice was unwritten, S’s claim would be far less plausible and relationship breakdown would more likely follow.

See supra, note 185.

J. E. PALMIERI, loc. cit., note 84, 9.

See supra, text accompanying note 100 (noting that the WSRs have become custom); Arbitration Practice Upheld, 3 Silkworm 137-138 (July 1921) (noting that “practically all raw silk is sold under and pursuant to these rules, so that they have come to bear the dignity of a trade custom”).
cotton industry sponsors an introductory nine week course at Rhodes College that includes instruction in “the ethical implications of contracts, contract law, and the importance of contract sanctity; [...] typical trade terms and conditions in the various markets; [...] [and] international trading rules.” Similarly, the NGFA sponsors trade rules seminars, in which participants study the trade rules, read arbitration opinions, work through hypothetical trading situations, and listen to speeches about trade ethics and proper trading practices. In most industries these efforts are and were complemented by extraordinarily active trade presses.

The opinions issued by merchant arbitrators also serve to promote common knowledge. Although no merchant-run arbitration tribunals give formal precedential weight to their prior decisions, the tribunals discussed here, along with many others, nonetheless produced, published, and widely disseminated arbitration opinions. Although these opinions served a variety of functions, perhaps the most important was


228 Similar seminars are run by the Texas Grain and Feed Association. See Nuts & Bolts of the Grain Trade (flyer describing seminar on contracts, trade rules, customs, and arbitration, including panel on “Customs of the Trade,” “Contract Law,” “Arbitration,” and “Trade Rules”). See also Trade Rules as Amended at Des Moines Meeting, 13 WWGT 26 (Nov. 20, 1923-24) (“[F]rom the standpoint of education it can with truth be said that the annual trade rules article is about as valuable a contribution as could be made to the members through their official organ.”).

229 Opinions were viewed as a way of improving the quality of arbitral decisions by forcing arbitrators to articulate reasons for their award. They were also a useful, albeit imperfect, check on certain types of arbitral bias. An arbitrator who signed his name to a widely-circulated opinion that was perceived to be biased risked reputational harm and damage to his own business. In addition, especially in the early days, the production of well-reasoned opinions was one of the ways that the association-run tribunals attempted to establish their own legitimacy. In industries that published the names of the disputing parties in the opinions, the prospect that an opinion would be written was also a way of encouraging settlement. Opinions often noted that even the prevailing party acted improperly, Cook Industries v. Tripple “F” Feeds, NGFA App. Case No 1532 (1977) (affirming the primary arbitration panel’s judgment against the defendant while noting that the “Arbitration Appeals Committee observes regrettable that both parties to the transaction left much to be desired in preparation and performance of the contract.”), thereby giving both parties a reason to settle. See NHA, Report of the Second Annual Meeting 16 (1895)
to educate members of the trade about the content of the rules and the contours of proper business practices, a function that is still regarded as important today. The interpretation and discussion of these opinions in the industry-sponsored trade press furthered and continues to further this norm-inculcation effort.

The membership rules of these associations also facilitate this method of establishing relationships. Most

(suggesting that because the Kansas City Hay Association has its opinions “published in the press [...] the result has been that today we do not have any complaints in Kansas City between shippers and receivers.”)

230 See, for example, the excerpt from the report of C.F. KRAEMER, Chairman of the NAWLA Arbitration Committee in Convention “Highlights,” 3 Service 1 (May 26, 1930) (NAWLA arbitration opinions help “to avoid repetition of disputes through the more careful handling of situations likely to develop into controversies [...] we all profit by our mistakes of the past and when it is pointed out to us where we have erred either by neglecting to fully protect ourselves in the original contract or through some oversight and careless attention to conditions arising subsequent to the contract, we can certainly be careful to guard against repetition. That, we conceive, to be one of the important functions of Arbitration.”). See also SAA Fifty-Fifth Annual Report 35 (1926-27) (“Arbitration has been helpful [...]. It has been found to clarify, in the thought of the trade, the ethics of a situation; and has also, to a certain degree, taught the application of official trade practices as compiled by the several Association divisions.”).

Although merchant arbitrators tended to decide the cases strictly on the basis of Trade Rules and written contracts, in many industries their opinions sometimes note that the prevailing party nonetheless engaged in unacceptable business practices. In addition, while none of the trade rules studied imposed a general duty of good faith and fair dealing, and it was quite rare for arbitrators to take good faith into account, see L. BERNSTEIN, Private Commercial Law, op. cit., note 13 (reporting that at the Memphis Cotton Exchange only three of the opinions written from 1944-90 mentioned good faith), opinions sometimes noted that one party (sometimes even the prevailing party) or both parties acted in bad faith.

