Volume 36
Number 1
2017

QUINNIPAC
LAW REVIEW

Contents

Articles
Food and Fertile Ground: Improving Chinese Food Safety Through Environmental Regulation
  Brent Domann  1

Justice in Full is Time Well Spent: Why the Supreme Court Should Ban Sua Sponte Dismissals
  Michael J. Donaldson  25

The Paradox of Military Tribunals That Try Civilians
  Golan Luzon  53

A Sound Taxonomy of Remedies
  Tim Kaye  79
Articles

FOOD AND FERTILE GROUND: IMPROVING CHINESE FOOD SAFETY THROUGH ENVIRONMENTAL REGULATION

Brent Domann*

I. INTRODUCTION ...................................................................................... 2
II. CHINESE FOOD SAFETY ........................................................................ 3
   A. Food Safety Goals ..................................................................... 3
   B. History of Food Safety in China ................................................ 5
      3. 1993–2015: Regulation of a Single but Segmented Industry .............................................................. 10
   C. Current Regulations and their Effectiveness ....................... 11
III. CHINESE INDUSTRIALIZATION AND ENVIRONMENTAL PROTECTION ...... 12
    A. History of Industrialization in China ......................................... 12
    B. Urbanization and the Bifurcation of Chinese Demographics 14
    C. Environmental Quality Problems and their Effects on Food Supply .......................................................... 15
IV. STATEMENT OF THE PROBLEM AND SOLUTION .......................... 19
V. COUNTERARGUMENTS ....................................................................... 21
VI. CONCLUSION .................................................................................... 23

* Reference Librarian and Adjunct Faculty, Global Food Law Program, Michigan State University College of Law. The author would like to thank Michael T. Roberts for his invaluable feedback and input through multiple drafts of this article and Jane Meland for her feedback and advice. Edison Tang and Yuan Liu also provided very helpful perspectives for which the author is very grateful. Any errors in this article are the sole responsibility of the author.
I. INTRODUCTION

Food safety in China has a mixed reputation. Although it has a brief history, it has made great strides in incorporating new solutions to classic problems of sufficiency, security, and safety. Still, it is imperfect. From tainted baby formula to contaminated strawberries, both domestic and international Chinese food scandals have garnered a lot of attention. From the Chinese government’s perspective, faith in Chinese food safety is important for economic reasons, as foreign importers seek Chinese food. Additionally, the Chinese government’s domestic reputation and political security depend on a populace that trusts the efforts of the sovereign state.

As its consumer class rapidly expands, China is building an increasingly effective and detailed model for food regulation. Issues of food safety, food security, and accuracy in representing the contents and methods of production of food items are moving to the front and center of both consumer expectations and regulatory attention.

---


2 See GUANQI ZHOU, THE REGULATORY REGIME OF FOOD SAFETY IN CHINA: GOVERNANCE AND SEGMENTATION 4 (2017) (noting that the current regulatory scheme has only been around since 1950, but has stabilized and institutionalized food safety in China); see also Fu, supra note 1 (discussing China’s commitment to strengthening regulations and oversight across the entire supply chain); see also FRANCIS SNYDER, FOOD SAFETY LAW IN CHINA: MAKING TRANSNATIONAL LAW 474 (2015) (“China has made tremendous strides since the 1995 Food Hygiene Law in improving its system of food safety regulation . . . .”).

3 For an in-depth discussion of Chinese food scandals, see infra Part II(c).

4 See Fu, supra note 1 (“Improving food safety in China is also important for international consumers because food and ingredients from China can be found on supermarket shelves all over the world.”).

5 Id.

6 See Kim Iskyan, China’s Middle Class is Exploding, BUS. INSIDER (Aug. 27, 2016, 9:09 PM), http://www.businessinsider.com/chinas-middle-class-is-exploding-2016-8 (discussing China’s rapidly expanding middle class and the growing rate of Chinese consumption); see also John Balzano, Revised Food Safety Law in China Signals Many Changes and Some Surprises, FORBES ASIA (May 3, 2015, 11:06 PM), https://www.forbes.com/sites/johnbalzano/2015/05/03/revised-food-safety-law-in-china-signals-many-changes-and-some-surprises/ (discussing how the revision of the Food Safety Law has the ability to make China’s regulation of food safety more effective).

7 See Andrew Sim, China: An Overview of the New Food Safety Law, FOOD SAFETY MAG. (Apr. 19, 2016), https://www.foodsafetymagazine.com/eneWsletter/china-an-overview-of-the-new-food-safety-law/ (“[The new Food Safety Law] shows the proactive attitude of the Chinese government in cracking down on food scandals that have affected the country in
China is not, however, just an urban, consumption-based economy. Despite an incredible migration toward urbanization and transition to manufacturing, a rural population remains; additionally, there are citizens that live in rapidly urbanizing areas that nevertheless employ some traditional means of food production by growing their own food supply. These rural and semi-rural citizens benefit from food regulation in the aggregate, but are particularly prone to suffering the negative results of regulatory failures elsewhere in China’s economy. More specifically, China’s comparatively lax regulation of environmental quality has a direct impact on the food supply for those who eat what they grow. Without a defined and symbiotic relationship between food safety and environmental regulations, China’s regulatory scheme for the food supply can never be complete.

This paper explores the historical development of food safety regulation in China and highlights how the bifurcation of urban and rural populations leaves a regulatory gap that limits the efficacy of food safety regulation. Part II discusses food safety in China as the nation’s politics have evolved over the last century. Part III surveys China’s industrialization and urbanization, examines the bifurcation of urban and rural populations, and discusses current environmental issues. Part IV defines the resulting problem and suggests a solution: unless environmental regulations related to food-producing natural resources are a part of China’s regulatory scheme, consumer-focused food safety regulation cannot protect those who produce their own food. Part V considers counterarguments before this paper concludes in Part VI.

II. CHINESE FOOD SAFETY

A. Food Safety Goals

Food safety laws are designed for several purposes: safety, security, and public health and economics (as a paired set of concerns). Each of
these purposes, or pillars, benefits consumers in a unique way and, as a whole, serve a holistic role in fostering an optimal food marketing or distribution scheme.

An ancient concept for healthy consumption, food safety includes, “many facets of handling, preparation and storage of food to prevent illness and injury [including chemical], microphysical and microbiological aspects.” The focus of a food safety regime is to avoid injury or sickness as a result of contamination, improper handling, or adulteration. In pursuit of these aims, major attention is given to labeling, hygiene, and chemicals (such as additives and pesticides), with even more attention given to providing for safe food delivery and preparation.

Food security relates to food’s availability to a population. Put another way: to provide for greater food security, an entity must ensure that people have enough food available and secure access to that food.

Public health and economics are two major considerations that, together, inform food safety laws and serve as goals for those laws. Public health is a fundamental goal and primary responsibility for government. Governments with healthy, well-supplied populations

---

10 Takashi Uemura & Md. Latiful Bari, History and Safety of Food: Past, Present and Future, in FOODBORNE PATHOGENS AND FOOD SAFETY 2–3 (Md. Latiful Bari & Dike O. Ukwu eds., 2015) (“[F]ood safety . . . is probably nearly as old as human history itself and may have started with the recognition and subsequent avoidance of foods that were naturally toxic.”).
12 Id.
14 FOOD AND AGRIC. ORG. OF THE UNITED NATIONS, TRADE REFORMS AND FOOD SECURITY: CONCEPTUALIZING THE LINKAGES 29 (2003) (“Food security exists when all people, at all times, have physical, social and economic access to sufficient, safe and nutritious food which meets their dietary needs and food preferences for an active and healthy life.”).
15 Food safety laws help avoid foodborne illnesses and food security laws help avoid malnutrition and undernutrition. See Irene B. Hanning et al., Food Safety and Food Security, NATURE EDUC. KNOWLEDGE (2012), https://www.nature.com/scitable/knowledge/library/food-safety-and-food-security-68168348 (last visited Sept. 23, 2017). Economics are implicated especially in socialized economies such as China’s, where public health is subsidized and the State may also have financial stakes in food production enterprises; “soft financial constraints” and other economic pressure may even be used as an enforcement mechanism. Peng Liu, Tracing and Periodizing China’s Food Safety Regulation: A Study on China’s Food Safety Regime Change, 4 REG. & GOVERNANCE 244, 247–50 (2010).
16 Michael T. Roberts, Role of Regulation in Minimizing Terrorist Threats Against the
enjoy the benefits of greater popular support. Economically, strongly established infrastructure and trade channels better equip populaces when faced with droughts, shortages, or military needs. In socialized states, these goals and their benefits (and risks) are magnified: public health becomes not just a consideration of the government, but a consideration of primarily, or even solely, the government. Similarly, although other government types feel pressure to establish economic efficiency, socialized states need to incorporate and perfect such efficiency to ensure their own budget goals are met.

B. History of Food Safety in China

China’s history with food law stretches back thousands of years. Ancient texts highlighted food safety perhaps as early as the 10th century BC. I Ching, a “classic” text of major historical importance, states that through “firm correctness” in the preparation of food, good luck will be preserved. Confucius set special diet restrictions during purification periods, which included refusing to eat rancid fish, spoiled meat, anything looking or smelling bad, out-of-season food, and the like. The Tang Dynasty (7th to 10th centuries AD) saw refrigeration,

---

\(^{17}\) See Anna Lora-Wainwright, Dying for Development: Pollution, Illness and the Limits of Citizens’ Agency in China, 214 CHINA Q. 243, 245 (2013) (“It is not only physical health that is affected but also the texture of communities and relations between individuals, communities and the state. . . . [i]t is not solely an economic question over where resources to mitigate these [pollution-related] risks should be drawn from, but also a deeply moral question.” Further, highlighting the deep political connection between moral responsibility and sovereign authority, “[i]f state agents are perceived to be responsible for reducing pollution and yet are unresponsive or ineffective at doing so, citizens’ trust in the state and its legitimacy may be severely threatened.”).