231 The NGFA, for example, whose contemporary opinions are among the most detailed and comprehensive published, takes the position that

“most importantly, the [arbitration] decisions serve as an educational tool for the industry by communicating how particular disputes have been resolved. While arbitration decisions are not formal precedent as to subsequent disputes, they are especially instructive to the membership because arbitration cases often involve issues faced by others on a daily basis.” R. C. GORDON, op. cit., note 56, p.57.
associations have substantial membership fees\textsuperscript{232} and strict membership requirements, permitting just a few dissenting votes to block a candidate’s admission.\textsuperscript{233} As a consequence, a transactor who behaved badly in previous deals with an association member is unlikely to be admitted. In addition, by joining an association, a transactor implicitly binds himself to make more aspects of his commercial behavior observable to market participants. For example, most associations require members to arbitrate all disputes with other members in the association’s own tribunals and have a variety of ways of making the outcomes of cases quickly known to association members. Moreover, members who either refuse to arbitrate, or who do not comply with an award, are usually either suspended or expelled from the association and will typically have their names posted on the exchange floor\textsuperscript{234} and/or mentioned in the association’s newsletter along with a description of their wrongdoing.\textsuperscript{235} In most industries these sanctions are severe. Transactors are quite reluctant to deal with someone who has been expelled from an industry association, so these sanctions sometimes put the offending transactor entirely out of business.\textsuperscript{236} The existence of these rules and procedures turns membership into a credible signal that a transactor has a low discount rate since seeking membership, and thereby opening up his future dealings to more exacting scrutiny, would be irrational if he did not.

\textsuperscript{232} See, for example, Interview with Southern Cotton Association Executive (1996) (noting that dues and fees can, depending on the size of the firm, be well over $100,000 per year).

\textsuperscript{233} Southern Cotton Association, By-Laws & Trade Rules Art. 2, §§ 1, 2 (requiring that membership candidates be nominated by five members of the association, that all members be given opportunity to object, that two-thirds approval of the Committee on Membership be obtained, and that two-thirds of the Board of Directors vote in favor); see also Interview with Association Executive (June 1996) (a few objecting members block admission).

\textsuperscript{234} L. Bernstein, loc. cit., note 171, 138-143 (discussing diamond industry nonlegal sanctions).

\textsuperscript{235} See, for example, Ten Cases Cleaned Up: Arbitration Committees Decide Six Disputes—Two are Settled Direct and Two are Expelled, 10 WWGT 39, 40 (Dec. 5, 1920-21) (listing the names of those expelled from the association for failure to arbitrate or failure to pay an arbitration award).

\textsuperscript{236} See L. Bernstein, loc. cit., note 171, 141.
Finally, contracting practices both in merchant industries
governed by private legal systems as well as in those governed
by the Code might also play a role in enhancing cooperation.
Buyers’ and sellers’ confirmations often include terms not
found in the others’ confirmations. These additional terms are
often unenforceable under the Code and may not be
enforceable under many industries’ approaches to the “battle of
the forms.” The standard explanation for these terms is that each
transactor hopes that the other will either not know the terms are
unenforceable or that the other party will not risk violating them
since he does not have the financial resources to defend himself
if they are litigated. Yet these additional terms can also be
understood as setting the terms that the other transactor must
comply with if he wants the contracting relationship to continue.
In some industries, these terms are supplied not in fine print on
contracts, but on posters at the buyer’s place of business, or in
manuals or circulars the buyer produces and distributes to his
suppliers. The terms are, in effect, providing a more efficient
substitute for common knowledge as to the ranges of practices
considered acceptable. These additional terms are really saying
to sellers, “It is all well and good for you to comply with the
contract, that protects you from suit, but if you want to get
repeat business from us, these are the terms you must meet.”

**Conclusion.** In sum, weak-form customs are neither
optimal end-game norms nor, for the most part, optimal RPNs.
Rather, they may, under certain conditions, and at certain stages
of market development, provide transactors with a set of
imperfect, yet nonetheless workable, relationship-creating
norms. The availability of these norms—particularly as
enhanced by trade rules and other institutional features of trade-
association-run private legal systems is particularly important in
markets where reputation information is not, standing alone,
sufficient either to support exchange between strangers and/or
to induce traders to seek the optimal number of transactional
partners. More generally, the divergence between the end-game
provisions contained in trade rules and written contracts, on the
one hand, and the relationship-creating and relationship-
preserving norms that transactors follow in their work-a-day
interactions, on the other, together with the acontextual

---
237 U.C.C. § 2-207.
approach of most merchant tribunals, enables transactors in many industries governed by merchant-run private legal systems to capture the benefits not only of repeat-dealing but also of contracting under a dual set of legally unenforceable relationship-preserving norms and legally enforceable end-game norms, each better adapted to the situations they are designed to govern than either would be to governing the domain of the contracting relationship as a whole. In short, this alternative account of the important role played by weak-form customs in commercial relationships does not depend on courts or other third-party arbiters looking to custom to fill contractual gaps or interpret the meaning of contractual provisions; yet it helps to explain why merchants consider an understanding of customs, at least weak form customs, to be central to the conduct of successful business operations.