\(^{18}\) See Timothy McDaniels et al., Fostering Resilience to Extreme Events Within Infrastructure Systems: Characterizing Decision Contexts for Mitigation and Adaptation, 18 GLOBAL ENVT’L CHANGE 310, 310–11 (2008).


\(^{20}\) See MICHAEL NYLAN, FIVE CONFUCIAN CLASSICS 218 (2001) (noting that “[s]ome modern scholars” believe the text may have existed that early as a collaborative work, but that “it was [not] stabilized [until] the first or second century AD.”); see also RICHARD J. SMITH, THE I CHING: A BIOGRAPHY 4 (2012) (“The Changes [or I Ching] first took shape about three thousand years ago . . .”).


\(^{22}\) CONFUCIUS, ANALECTS 98 (David Hinton trans. 2014). Confucius seems to have been especially picky about meats—in addition to avoiding rancid meats, he also limited his intake,
salting, brining, pickling, and even jerky-making.\textsuperscript{23} The Song Dynasty (10th to 13th centuries AD) encouraged the avoidance of fly-contaminated food and recommended boiling water.\textsuperscript{24} Food safety measures continued to appear in various instances up until the fall of the Republic of China in 1949.\textsuperscript{25}

Although food safety regulation might be understood in a more contemporary context, the story of China’s food regulation begins its modern journey in 1949.\textsuperscript{26} The development of China’s food safety laws and regulations has been evolutionary, with many steps along the way.\textsuperscript{27} For convenience’s sake though, it may be best to break China’s regulatory development in this area into distinct periods.\textsuperscript{28} For instance, Peng Liu breaks these into three “regimes” since the revolution in 1949: a command regime (1949–1978), an intermediate regime of mixed instruments (1979–1992), and a new regime of regulatory governance (1993–ongoing).\textsuperscript{29} Linhai Wu and Dian Zhu prefer to discuss each major enactment of legislation or regulation as a period unto itself.\textsuperscript{30} Still, others use events like the Korean War,\textsuperscript{31} Cultural Revolution,\textsuperscript{32} economic reforms of the late 1970s,\textsuperscript{33} and the shift to much more detailed regulatory language in 1995\textsuperscript{34} as markers. A brief history of this

\textsuperscript{23} CHARLES D. BENN, CHINA’S GOLDEN AGE: EVERYDAY LIFE IN THE TANG DYNASTY 126–28 (2002). Benn describes imperial ice pits, but also states that poor farmers without access to ice could preserve fruits and vegetables for use throughout the winter season by digging trenches or creating underground storage units. Id.
\textsuperscript{24} Yu & Liu, supra note 19, at 6.
\textsuperscript{25} Id.
\textsuperscript{26} ZHOU, supra note 2, at 71 (describing the beginning of food safety regulation in the People’s Republic of China).
\textsuperscript{27} See LINHAI WU & DIAN ZHU, FOOD SAFETY IN CHINA: A COMPREHENSIVE REVIEW 170 (2015).
\textsuperscript{28} See, e.g., Peng Liu, supra note 15, at 246. For examples of other periodization schemes, see WU & ZHU, supra note 27, at 170–72; Yu & Liu, supra note 19, at 4–9.
\textsuperscript{29} Peng Liu, supra note 15, at 244, 247, 249, 252 (a summary appears in Liu’s abstract and regime headings appear on following pages; note that summary dates and heading dates are not exact, but are reasonably close).
\textsuperscript{30} WU & ZHU, supra note 27, at 170–72.
\textsuperscript{31} Chenglin Liu, The Obstacles of Outsourcing Imported Food Safety to China, 43 CORNELL INT’L L.J. 249, 281 (2010).
\textsuperscript{33} Id.
\textsuperscript{34} Chenglin Liu, supra note 31, at 283.

1. 1949–1977: From Revolution to Economic Reform

In 1949, the People’s Republic of China was founded as Communist authorities consolidated power at the end of a civil war. Early in this political shift, the population soared and starvation was a real concern; the focus of the Chinese government in food issues was one of quantity. Most food safety incidents were poisonings from the consumption of unsafe food (likely due to hygiene issues) because of scarcity in the food supply.

Early food safety regulation was modeled after Soviet Union systems. Rudimentary agencies were defined and given tasks—including food hygiene—through a special regulation in 1954, and by 1956, agency stations were present nation-wide. A more developed regulatory framework was in place by 1959. Though formal regulations beyond the tasks of early agencies were sparse (and food hygiene was only a minor function), enforcing standards was not impossible; tainted meat sold to Chinese troops in the Korean War resulted in the execution of factory owners involved, sending a “powerful message” to food producers.

In 1956, the Chinese government, in an effort to moderate rapidly changing political and economic circumstances, decentralized control...
over economic planning. This move increased pressure on agriculture through a reduction in labor and available land, and contributed to China’s devolution into the Great Leap Forward. A massive famine ensued, killing tens of millions of people. In 1965, several ministerial bodies in China collaborated to create the *Administrative Regulations on Food Hygiene for Trial Implementation*. This was the first regulation dealing with food hygiene after the Chinese Revolution. Its implementation marked the first formal effort at regulating food hygiene and helped to transition private food factories to state control, maintaining tight tolerances for safety despite low productivity. Enforcement was flexible, however, as China had not forgotten about its recent famine.

China’s Cultural Revolution, beginning in 1967, resulted in the abrogation or suspension of existing law and efforts to reform it were abandoned; economically, however, agricultural production was not disrupted. A working—if inadequately productive—system was in place. Regulations specifically regarding food content or production safety would not appear for several years, as tight state control left poor consumption habits as the main reason for food safety incidents in the following decade.

This early period saw the Chinese government dealing with the repercussions of major political and economic change, and food supply frequently took priority over hygiene or safety from a regulatory standpoint. As external pressures began to stabilize, China saw the implementation of its earliest formal food safety regulations. A new period approached, however, and it would help bring to light the need for food safety regulation as a consolidated enterprise.

---

44 *Michel Aglietta & Guo Bai, China’s Development: Capitalism and Empire* 55 (2013).
45 *Id.*
46 *Id.;* Hon-Ming Lam et al., *Food Supply and Food Safety Issues in China*, 381 LANCET 2044, 2044 (2013); Chenglin Liu, *supra* note 31, at 282.
47 *Wu & Zhu, supra* note 27, at 170.
48 *Id.; see also* Chenglin Liu, *supra* note 31, at 282.
49 *Chenglin Liu, supra* note 31, at 282; *see also* *Wu & Zhu, supra* note 27, at 170.
50 *See* Chenglin Liu, *supra* note 31, at 282 (stating that “[the first regulation] did not set forth requirements for food content because China was still recovering from a catastrophic famine”).
51 Roberts, *supra* note 32, at 4; *Aglietta & Bai, supra* note 44, at 56 (for date information).
52 *Aglietta & Bai, supra* note 44, at 56.
53 *Peng Liu, supra* note 15, at 248–49.
54 *Wu & Zhu, supra* note 27, at 170.
55 *Id.*

China’s food supply issues did not disappear with the end of its famine in 1961, nor did they end after the Cultural Revolution. Instead, they would continue until economic reforms were implemented in the late 1970s. Mao Tse Tong died in 1976 and, in 1978, China instituted reformed economic policies that would lead to rapid growth in its food industry. In addition to production growth and ownership diversification, China also saw an increase in the number of restaurants. In the midst of this growth, the existing food safety regime and its agencies grew less effective as its bureaucratic structure failed to keep pace with a growing and diversifying industry.

The existing food safety regime’s waning effectiveness resulted in an increase in food safety incidents throughout the early 1980s. As a result, China passed the Food Sanitation Law in 1982, which took provisional effect in 1983. This law had a wider scope than the 1965 regulation and included standards regarding food content and additives, among others. Although subject to “shortcomings and limitations” regarding regulatory power, messy enforcement across disparate agencies, and economic pressures, this transitional period saw an early model of testing and updating food safety regulation in anticipation of Chinese economic growth both domestically and internationally.

57 See id. (“[A]verage rural diets continued to fall short of basic nutrition standards . . . . food supplies for millions of Chinese villagers were no better in the 1970s than in the 1930s.”) (citation omitted).
58 AGLIETTA & BAI, supra note 44, at 56.
59 Peng Liu, supra note 15, at 249 (noting an average annual growth of 9.3% in China’s food industry gross product between 1979 and 1984); see also Brandt & Rawski, supra note 56, at 8 (“Everyone recognized the death of Mao [Tse Tong] in 1976 as a major turning point for the People’s Republic.”).
60 Peng Liu, supra note 15, at 250.
61 See id. (describing the shift from a State-owned food industry to a “diverse ownership structure” during period of economic reform, causing the structure to become “steadily weaker”).
62 Id.
63 Id.; WU & ZHU, supra note 27, at 170–71 (using the term “Food Hygiene Law,” though other sources refer to it as the “Food Sanitation Law”).
64 Chenglin Liu, supra note 31, at 282.
65 Peng Liu, supra note 15, at 251.
66 See, e.g., id. at 249–52 (noting that the Food Sanitation Law, or Food Hygiene Law, was passed during a time of transition and economic reform, which prompted the utilization of new policy methods, such as the creation of a national food hygiene regulation system).
3. 1993–2015: Regulation of a Single but Segmented Industry

A further update to the regulation of food safety came in 1995, when the Food Sanitation Law designated a new enforcement agency, the Ministry of Health, and led to the implementation of over 500 rules and regulations. This new law was the culmination of several efforts. First, food industry enterprises grew significantly in the decade after the implementation of the Food Sanitation Law in 1983 and were increasingly independent from government ownership or control. Second, the Chinese government proposed the establishment of a market economy in 1992, which also included institutional reform. As a result, these institutional reforms eliminated several of the industrial ministries that had maintained control over some aspect of food safety regulation dating back to the 1950s.

This elimination of (some) industrial ministries forced the centralization of regulatory power in a market that had shifted towards a decentralized group of producers. Put another way, where many State-owned producers were previously regulated by respective industrial ministries or similar entities, the new paradigm revolved around a wide array of private producers that could be regulated by a central, “third-party” governmental authority. There were a few adjustments made in the following years, but they did not demonstrate any major shifts in policy. The adjustments did, however, begin to segment the regulatory system by focusing certain departments’ supervision on particular stages of food production, like agricultural production, industrial food processing, food circulation, and food consumption.

The 1995 law started to show its limitations in the new segmented regime in the wake of a high-profile scandal in the early 2000s.

---

67 Chenglin Liu, supra note 31, at 283.
68 Wu & Zhu, supra note 27, at 171 (using the term “Food Hygiene Law,” though other sources refer to it as the “Food Sanitation Law”).
69 Id.
70 See Peng Liu, supra note 15, at 252.
71 See id. (describing centralized government regulation reforms of food industries that had been previously decentralized and governed by local authorities).
72 Id.
73 See Wu & Zhu, supra note 27, at 171 (describing institutional reforms in 1998 and 2003 that adjusted administrative responsibilities in several ways).
74 Id. at 171–72.
75 See id. at 171 (discussing the milk powder scandal in Fuyang, Anhui); see also Peng Liu, supra note 15, at 253 (“The milk powder scandal in Fuyang, Anhui in 2003 and 2004 showed that major problems remained in China’s food safety regime.”).
Following lengthy deliberations, but likely hastened by another major milk scandal in 2008, the Food Safety Law of 2009 was a much broader and more detailed attempt at food safety regulation. It invoked the many levels of Chinese government, from the township, up to the central government in Beijing, and included expanded standards and enforcement mechanisms.

C. Current Regulations and their Effectiveness

China’s food safety laws, despite their rapid development and strong mandates, are imperfect. Several major scandals have hit Chinese food producers in recent years, including problems with adulterated food, counterfeit food and drugs, and contaminated food. The following two examples help to illustrate a major issue.

In 2008, Chinese-manufactured baby formula was contaminated with melamine. Several babies died and nearly three hundred thousand were sickened by the tainted milk. The baby formula, according to the World Health Organization, was diluted to save on cost and then adulterated with melamine to appear protein-rich enough to pass quality control checks.

Another major scandal involved the sale of contaminated strawberries to schoolchildren. This scandal resulted in the sickening of over eleven thousand children in Germany in 2012. In this case, a catering firm supplied food made from a batch of the frozen fruit that

76 Peng Liu, supra note 15, at 253.
77 Chenglin Liu, supra note 31, at 283 (referring to a 2008 melamine-tainted milk scandal discussed infra Part II(C)).
78 See id. at 283–85 (describing China’s hierarchical government and the role local and national government play in food safety regulation); see also Austin Ramzy, Will China’s New Food-Safety Laws Work?, TIME (Mar. 3 2009), http://content.time.com/time/world/article/0,8599,1882711,00.html.
79 Chenglin Liu, supra note 31, at 284–89.
81 Branigan, supra note 80; Chenglin Liu, supra note 80, at 372.
82 Chenglin Liu, supra note 80, at 371–72.
was contaminated with norovirus. 84

What these scandals (and many others) have in common is that they occurred in settings involving commercialized food. The formula in 2008 and the strawberries in 2012 were food products manufactured (grown, processed, etc.) for commercial sale to domestic and international food consumers. Chinese food regulation is focused very heavily on alleviating consumer concerns through addressing the production and distribution of commercialized food directly. 85 Food grown in gardens or otherwise at home, however, is not and has not been the focus of food regulation in China. Instead, consumers of home-grown food are at the mercy of their own production or cultivation techniques and the limits and dangers of the natural resources they have at their disposal.

III. CHINESE INDUSTRIALIZATION AND ENVIRONMENTAL PROTECTION

A. History of Industrialization in China

Like its development in food safety, China’s industrialization is readily understood to have come in periods or steps. 86 Though the histories of food safety and industrialization are interesting to explore independently, they are tightly intertwined. 87 Early modes of food regulation implicated light industry, and the political and economic pressures to build industry had effects on food safety regulation. 88 Recognizing the interplay between food regulation and industry is 84 Id.
85 See Roberts, supra note 80, at 407–08 (discussing the Food Safety Law’s aim to build trust between government and consumers, especially by enabling anyone to report and act during production and trade that violates the law).
86 See, e.g., Xu Nan, A Decade of Food Safety in China, in FOOD SAFETY IN CHINA 1 (Zhou Wei ed., 2012) (“Researchers tend to divide China’s food-safety history since 1949 into four stages.”).
87 For instance, the Provisional Food Hygiene Control Ordinance utilized a control regime that was shared by the industrial ministries and health agencies. Peng Liu, supra note 15, at 249; see also AGLIETTA & BAI, supra note 44, at 56 (discussing the interaction between industrial and agricultural growth).
88 See Peng Liu, supra note 15, at 247 (detailing light industry sectors and ministries that regulated aspects of food safety); see also AGLIETTA & BAI, supra note 44, at 55–56 (“[R]ural communes were restructured on smaller groups of households and inefficient rural factories were cut back. Meanwhile, industrialization was focused on the creation of an industrial base in inland provinces for military reasons.”); CHRIS BRAMALL, THE INDUSTRIALIZATION OF RURAL CHINA 10 (2007) (“[F]rom a purely military point of view, it was essential to develop a new industrial base capable of meeting China’s defence and other economic requirements in the [chiefly agricultural and pre-industrialized] hinterland.”).
crucial to understanding how volatility in one can alter the effects of the other on populations.89

After an initial expectation of slow development, China began to speed up the transition to a completely socialist economy following its involvement in the Korean War.90 By 1956, additional policy changes resulted in “full nationalization of private businesses.”91 This approach treated agriculture and industry differently92 and created a long-lasting tension between the two.93 By 1958, an attempt to moderate dramatic and unsteady growth resulted in the Great Leap Forward; in this period the move to industrialize rural areas had massive negative effects on the food supply.94 By 1960, starvation was a major issue and famine remained until 1962.95 Despite agricultural limitations, a move towards industrialization continued, especially in rural areas, which continued to serve both the modernization of agricultural sectors of the economy and the widespread rooting of industrialized areas for defense purposes.96 The Cultural Revolution began in 1967 and slowed this industrial push, but tensions between agricultural growth and industrialization resurfaced by its end in 1969.97

These tensions contributed to political struggles in China until Mao Tse Tong’s death in 1976.98 Shortly after, China’s 1978 economic reforms ushered in a new period of growth.99 China’s rapid

89 See, e.g., Xizhe Peng, Demographic Consequences of the Great Leap Forward, 13 POPULATION & DEV. REV. 639, 639 (1987) (“Heavy industry, especially steel production, was accorded high priority at the expense of agriculture and light industry . . . . Millions of peasant laborers moved into cities to work in factories. In the countryside the formation of people’s communes was praised as a ‘golden bridge’ toward communist society. Unfortunately, nothing worked as expected.”).
90 See AGLIETTA & BAI, supra note 44, at 54 (explaining that China’s difficulty in establishing a currency caused the Communist Party to anticipate a slow transition to Soviet-style industrial planning).
91 Id.
92 Id.
93 Id. at 55 (“[D]ecentralization of planning rights facilitated irrational investments which dramatized the stresses between agriculture and industry.”); AGLIETTA & BAI, supra note 44, at 56 (“Industrial growth was much too fast for the rigid capacity of food supply.”).
94 See id. at 55 (describing the Great Leap Forward as “one of the most catastrophic experiences ever”); see also supra notes 44–47 and accompanying text.
95 AGLIETTA & BAI, supra note 44, at 55.
96 See id. at 56; see also BRAMALL, supra note 88, at 9–11.
97 AGLIETTA & BAI, supra note 44, at 56 (“After [the Cultural Revolution’s] end in 1969, the imbalance between industrial and agricultural growth re-emerged.”).
98 Id.
99 See id. (“Although Mao’s era was often used to contrast with the reforms after 1978, it actually laid down crucial foundations for the later dazzling economic performance.”).
industrialization and its tensions with food safety began to show visible effects on population demographics: China’s economic development from the late 1970s forward has increased both wealth and poverty simultaneously. As a result, significant portions of the Chinese population have been both economically and politically marginalized. This marginalized population, as will be demonstrated in Part III(B), is the group that suffers from incomplete food safety regulation in China’s current system.

B. Urbanization and the Bifurcation of Chinese Demographics

Industrialization’s major role in China’s economic expansion has had significant effects on the use of land. Mostly in the form of commodification through urbanization, land is being increasingly utilized by “capital-intensive” enterprises instead of the collectives and individuals that farmed them in the past. Farmers till dwindling acreage due to a lack of off-farm employment activities. Additionally, growing farming enterprises are designed to feed urban populations; related policy-driven land commodification serves to improve agricultural efficiency at the expense of reduced food self-sufficiency. In short, rural farmers are surviving on less and less.