Conclusion

The goal of this Article has been to raise, though by no means resolve, questions relating to the desirability of the Code’s incorporation strategy. It has presented evidence that the Code’s conception of widely-known commercial standards and usages that are geographically coextensive with the scope of trade does not correspond to merchant reality but rather is a legal fiction whose usefulness and desirability needs to be demonstrated and defended rather than assumed. Although arguments might be developed to justify the use of this legal fiction as a second-best type solution to problems of gap-filling and interpretation in commercial law, nothing in the legislative history of the Code, or the scholarly literature on commercial law, has yet put forth an adequate justification for the Code’s and the Permanent Editorial Board’s broad endorsement of the incorporation strategy.

Indeed, developing a justification for the Code’s approach would require separate consideration of the wide variety of ways in which incorporation influences commercial adjudication and, perhaps more importantly, ex ante contracting practices. In addition, given the highly local nature of most customs that do exist, any attempt to rehabilitate the incorporation strategy would have to be accompanied by the development of complex and explicit choice-of-custom rules. It would also have to take into account the fact that, as a practical
matter, the incorporation strategy is closer to the pole of a mandatory, rather than a default adjudicative approach, and that the consequences of the approach vary widely in its different and varied uses. For example, the Code’s use of “commercial standards” in determining whether a contract is definite enough to warrant legal enforcement\(^\text{238}\) might reinforce the undesirable tendency of merchants to under-specify their contracts, a tendency that even today merchant associations actively try to combat.\(^\text{239}\) Conversely, the Code provision specifying that when no time is given for an action to be taken under a contract a “reasonable time” is to be implied\(^\text{240}\) may be less problematic. The cost to transactors of specifying time frames for important aspects of contractual performance in the boiler-plate on their confirmations is quite low and might, in fact, be the type of practice we want to encourage through a “penalty default” type of incentive.\(^\text{241}\)

---

\(^{238}\) Id., at § 2-204 cmt.

\(^{239}\) See NGFA, Protecting Your Company’s Interest in Agricultural Commodities 13-16, 19, 30 (1995) (discussing the importance of entering into well-specified contracts and urging transactors to do so). See also supra, sources cited in note 48 (discussing the tendency of hay dealers to underspecify their contracts).

\(^{240}\) U.C.C. § 2-309.

\(^{241}\) A penalty default rule is a rule that is deliberately given content that many or a majority of transactors do not like, in an effort to induce them to reveal information and/or to adopt a provision better suited to their transactions. See Lucian AYRE BEBCUK and Steven SHAVELL, “Information and the Scope of Liability for Breach of Contract!: The Rule of Hadley v. Baxendale», (1991) 7 J. L. Econ. & Org. 284; Ian AYRES and Robert GERTNER, “Filling Gaps in Incomplete Contracts!: An Economic Theory of Default Rules”, (1989) 99 Yale L. J. 87. If, however, courts then vary the meaning of explicit timing provisions by reference to industry practice, the benefits of inducing explicit contracting might not be attainable under existing doctrine.

Although it has been suggested that the incorporation strategy itself might be an effective penalty default, inducing industries that could benefit from a more formalistic approach to opt-out and create private legal systems, it is important to note that the conditions that make opt-out desirable, and the pre-conditions necessary to overcome collective action problems and make opt-out possible, are quite different. As a consequence, industries that might prefer a more formalistic adjudicative approach may not be able to opt-out of the Code.
The battle of the forms situation, in contrast, is one where the Code’s incorporation strategy is problematic, especially if the Hayekian view of custom is incorrect and some customs, while not quite rising to the level of customs of adhesion, nevertheless favor certain types of transactors. Under the Code, trade usage is an important consideration in determining whether additional terms in a confirmation “materially alter”$^{242}$ the offer and are therefore unenforceable or whether they fail to do so and are therefore enforceable unless either the offer limits acceptance to its terms or the inclusion of additional terms amounts to “surprise or hardship.”$^{243}$ If customs favor the stronger party, but do not rise to the level of imposing a hardship and are not unconscionable, they will be enforced. This gives large firms, who will also have better resources to hire experts, the ability to get a significant contracting advantage.$^{244}$