“[M]ultiple bifurcations of policy” have resulted in the economic

---

101 Id.
103 Id.
104 Id.
105 See Siciliano, supra note 102, at 167 (“This process of urbanization of populations . . . is closely linked to a commodification of the land resource.”).
106 Id.
107 Id.
108 See Lu et al., supra note 102, at 1710.
110 Ying Liu et al., The Bittersweet Fruits of Industrialization in Rural China: The Cost of Environment and the Benefit from Off-Farm Employment, 38 CHINA ECON. REV. 1, 7–8 (2016).
111 Siciliano, supra note 102, at 176.
112 AGLIETTA & BAI, supra note 44, at 54.
division of China’s urban-rural population. As its economy develops and its agricultural policies reduce farmers’ resources, China faces increasing gaps in income between urban and rural populations.\textsuperscript{110} While the Chinese’s government’s priority is to reduce income gaps, the current situation remains one of inequality.\textsuperscript{111} As a result, rural populations continue to rely on very small-scale farming without the option to improve personal economic conditions, unless they move to towns or cities.\textsuperscript{112} Historically treated as low-status, industrialization has impacted these rural populations more negatively than it has impacted urban residents.\textsuperscript{113} These populations have limited economic and political power\textsuperscript{114} and, therefore, suffer the effects of policies that cannot be enforced completely across populations and geographical areas.\textsuperscript{115}

C. Environmental Quality Problems and their Effects on Food Supply

China has been home to rapid industrialization, urbanization, and economic growth for several decades, and environmental quality problems have grown as well.\textsuperscript{116} The relationship between environmental regulation and industry growth has been a popular academic focus.\textsuperscript{117} In this specific context, differing relationships have been hypothesized: either lax environmental regulation has played a part in this growth,\textsuperscript{118} or exploding pollution in air, water, and on land has

\textsuperscript{110} Siciliano, \textit{supra} note 102, at 167.
\textsuperscript{111} See id. (detailing income discrepancies and discussing the potential for urbanization to reduce income gaps); Prändl-Zika, \textit{supra} note 106, at 236–37 (noting that rural incomes and living standards “[lag] far behind the prosperous development in the cities and coastal areas” and are at “somewhat above subsistence level”).
\textsuperscript{112} Prändl-Zika, \textit{supra} note 106, at 237.
\textsuperscript{113} Liu et al., \textit{supra} note 107, at 8.
\textsuperscript{114} Craig G. Smith, \textit{Chinese Farmers Rebel Against Bureaucracy}, N.Y. TIMES (Sept. 17, 2000), http://www.nytimes.com/2000/09/17/world/chinese-farmers-rebel-against-bureaucracy.html; see also Hong, \textit{supra} note 100, at 108 (describing “Chinese farmers’ lack of rights to landed property” and the fact that they “have little or no say in the decision-making process [regarding property]”).
\textsuperscript{115} See Liu et al., \textit{supra} note 107, at 1–2 (outlining asymmetrical policymaking, enforcement, and monitoring of environmental issues between urban and rural areas).
\textsuperscript{116} See \textit{supra} Parts III(A) and III(B).
\textsuperscript{117} For example, Wang and Shen discuss the “Porter Hypothesis,” which states that properly written and enforced regulations could benefit both the environment and private industry. This, they note, is counter to a more popular understanding that regulations have a negative impact on industry growth. Yan Wang & Neng Shen, \textit{Environmental Regulation and Environmental Productivity: The Case of China}, 62 RENEWABLE & SUSTAINABLE ENERGY REV. 758, 759 (2016).
\textsuperscript{118} See, e.g., Hua Wang et al., \textit{Environmental Protection in China, in Pollution in China} 1 (Michael I. Chang ed., 2011).
been a result of it. Some have also pointed to corruption as a negative contributor to the state of Chinese environmental regulation. In any case, China faces major environmental issues in the midst of its industrial and urban development.

While China’s economy has developed over the last thirty years, its growth has contributed to a concerning deterioration in environmental conditions. It has been stated that “the foundations of effective environmental protection are a country’s legal regime and its implementation.” Other pressures from China’s massive economic transition, however, have produced mixed results as growth has created new environmental stresses and reduced regulatory capacity.

Water quality has deteriorated for decades as rapid economic growth polluted water sources. Soil has been polluted by industrial waste, agrochemicals, sewage, and polluted water. Heavy metals pollute nearly one-sixth of the cultivated land mass in China; areas on the outskirts of cities see cadmium, mercury, arsenic, copper, lead, nickel, and chromium soil contamination. Urban industrial waste plays a major role, but pesticide application also contributes. Major enterprises or industries are not the sole introducers of pollution: even if given safety instructions with agricultural-use chemicals, farmers may be unable to read them.

These environmental hazards strongly manifest in urban-rural

---

119 See, e.g., Wenfeng Mao & Shujuan Zhang, Impacts of the Economic Transition on Environmental Regulation in China, 5 J. ENVTL. ASSESSMENT POL’Y & MGMT. 183, 183–84 (2003) (discussing how economic growth, amongst other factors, have created new “environmental stresses”).

120 Ying Huang & Lei Liu, Fighting Corruption: A Long-Standing Challenge for Environmental Regulation in China, 12 ENVTL. DEV. 47, 47 (2014).

121 Wang et al., supra note 118, at 1.

122 Margherita Poto, Environmental Regulation in China through the Lens of the European Model, 18 ASIA PAC. J. ENVTL. L. 69, 73 (2016).

123 Mao & Zhang, supra note 119, at 183–84.

124 Wang et al., supra note 118, at 4.


126 See id.

127 See id. (“Reportedly, soil pollution associated with urban and industrial sources accounts for one-third to half of the total polluted arable area in the country, with a concentrated distribution in the peri-urban zones.”).

128 See id. (“Besides heavy metals, 15 kinds of polychlorinated biphenyls (PCBS) [which can be used as pesticide extenders] and organochlorine pesticides were detected in all of the tested paddy soils.”).

129 Prändl-Zika, supra note 106, at 238.
inequalities. The exact incidence of this is undetermined; it is often assumed that health risks related to pollution fall disproportionately on the poor, but in China this is yet to be formally studied across population groups. A lack of options and resources does put the poor at risk of health problems generally, though problems specifically related to pollution fall into a much more complex set of circumstances. Still, there is evidence that rural residents are the ones suffering more. Geography matters. Additionally, the difference in stringency of environmental policies between rural areas and urban areas plays a role. Furthermore, low population density and insufficient monitoring in rural areas are also contributing factors.

This increased rural exposure to pollution causes negative health effects by finding its way through the food supply. “As a country undergoing rapid changes in economic, social, cultural and political life, and with people in rural and urban areas in different parts of the country living and working in vastly different physical and social conditions, China faces a particularly complicated set of environment and health challenges.” Despite strong efforts and continual improvement, “environmental deteriorations have caused significant damages to China’s economy and its socio-welfare,” including the health of rural populations.

Returning to the goals of a food safety regime, the three pillars are each threatened by this environmental situation in China. Soil pollution has been a concern for food quality safety. Water is undrinkable in

---

130 Liu et al., supra note 107, at 1 (“In general, rural residents . . . have been found to suffer more environmental pollution from industry after controlling for other socioeconomic factors.”).
131 See Jennifer Holdaway, Environment and Health Research in China: The State of the Field, 214 CHINA Q. 255, 269 (2013); see also Liu et al., supra note 107, at 2–3 (stating that environmental inequality findings are “mixed”).
132 Holdaway, supra note 131, at 270.
133 See, e.g., Chunbo Ma, Who Bears the Environmental Burden in China—An Analysis of the Distribution of Industrial Pollution Sources?, 69 ECOLOGICAL ECON. 1869, 1874 (2010) (stating that “townships with a higher proportion of rural migrant residents are more likely to have a higher exposure to environmental pollution” and “rural townships and residents suffer more from pollution”).
134 Liu et al., supra note 107, at 3.
135 Ma, supra note 133, at 1874.
136 Liu et al., supra note 107, at 2.
137 Holdaway, supra note 131, at 256–57.
138 Wang et al., supra note 118, at 3.
139 Chen, supra note 125, at 9. Chen states, “soil not only functions as an indispensable physical base to provide humankind a majority of food, livestock feed, fibre and biotic fuel, but serves as both a source and a sink for green house gases and an integrated part of
many areas, but continues to be used for irrigation. Air pollution and overuse of agricultural chemicals are negatively affecting arable land. All of these pollutants make their way into the food supply, increasing health risks from eating contaminated food.

Food security is at risk due to lacking environmental regulation. Arable land is being reduced, threatening supply. Pollution from the deposit of acid rain is accelerating soil degradation, depleting nutrients, and heightening the bioavailability of heavy metals, jeopardizing yield and raising contamination risk simultaneously. Changes in soil utilization (from subsistence grains toward the raising of animals for meat and dairy) have tended toward decreasing agricultural efficiency by increasing the production of particularly land-consuming foods and risking future output by focusing production on particularly nutrient-hungry crops. These factors reduce access to sustainable, natural resources that are necessary for long-term food production, and thus threaten the stability of the food supply.

Public health is directly threatened by natural resource pollution, including that in the water, soil, and air. Indirectly, chemicals and heavy metals move through the food supply to cause sickness, injury, and worse. Biologically contaminated water can carry disease in addition to other environmental pollutants. Economically, this burdens biogeochemical cycles.

---

140 Id. at 2.
141 Id. at 9.
142 Prändl-Zika, supra note 106, at 238; see also Yang & Li, supra note 106, at 73; see also Chen, supra note 125, at 2–4.
143 Chen, supra note 125, at 10.
144 Prändl-Zika, supra note 106, at 238 (“The consequences [of shifting from grain production toward meat and dairy production] are that more land and more energy are required for those products delivering the same nutrition energy as grain products.”); Chen, supra note 125, at 11–12.
145 Chen, supra note 125, at 12 (“For instance, fruits and vegetables remove considerably more nutrients from the soils than cereal grains.”).
146 Wang et al., supra note 118, at 22.
147 See, e.g., Mary E. Zabik, Polychlorinated Biphenyls and Polybrominated Biphenyls in Foods, in THE SAFETY OF FOODS 444 (Horace D. Graham ed., 2d ed. 1980) (discussing polychlorinated biphenyls and its ability to accumulate in food chains); see also Islam Ejaz ul et al., Assessing Potential Dietary Toxicity of Heavy Metals in Selected Vegetables and Food Crops, 8 J. ZHEJIANG U. SCI. B 1, 2 (2007) (discussing heavy metals as cumulative poisons that “have damaging effects on human beings and other animals” and are “exceptionally toxic”).
public health resources,\textsuperscript{149} it affects the desirability of Chinese-produced agricultural products abroad,\textsuperscript{150} and it also reduces agricultural efficiency.\textsuperscript{151}

China’s efforts in food regulation are well-publicized in the wakes of its major food scandals.\textsuperscript{152} Chinese food regulation has in turn been treated with international respect, though support is often guarded in light of safety issues that have had international impacts.\textsuperscript{153} Unfortunately, China’s record with environmental regulations or protections is both comparatively and objectively woeful.\textsuperscript{154} Under the rapidly unfolding umbrella of industrialization, China has clearly placed environmental protections on a back burner when it comes to policymaking, and this has had an indirect effect on food safety, especially among the rural poor.

IV. STATEMENT OF THE PROBLEM AND SOLUTION

The triaging of environmental regulations has negative consequences in many arenas, and food safety is one that is often overlooked.\textsuperscript{155} While researchers and scholars are no strangers to the deleterious effects of contaminated environments on the safety of food cultivated in those very environments,\textsuperscript{156} China’s particularly rapid and thorough recent history of urbanization and industrialization have

\begin{itemize}
\item \textsuperscript{149} See Eleanor Albert & Beina Xu, \textit{China’s Environmental Crisis}, COUNCIL ON FOREIGN RELATIONS (Jan. 18, 2016), http://www.cfr.org/china/chinas-environmental-crisis/p12608 (“Environmental depredations pose a serious threat to China’s economic growth, costing the country roughly 3 to 10 percent of its gross national income. . . .”); see, e.g., \textit{WORLD BANK & STATE ENVIRONMENTAL PROTECTION ADMINISTRATION, P. R. CHINA, COST OF POLLUTION IN CHINA: ECONOMIC ESTIMATES OF PHYSICAL DAMAGES 19} (2007) (suggesting that urban air pollution results in public health and economic harm).
\item \textsuperscript{150} Peng Liu, \textit{supra} note 15, at 245 (“Maintaining the reputation of ‘Made in China’ has . . . become a serious policy issue for Chinese leaders.”).
\item \textsuperscript{151} Chen, \textit{supra} note 125, at 9 (noting a study that found heavy metal pollution causing “an annual grain yield loss of 10 million tons”).
\item \textsuperscript{152} See, e.g., Congressional Research Service, \textit{China’s Efforts to Address Ongoing Food Safety Concerns}, IN FOCUS (Sept. 9, 2016), http://nationalaglawcenter.org/wp-content/uploads/assets/crs/IF10465.pdf (discussing China’s response to recent food scandals).
\item \textsuperscript{153} See, e.g., \textit{id.} (discussing Congress’ concern over food safety in China despite recent efforts).
\item \textsuperscript{154} See Wang et al., \textit{supra} note 118, at 1 (stating that the environmental efforts in the past were insufficient and the evidence is clearly seen in the “unbearable pollution levels” and other economic and environmental concerns).
\item \textsuperscript{155} See Junfeng Zhang et al., \textit{Environmental Health in China: Progress Towards Clean Air and Safe Water}, 375 LANCET 1110, 1111 (2010).
\item \textsuperscript{156} See \textit{supra} Part III(C).
\end{itemize}
markedly marginalized the groups that would be most injured by contaminated resources and their effects on local food supply. Put more directly, China’s push to industrialize has forced some of the most vulnerable populations out of the regulatory scheme designed to protect citizens. The sources of this conflict are threefold. First, China has industrialized too quickly for regulatory schemes to keep up. Second, China’s environmental quality problems are persistent and directly affect the natural resources needed for food cultivation. Third, regulation has moved its focus with the bulk of the population as it has shifted toward urbanization, casting an ever-darker shadow on the likelihood of addressing environmental concerns in rural areas.

Chinese food regulation is focused heavily on the urban subset of the population to the detriment of the remainder. Those outside of the urban subset suffer not just because they fall outside of the scope of the regulations themselves, but because they are more likely to consume food from sources that are also under-regulated. Chinese environmental regulation is proving ineffective, and many people find themselves farming on contaminated land or with contaminated water. Despite the growing strength of Chinese food safety laws and their implementation, this population still faces the risk of eating unsafe food.

The solution to this problem is to introduce or rework environmental regulations to deal directly with substances or resources in the food supply chain. Hazardous heavy metals, chemicals, and the like, which may have negative effects on the soil or water, along with the soil and water themselves, are fertile ground for regulation that could save the lives of those who depend on their proper management.

China’s 2009 Food Safety Law was written by the State Council, and it required that the Council establish a Food Safety Commission (“FSC”) to oversee the administration of the law. The FSC remains in the 2015 Food Safety Law, which brings other “health related ministries, commissions, and departments” under the umbrella of the FSC in enforcing the law and tasks “local people’s governments above the county level” with administering food safety regulations directly. This

---

157 Banister, supra note 148, at 997 (“Much of China’s food is grown and consumed in villages without the involvement of a market, so it is only localities and households that can prevent its contamination.”).

158 See supra Part III (C).

159 Chenglin Liu, supra note 31, at 283–85.


161 Id. at 4, art. 6.
local power of administration includes the authority to determine the duties of both the Food and Drug Administration and the Health Administration in China, along with other departments at the same governmental level.\footnote{Id.} The Ministry of Environmental Protection is, like the National Health and Family Planning Commission,\footnote{See Shan Juan, \textit{New Health Commission Set to be Established}, CHINA DAILY (last updated Mar. 11, 2013), http://usa.chinadaily.com.cn/china/2013-03/11/content_16296626.htm.} a cabinet-level department.\footnote{See, e.g., \textit{People’s Republic of China Ministry of Environmental Protection}, AECEN, http://www.aecen.org/stories/people%E2%80%99s-republic-china-ministry-environmental-protection (last visited Sept. 21, 2017) (stating that the Ministry of Environmental Protection was “[e]stablished by the elevation of the State Environmental Protection Agency to the cabinet level in 2008”); \textit{The Organizational Structure of the State Council}, CENT. PEOPLE’S GOV’T., http://www.gov.cn/english/links.htm#1(last visited Sept. 8, 2017).}

Regulating environmental pollution will protect natural resources upon which agricultural self-producers rely. Such regulation would serve to protect both pipelines for food in China—the commercial and the traditional. Only by protecting both pipelines can the goals of food security, food safety, and public health and economic efficiency be adequately met for all citizens, urban and rural.

\section*{V. COUNTERARGUMENTS}

One major counterargument that might be raised is that China is a socialized nation with an ongoing plan to both provide for the needs of its citizens and control the means by which those citizens are supplied. Specifically, China’s government has the power and intent to commercialize farming nationally and provide for every last one of its citizens through its state-controlled production and distribution channels. By doing so, the need for subsistence agriculture is eliminated along with the risks that would prompt relevant (environmental) safety regulation.

While this is a strong argument, it misses the mark in two ways. First, regardless of food supply issues, China is in desperate need of improved environmental regulation and enforcement. To tie this effort to food safety regulations (existing or future) may require only a marginal increase in cost. Just as enforcement of commercial food safety regulations has proven difficult in such a broad and diverse market, there will doubtless be some sustenance farmers that go untouched by even perfect food safety regulations. For near-zero cost, a safety net for these
citizens is difficult to argue against. Second, ubiquitous commercial farming on a large scale is a massive undertaking. It will not happen instantly, and so, even as merely a stop-gap measure, some food regulations should exist that take into account hand-to-mouth cultivators.

Another potential counterargument is that the changes necessary to accomplish integration between environmental regulation and food safety regulation are simply too great to accomplish. The two spheres are regulated by different ministries unlikely to yield political power to one another; in reality, the FSC lacks the authority implied by the Food Safety Law, and China has little incentive to spend limited resources on marginalized minority populations.

This counterargument also fails. While different ministries may indeed refuse to cooperate, and even if the FSC lacks authority, China is exploring other avenues to regulation that could bridge the gaps. The first is the concept of an integrated approach through social governance. Under this concept and using this approach, China could tackle holes in food safety regulation by encouraging better practices socially; that is, by addressing some social issues, such as disparity in information access, helping outsider groups understand and accept best practices, and even simply encouraging some political involvement on a grassroots level, China can more quickly mitigate some of the issues that formal lawmaking might eventually address. As for incentives, there is a growing trend of urban agriculture migrating from the West, particularly the United States.

---


166 Because local governments are tasked with oversight of food safety, and rural subsistence agriculture goes largely unregulated, the FSC is unlikely to be able to control all aspects of food safety throughout China. See supra Parts II(C) and IV.


168 See Pinghui, supra note 8; see also Yan Shi et al., Safe Food, Green Food, Good Food: Chinese Community Supported Agriculture and the Rising Middle Class, 9 INT’L J. AGRIC. SUSTAINABILITY 551, 553 (2011).
classes are also growing their own food, and increasingly so.\textsuperscript{169} Urban areas in China are frequently heavily polluted,\textsuperscript{170} and so these higher, majority classes are also at risk for contamination. As such, China’s incentive to regulate is larger than it may initially appear.