$^{242}$ U.C.C. § 2-207(2)(b) & cmts 3, 4, 5. See, for example, Avedon Engineering Inc. v. Seatex, 126 F.3d 1279, 1284-1285 (10th Cir. 1997) (in deciding whether an arbitration provision in an unsigned textile sales confirmation form was a “material alteration” under U.C.C. §2-207 (which would render it unenforceable), the court looked to see whether the inclusion of an arbitration clause was customary in the trade, explaining that if it was the provision would be enforceable); Suzy Phillips Originals, Inc. v. Coville Inc., 939 F. Supp. 1012 (E. D. N. Y. 1996), affd., 125 F.3d 845 (2d Cir. 1997) (finding a limitation of liability clause in the seller’s confirmation to be “standard trade practice,” as evidenced by the WSRs, the court held that it therefore neither materially altered the terms of the offer nor inflicted “unreasonable surprise”); Graphics v. Peck Industries, Inc., 304 S.C. 101, 403 SE.2d 146, 150 (SC App. 1991) (same).

$^{243}$ U.C.C. § 2-207 cmt 4. See also Wilson Fertilizer & Grain, Inc. v. ADM Milling Co., 654 N.E.2d 848, 852-854 (Ind. App. 1995). In holding that an arbitration provision in buyer’s order did not materially alter the offer so as to render the arbitration provision unenforceable, the court explained that the Code’s

“Comments suggest that hardship or surprise may be created by terms that deviate from customary trade standards and practices, but may not be created by terms that operate within the accepted norms of the parties’ particular trade [...] while evidence of usage of trade and course of dealings are not conclusive on the question of surprise [...] such proof is significant to the issue [...]. In short, a party should not be surprised to find in a confirmation a clause of a type that is customarily used within the trade, whereas the clause is an unreasonable surprise where it represents an unreasonable or harsh deviation from custom.”
Determining the desirability of other provisions that rely on incorporation is somewhat more complicated. For example, when the issue is whether usage should be admissible to show acceptable customary ranges of quality, it may be that a penalty-type approach to inducing parties to include more detailed quality specifications in their agreements might work. If, however, it is difficult to describe a quality standard that a nonexpert court will be able to apply, it might be better to encourage transactors to include a quality-determining wiseman provision in their contract. Whether the inclusion of wiseman provisions in appropriate circumstances is more likely to be induced by a default rule that enforces every clause precisely as written, or one that interprets written clauses by customary practices, is just one of many open questions relating to the Code’s incorporation strategy that are in need of further exploration.

More generally, it is important to note that one of the primary justifications for looking to custom to fill gaps and interpret contracts is that the customs themselves provide useful information to generalist judges about the intent of the parties, or failing that, about a range of practices whose widespread use suggests that they are viewed as reasonable by industry transactors. However, recognizing that the customs often evolve to govern situations where transactors trust one another and want to continue dealing, but that cases arise when the very trust that makes the custom workable has broken down, suggests that there is no reason to suppose that customs will provide useful information about what contracting parties would have agreed to had they included a provision stating how the matter at issue was to be dealt with if third party adjudication were required. In addition, because so many customs have an implied precondition that the transactors trust one another, and because many customs condition on this and other information that is inexpensively observable, but that may be quite expensive to verify (that is, to prove to a court with reasonable accuracy), even customs that are widely followed may be poor candidates for judicial enforcement. First, from an \textit{ex ante} perspective,
transactors are unlikely to want to spend an infinite amount of money establishing their case if a dispute arises. And second, third-party application of these types of customs may well be error prone. Transactors may therefore prefer a third party to apply a very different rule, one that may be less well-tailored to their relationship, but easier and less costly to dispute under should a problem arise.

In sum, while the fact that the Code is built on a highly questionable empirical basis does not necessarily mean that its drafters erred, it does mean that it is necessary to inquire into whether a justification for the incorporation strategy can be constructed that takes merchant reality as its starting point. This inquiry is particularly timely, not only because the revised Code—which extends the incorporation principle—will soon be submitted to the States for their approval, but also because proposals to create commercial law in transition economies often take as their starting point the desirability of codifying existing commercial practices.

---

245 Even if in merchant industries that adopted trade rules customs arose as ranges around these focal point rules, this would still not justify the incorporation strategy. The industries that have adopted such rules are, for the most part, precisely those that have also chosen to opt-out of the public legal system, making it much less common for their disputes to wind up being governed by the Code in the first place.