VI. CONCLUSION

China’s food safety regulation is imperfect, suffering from occasional high-profile scandals.\textsuperscript{171} Food safety regulation, however, makes a positive impact on the food market, both domestically and internationally, and has been remarkably successful given its quick implementation and the sheer scale of the Chinese food market.\textsuperscript{172} China’s environmental regulations are far less effective, and pollution is rampant in soil, water, and air.\textsuperscript{173} China’s rapid industrialization has lifted millions from poverty and put them under the protection of commercial food regulation, but it has simultaneously accelerated pollution of land, water, and other natural resources.\textsuperscript{174} Ensuing urbanization has left many people unprotected as they remain in under-regulated rural areas.\textsuperscript{175} The sheer scale of the Chinese government’s focus on its urban population has left those growing their own food particularly vulnerable and marginalized, facing disproportionate risks that come from eating food grown with contaminated resources.\textsuperscript{176}

The solution to this problem is to introduce, rework, or find food safety applications for environmental regulations that can affect substances or resources in the food supply chain. Attention must be directed to hazardous heavy metals, chemicals, and the like that may have negative effects on or through soil, water, and air. If dealt with effectively, a harmonization between food safety regulation and environmental regulation could save many of those whose lives directly depend on natural resources for food.

\textsuperscript{169} See Pinghui, supra note 8 (“[M]any mainlanders—frustrated about numerous food scandals that have plagued even staple elements of Chinese dinner tables . . . —are taking matters into their own hands to keep their diets safe.”).

\textsuperscript{170} Hong Yang et al., Soil Pollution: Urban Brownfields, 344 SCI. 691, 691 (2014).

\textsuperscript{171} See supra Part II(C).


\textsuperscript{173} See generally, Wang et al., supra note 118.


\textsuperscript{175} See supra notes 109–115 and accompanying text.

\textsuperscript{176} Liu et al., supra note 107, at 1–2.
JUSTICE IN FULL IS TIME WELL SPENT: WHY THE SUPREME COURT SHOULD BAN SUA SPONTE DISMISSALS

Michael J. Donaldson*

Sometimes I ask students to say whom they consider to be the most important person in a court room. Many pick the judge; others give a variety of answers. One even opted for the usher, without being able to explain why. My answer, given unhesitatingly, is that it is the litigant who is going to lose. Naturally he will usually not know this until the case is at an end. But when the end comes, will he go away feeling that he has had a fair run and a full hearing? . . . One of the important duties of the courts is to send away defeated litigants who feel no justifiable sense of injustice in the judicial process. . . . Justice in full takes time: but often it is time well spent.1

—Sir Robert Megarry

I. INTRODUCTION ................................................................. 26
II. COLLINS V. MERRILL: AN EXAMPLE OF SUA SPONTE DISMISSAL IN ACTION ................................. 28
III. THE GREAT AMERICAN TRADITION OF APPELLATE SUA SPONTE DECISION-MAKING ................................. 29
IV. THE CIRCUIT SPLIT ON DISTRICT COURT SUA SPONTE DISMISSALS .......................................................... 32
V. THE SUPREME COURT’S FAILURE TO REIN IN SUA SPONTE

*Partner, Burnet, Duckworth & Palmer, LLP, Calgary, Canada. LL.M. (Columbia), LL.B. (U.B.C.), B.A. (Queen’s). My thanks to Elizabeth M. Sacksteder of Paul, Weiss, Rifkind, Wharton & Garrison, LLP, for an interesting discussion that led to me writing this article, to Roberta A Kaplan of Kaplan & Company, LLP, for leading the very interesting seminar in which that discussion took place, and to Arundhati Katju for discussing and debating these and other issues. Thanks also to my BD&P colleague, Julia Lisztwan, and the editorial team at the Quinnipiac Law Review for their help with editing. As always, I take sole credit for any and all errors.

I. INTRODUCTION

“The fundamental requisite of due process of law is the opportunity to be heard.’ This right to be heard has little reality or worth unless one is informed that the matter is pending and can choose for himself whether to appear or default, acquiesce or contest.”2 Does this mean that a judge has to give every plaintiff notice and an opportunity to be heard before dismissing a case? Surprisingly, according to some U.S. courts, the answer is “no.”3

How can this be? Appearances matter tremendously in court. The legal system shouldn’t just be fair; it should also appear to be fair. As Lord Chief Justice Hewart put it in the famous Sussex Justices case, “justice should not only be done, but should manifestly and undoubtedly be seen to be done.”4 This article is about a troubling practice in which some federal district courts sacrifice due process in the name of efficiency and, with the explicit blessing of their circuit courts of appeals, dismiss supposedly hopeless cases without giving the plaintiff notice of the impending dismissal or an opportunity to be heard.

There is a vast difference between a court raising an issue sua sponte and the court deciding an issue sua sponte—by which I mean deciding an issue that only the court has raised, without hearing from the

---

3 See, e.g., Baker v. Dir. U.S. Parole Comm’n, 916 F.2d 725, 725–727 (D.C. Cir. 1990) (upholding district court’s dismissal of complaint without providing plaintiff with notice or the opportunity to respond); see also Curley v. Perry, 246 F.3d 1278, 1281–82, 1284 (10th Cir. 2001) (citing Hall v. Bellmon, 935 F.2d 1106, 1110 (10th Cir. 1991)) (permitting dismissal without notifying the plaintiff or permitting an opportunity to be heard is proper when patently obvious that the plaintiff cannot prevail and amending the complaint is futile).
parties on that issue. This article deals only with the latter. Is it ever appropriate for a district court, after identifying what it sees as an important issue the parties have missed, to decide that issue without first giving the parties notice and an opportunity to be heard? This is an issue on which the circuit courts are split and on which the Supreme Court has so far declined to give a clear answer.

Much has been written about the perils of appellate courts deciding issues sua sponte. Far less has been written about trial courts doing so, even though “criticisms” about the practice “apply with equal force to the sua sponte decisions of both appellate and trial courts.” But isn’t the problem even more serious when it happens in a trial court? In comparison with appellate courts, trial courts have much more frequent

---

5 The term “sua sponte” means “[w]ithout prompting or suggestion; on its own motion.” Sua sponte, BLACK’S LAW DICTIONARY (10th ed. 2014). In using the term “sua sponte dismissal,” I will follow the lead of Adam A. Milani and Michael R. Smith:

Bryan Garner, an expert on word usage in legal writing, states that the phrase “sua sponte” refers to how an issue is raised and suggests that it is not proper to use the phrase “sua sponte” to refer to how an issue was decided. Despite Garner’s admonishment, we use the phrase “sua sponte decision” in this Article to refer to situations in which an issue was not only raised by the court but was also decided by the court alone, without the input of the parties.


6 See infra Part IV.

7 See infra Part V. The Supreme Court expressly left open whether an appellate court must give the parties an opportunity to be heard in 1940, and left the district court question open in 1989. But, as we will see, the Court itself has a long and illustrious history of deciding important cases sua sponte, so if the Court did require such notice, it would be telling the lower courts to “do as I say, not as I do.”


10 Milani & Smith, supra note 5, at 251 n.17.
contact with ordinary citizens. Most people who are unfortunate enough to see the inside of a federal district court will never encounter an appellate court; according to one study, about 80% of civil cases that end with a dispositive order before trial are never appealed. The district courts are the front line of the interface between the federal courts and the public, and are often the only face of the federal judicial system that litigants will ever see. If the district court’s procedures do not seem fair, the theoretical availability of a fair hearing in some other court is likely to do little to inspire confidence in the system.

II. COLLINS V. MERRILL: AN EXAMPLE OF SUA SPONTE DISMISSAL IN ACTION

On December 5, 2016, Paula Collins, a registered New York voter, filed a pro se claim against dozens of state and federal officials seeking “declaratory and injunctive relief restraining Defendants from certifying the vote of the Electoral College ‘in any way that is inconsistent with the results of the nationwide popular vote as it was decided’ in the presidential election that concluded on November 8, 2016.” Two days later, Ms. Collins was notified that her claim had been dismissed for lack of standing. She does not appear to have been notified that the court was considering a dismissal.

While the dismissal was without prejudice, refiling the same claim would presumably have resulted in another dismissal on identical grounds and perhaps would have invited sanctions. If Ms. Collins did have some viable legal theory in support of her claim, she would, therefore, have had no way to present it to the district court. If she wanted to have a hearing (even a paper hearing) at which she could argue why she did have standing and respond to the issues the district

---

13 Id. at *2.
14 This seems obvious from the fact that the dismissal order was entered two days after filing; it also appears to be the case from a review of the docket entries. Id. at 1.
15 Id. at *2. This is due to Carter v. HealthPort Techs., which mandates that dismissals for lack of Article III standing are dismissed without prejudice on the theory that “without jurisdiction, the district court lacks the power to adjudicate the merits of the case.” Carter v. HealthPort Techs., LLC, 822 F.3d 47, 54–55 (2d Cir. 2016); see also Collins, 2016 WL 7176651, at *2.
court raised in its brief opinion, Ms. Collins’ only remedy would have been to appeal to the Second Circuit.16

There is little doubt that the district court’s decision was correct and that Ms. Collins did not meet the requirements for Article III standing—at least as her claim had been framed. Dismissal was the correct result, but did Ms. Collins receive due process? What if the district court had overlooked some authority supporting her position? What if she had a brilliant argument that might have resulted in the court viewing either the facts or the law differently? Is it good enough to tell a disappointed litigant that if they want a fair hearing they can get one in an appellate court?

While the district court didn’t address these questions in Collins v. Merrill, its actions were consistent with a long and storied tradition of sua sponte decision-making in American law—although more so in appellate courts than in trial courts.17 This is one of the great ironies uncovered by the issue of sua sponte dismissals: although some appellate courts are critical of lower court judges for deciding cases sua sponte, they engage in the practice themselves, frequently without acknowledging they are doing so.18

III. THE GREAT AMERICAN TRADITION OF APPELLATE SUA SPONTE DECISION-MAKING

In a section of their article aptly entitled, “Sua Sponte Decisions by Appellate Courts Include Some of the Most Far-Reaching Cases in American Law,” Adam A. Milani and Michael R. Smith are quick to

---

16 Another (theoretical) remedy would be a motion under Rule 60(b), discussed infra Part VII. But Ms. Collins might be forgiven for wondering whether such a hypothetical hearing would be useful given that the court had already decided against her without hearing from her. It should be noted, however, that even if the court had subject matter jurisdiction to entertain her suit, Ms. Collins’ claim would still fail. Collins, 2016 WL 7176651, at *2 n.1 (“Plaintiff would clearly fail to state a claim on which relief can be granted. ‘[T]he electoral college cannot be questioned constitutionally because it is established by the Constitution,’ and the Court is ‘not empowered to strike the document’s text on the basis that it is offensive to itself or is in some way internally inconsistent.’”) (alteration in original) (quoting New v. Aschcroft, 293 F.Supp.2d 256, 259 (E.D.N.Y. 2003)).

17 See infra Part III.

18 See Vestal, supra note 8, at 497; Joan Steinman, Appellate Courts as First Responders: The Constitutionality and Propriety of Appellate Courts’ Resolving Issues in the First Instance, 87 NOTRE DAME L. REV. 1521, 1549 (2012) (providing evidence that sources like The Federalists Papers never contemplated appellate courts having the ability to decide issues sua sponte, yet the Supreme Court repeatedly raises new issues on appeal despite some justices feeling uneasy about doing so).
note the significant impact of sua sponte decision-making in American courts. They say, for example:

[T]wo of the most significant cases in American civil procedure and criminal procedure—Erie R.R. Co. v. Tompkins and Mapp v. Ohio, respectively—were decided on issues that were neither briefed nor argued before the Court. Accordingly, the losing parties never had an opportunity to be heard on the issues that ultimately determined the cases’ outcomes. This likely will—and should—shock not only many of the civil and criminal litigators who live with the consequences of those decisions on a daily basis, but also lawyers whose only encounter with these cases was reading them as part of required classes in law school.

_Erie_, of course, established when and how federal courts are required to apply state law. In doing so, the Court overruled its long-standing decision in _Swift v. Tyson_. The majority opinion in _Erie_ begins by saying: “The question for decision is whether the oft-challenged doctrine of _Swift v. Tyson_ shall now be disapproved.” But neither of the parties in _Erie_ had argued that _Swift_ should be overturned—the railroad went so far as to say in its brief, “We do not question the finality of the holding of this court in _Swift v. Tyson_.” Nonetheless, the court overruled _Swift_ sua sponte without providing the parties with notice or the opportunity of a rehearing or further briefing. This caused Justice Butler to suggest in his dissent that “the Court should move cautiously, seek assistance of counsel, act only after ample deliberation, [and] show that the question is before the court,” noting that he and the other dissenters had called for reargument, but had been overruled by the majority. Finally, Justice Butler protested, “[i]t may not justly be assumed that the labor and argument of counsel for the parties would not disclose the right conclusion and aid the Court in the statement of the reasons to support it.”

---

19 Milani & Smith, supra note 5, at 253.
20 Id. at 253–54.
22 Id. at 79–80.
23 Id. at 69 (footnote omitted).
25 See Milani & Smith, supra note 5, at 254.
26 Erie, 304 U.S. at 87 (Butler, J., dissenting).
27 Id. at 88.
28 Id.
As in Erie, in Mapp v. Ohio, the Court overruled its own precedent, Wolf v. Colorado, even though the appellant never argued that Wolf should be overruled in its brief or in oral argument. Instead, “when pressed by questioning from the bench whether he was not in fact urging us to overrule Wolf, counsel [for appellant] expressly disavowed any such purpose.” Milani and Smith note that, according to various sources, the first serious discussion about overruling Wolf occurred not in a brief or in the courtroom, but rather after argument and after the justices’ conference, in an elevator discussion between Justices Clark, Black, and Brennan. Justice Harlan’s dissent in Mapp points out the unfairness and “unwisdom of overruling Wolf without full-dress argument.”

The occasion which the Court has taken here is in the context of a case where the question was briefed not at all and argued only extremely tangentially. . . . I would think that our obligation to the States, on whom we impose this new rule, as well as the obligation of orderly adherence to our own processes would demand that we seek that aid which adequate briefing and argument lends to the determination of an important issue.

. . . [A]t the very least, the present case should have been set down for reargument, in view of the inadequate briefing and argument we have received on the Wolf point. To all intents and purposes the Court’s present action amounts to a summary reversal of Wolf, without argument.

I am bound to say that what has been done is not likely to promote respect either for the Court’s adjudicatory process or for the stability of its decisions.

What is unusual about Erie and Mapp is that the Court actually acknowledged it had decided the major issues without argument. The reality is that with most sua sponte decisions, “[u]nless there is a dissenting opinion noting the fact, only the attorneys for the litigants will be aware that the court has decided the case on issues not argued to the court.” (One study of 112 majority opinions from state appellate courts over the course of a year found that sua sponte decision-making occurred in about 15% of cases, but was only identified as such

31 Mapp, 367 U.S. at 676 n.6 (Harlan, J., dissenting).
32 Milani & Smith, supra note 5, at 258.
33 Mapp, 367 U.S. at 676 (Harlan, J., dissenting).
34 Id. at 676–77.
35 Vestal, supra note 8, at 497 (footnote omitted).
explicitly in 2.5% of the cases.) With examples like this, it is perhaps little wonder that district courts feel emboldened to dismiss cases without providing the litigants with an opportunity to make submissions where the court believes that the issues are obvious and the result foreordained. Yet it is not just *Erie*, *Mapp*, and similar rulings by appellate courts that implicitly send a message that sometimes due process can be dispensed with—several federal courts of appeals explicitly endorse the practice.

IV. THE CIRCUIT SPLIT ON DISTRICT COURT SUA SPONTE DISMISSALS

In all circuits, a judge is permitted to dismiss a case as a result of raising jurisdictional or other issues sua sponte. The circuits are split, however, on whether a judge must give a plaintiff notice and an opportunity to be heard before doing so. And they are further split on the consequences of a judge’s failure to give notice.

The First, Second, Sixth, and Eleventh Circuits regard a
district court’s failure to give notice as a basic denial of due process that voids the dismissal order and requires a remand. The Second Circuit in *Perez v. Ortiz* explained its position this way:

> Although *sua sponte* dismissals may be appropriate in some circumstances, particularly in cases involving frivolous *in forma pauperis* complaints, or frivolous habeas petitions, the general rule is that a “district court has no authority to dismiss a complaint for failure to state a claim upon which relief can be granted without giving the plaintiff an opportunity to be heard.” . . . “Failure to afford an opportunity to address the court’s *sua sponte* motion to dismiss is by itself, grounds for reversal” . . .

> As we have noted before, “adequate notice helps the court secure a just determination” by giving parties moved against the opportunity to present their best arguments in opposition. *Sua sponte* dismissals, especially those entered without notice, also deviate from the traditions of the adversarial system by making the judge “a proponent rather than an independent entity”. Finally, such dismissals may tend to produce the very effect they seek to avoid—a waste of judicial resources—by leading to appeals and remands. 43

Recall that the district court in *Collins v. Merrill* did not provide such notice, 44 so it may be that, in the busy district courts, the practice is not always followed, even in those circuits that explicitly require it. The use of language such as “in some circumstances” and “generally” in the above quote from *Perez* may also be read as exempting hopeless cases, but this seems inconsistent with much of the Second Circuit’s jurisprudence on the point.

The Fifth, 45 Seventh, 46 and Eighth Circuits 47 regard giving notice as

---

42 See, e.g., Jefferson Fourteenth Assocs. v. Wometco de P.R., Inc., 695 F.2d 524, 527 (11th Cir. 1983) (requiring that plaintiff have the opportunity to file a written response and present arguments at a hearing before *sua sponte* dismissal); cf. Surtain v. Hamlin Terrace Found., 789 F.3d 1239, 1248 (11th Cir. 2015) (requiring notice and opportunity to respond prior to a dismissal *sua sponte*, but also stating that this requirement is excepted “when amending the complaint would be futile or when the complaint is patently frivolous”).

43 *Perez*, 849 F.2d at 797 (citations omitted).

44 See *supra* note 14 and accompanying text.

45 See, e.g., Lozano v. Ocwen Fed. Bank, FSB, 489 F.3d 636, 643 (5th Cir. 2007) (noting that case law suggests requiring notice and an opportunity to respond).

46 See, e.g., Joyce v. Joyce, 975 F.2d 379, 386 (7th Cir. 1992) (“[W]hile we reiterate the caution necessary in *sua sponte* dismissal without notice of hearing, we find no error in the district court’s action in this case [due to the defect being incurable].”).

47 See, e.g., Smith v. Boyd, 945 F.2d 1041, 1043 (8th Cir. 1991) (stating that failure to provide pre-dismissal notice does not require a reversal, even though courts “should provide pre-dismissal notice”).
a best practice that district courts “should” follow, warning that dismissals sua sponte without notice are “hazardous,” but these courts do not regard failure to give notice as reversible error. Instead, where the district court fails to give notice, the appellate courts purport to cure the denial of due process by reviewing the district court decision de novo.

The Eighth Circuit described the rule as follows in Smith v. Boyd:

Smith next contends the district court committed reversible error by dismissing his complaint without giving him prior notice and an opportunity to respond. We disagree and hold that the failure to give such notice is not per se reversible error when it is patently obvious the plaintiff could not prevail based on the facts alleged in the complaint. Though district courts should provide pre-dismissal notice, we decline to hold that the failure to give such notice mandates reversal. Rather, we will review the allegations of the complaint de novo . . . .

These courts do not address the unfortunate reality that a plaintiff whose case is dismissed without notice must pursue an appeal in order to secure the fair hearing she was entitled to have in the court of first instance. Moreover, they fail to answer several elephant-in-the-room questions: How can any court possibly conclude that “it is patently obvious the plaintiff could not prevail based on the facts alleged in the complaint” without first hearing argument on the issue? Something might appear patently obvious to a judge based on whatever the judge is aware of when she reads the complaint, but what about case law, statutes, or arguments she isn’t aware of? As we will see, the risk of overlooking such important information in cases dismissed sua sponte is real and not infrequent. And when, exactly, is a court permitted to invoke this “patently obvious” standard? To whom must the obviousness be patent—the district court judge, the appellate judges who might review it, the plaintiff and/or her attorney, all or none of the above? These are all questions about which the district court judges—and litigants—in these circuits are left to wonder.

The Third, Ninth, Tenth, and D.C. Circuits go further,

48 Id. In Lozano, however, the Fifth Circuit stated, “We do not always require notice prior to sua sponte dismissal . . . as long as the plaintiff has alleged his ‘best case.’” Lozano, 489 F.3d at 643 (citations omitted).
49 Shockley v. Jones, 823 F.2d 1068, 1072 (7th Cir. 1987) (quoting Tamari v. Bache & Co., 565 F.2d 1194, 1198 (7th Cir. 1977)).
50 See Smith, 945 F.2d at 1043.
51 Id. (footnote omitted) (citations omitted).
52 See, e.g., Goodwin v. Castille, 465 F. App’x 157, 163 (3d Cir. 2012) (stating that sua sponte dismissal was proper because the appellant could not provide additional facts that
suggesting that there is nothing wrong with a district court dismissing a complaint without notice “where the claimant cannot possibly win relief.”56 These courts argue that insisting on district courts providing a plaintiff with notice and an opportunity to be heard in such a case “can only lead to a waste of judicial resources.”57 The courts’ opinions in these cases, however, rely on earlier cases that strongly suggest just the opposite.58 Somehow, reliance on these earlier cases, which insist on procedural fairness, has resulted in a rule that discards procedural fairness. In addition, this rule appears to have migrated from a narrow class of prisoners’ pro se habeas petitions59 to almost anything a judge can conceive of as a basis for dismissal under Rule 12(b)(6) of the Federal Rules of Civil Procedure. The D.C. Circuit stated the rule and its rationale in a manner that expanded dismissal sua sponte under Rule 12(b)(6) in Baker, relying on the Ninth Circuit’s opinion in Omar:

would dispute a finding of legislative immunity).

53 See, e.g., Omar v. Sea-Land Serv., Inc., 813 F.2d 986, 991 (9th Cir. 1987) (stating that a trial court may dismiss a claim sua sponte without notice where there is no possible way for a claimant to win relief).

54 See, e.g., McKinney v. Okla. Dep’t of Human Servs., 925 F.2d 363, 365 (10th Cir. 1991) (stating that sua sponte dismissal under Rule 12(b)(6) of the Federal Rules of Civil Procedure is proper where it is patently obvious that plaintiff cannot prevail on the merits).


56 Baker, 916 F.2d at 726 (quoting Omar v. Sea-Land Serv., Inc., 813 F.2d 986, 991 (9th Cir. 1987)).

57 Id.

58 For example, the court in Goodwin relies on Oatess v. Sabolevitch, 914 F.2d 428, 430 (3d Cir. 1990), for the proposition that “sua sponte dismissal under Rule 12(b)(6) is proper after service of process.” Goodwin v. Castille, 465 F. App’x 157, 163 (3d Cir. 2012). Oatess, however, raises concerns about any dismissal on the merits before service; specifically stating “[r]ather than promoting efficient case management, premature dismissal often results in greater inefficiency.” Oatess, 914 F.2d at 431. Oatess also emphasizes upholding the “tradition of adversarial proceedings,” which are bypassed when “dismissal under Rule 12(b)(6) occurs before service of process.” Id. Additionally, the court in Baker stated that its holding did not conflict with the holding in Brandon v. District of Columbia Bd. Of Parole, 734 F.2d 56 (D.C. Cir. 1984) because Brandon did not suggest that courts could not dismiss cases under Rule 12(b)(6) in the absence of a defendant’s motion. Baker, 916 F.2d at 726–27. Upon examination of Brandon, however, the court appears to be hesitant about sua sponte dismissals for failure to state a claim, and even suggests that a complaint should not be dismissed sua sponte even if it does not “indisputably state a cause of action.” Brandon, 734 F.2d at 59.

59 See, e.g., Williams v. Kullman, 722 F.2d 1048, 1049, 1052 (2nd Cir. 1983) (affirming trial court decision to dismiss pro se prisoner’s habeas petition for failure to state claim entitling petitioner to relief).
The Ninth Circuit refused to find reversible error holding that a trial court may dismiss a claim *sua sponte* without notice “where the claimant cannot possibly win relief.” We adopt the position taken by our sister circuit: it is practical and fully consistent with plaintiffs’ rights and the efficient use of judicial resources.

We note that certain other circuits enforce a strict notice requirement with regard to *sua sponte* dismissals pursuant to Rule 12(b)(6) and mandate reversal for noncompliance with procedural steps dictated by the court. Such an approach, however, in cases where the plaintiff has not advanced a shred of a valid claim can only lead to a waste of judicial resources.

. . . [T]he “sharply honed adversarial exchange” involved in a Rule 12(b)(6) motion . . . does not suggest that Rule 12(b)(6) may never be used as a basis for dismissal except upon motion by the defendant.

Because it is patently obvious that Baker could not have prevailed on the facts alleged in his complaint, we find that *sua sponte* dismissal was appropriate.

Where did this new rule dispensing with notice come from? In *Omar* the Ninth Circuit did affirm a *sua sponte* dismissal without notice, relying on the Ninth Circuit’s earlier decision in *Wong v. Bell*. But *Wong* had criticized the district court for failing to give notice prior to dismissal for lack of standing, explicitly affirming the Ninth Circuit’s earlier holdings that “the court must give notice of its *sua sponte* intention to invoke Rule 12(b)(6) and afford plaintiffs ‘an opportunity to at least submit a written memorandum in opposition to such motion.’”

But then, in a cruelly ironic twist, the *Wong* court *did the very thing it had just said the district court should not do*—it decided the case on an issue the appellate court raised *sua sponte* and on which the parties do not appear to have made any argument. Rather than deciding the standing issue presented by the case—the issue on which the district court had dismissed and on which the parties had made their arguments on appeal—the court instead assessed *sua sponte* whether the claim could be dismissed under Rule 12(b)(6):

The only question which remains is whether we can affirm on the basis of

---

60 *Baker*, 916 F.2d at 726–27 (citations omitted).
61 *Omar v. Sea-Land Serv., Inc.,* 813 F.2d 986, 991 (9th Cir. 1987) (citing *Wong v. Bell*, 642 F.2d 359, 361–62 (9th Cir. 1981)).
62 *Wong v. Bell*, 642 F.2d 359, 362 (9th Cir. 1981) (citing *Crawford v. Bell*, 599 F.2d 890, 893 (9th Cir. 1979)) (quoting *Potter v. McCall*, 433 F.2d 1087, 1088 (9th Cir. 1970)).
63 “The parties to this appeal devote their efforts to arguing whether Wong’s brother and mother have standing to bring this action. Because resolution of that issue would require this court to formulate important new law, inapplicable to plaintiffs on the facts of this case, we choose an alternative mode of disposition.” *Id.* at 361.
Rule 12(b)(6) a dismissal based on other grounds. A trial court may act on its own initiative to note the inadequacy of a complaint and dismiss it for failure to state a claim, but the court must give notice of its sua sponte intention to invoke Rule 12(b)(6) and afford plaintiffs “an opportunity to at least submit a written memorandum in opposition to such motion.” Such was not done here. Nonetheless, the dismissal may properly be affirmed on the basis of Rule 12(b)(6). Plaintiffs cannot possibly win relief under the statute they have urged. In such an instance, “a federal appellate court is justified in resolving an issue not passed on below.”

And thus, it appears, the practice of dismissal sua sponte under Rule 12(b)(6) without notice or an opportunity to be heard was born.

V. THE SUPREME COURT’S FAILURE TO REIN IN SUA SPONTE DISMISSALS

The Supreme Court has not taken the opportunity to resolve the current circuit split on whether it is reversible error for a district court to fail to give a litigant notice and an opportunity to be heard before dismissing the case. On two occasions, the Supreme Court has passed on the opportunity to condemn the practice—once where an appellate court did so, and once with a district court dismissal.65 (Some writers suggest this is because the Supreme Court occasionally resorts to sua sponte decision-making, and may find it inconvenient to outlaw the practice altogether.)

In 1940, in LeTulle v. Scofield, the Supreme Court noted that it “took this case because the petition for certiorari alleged that the Circuit Court of Appeals had based its decision on a point not presented or argued by the litigants, which the petitioner had never had an

64 Id. at 361–62 (citations omitted).
65 See LeTulle v. Scofield, 308 U.S. 415, 419 (1940) (declining to rule on petitioner’s claim that circuit court’s sua sponte ruling was wrongful); Neitzke v. Williams, 490 U.S. 319, 321, 328 n.8 (1989) (“We have no occasion to pass judgment, however, on the permissible scope, if any, of sua sponte dismissals under Rule 12(b)(6).”).
66 See, e.g., Kevin T. McGuire & Barbara Palmer, Issue Fluidity on the U.S. Supreme Court, 89 AM. POL. SCI. REV. 691 (1995) (analyzing the “issue fluidity” with which the Supreme Court makes decisions on the merits, concluding that the Court is “often willing to provide authoritative answers to questions that have not been asked and to disregard issues that the parties have presented”); Shannon, supra note 8, at 27 (noting that the Supreme Court has confronted the issue of sua sponte decision-making “without much fanfare”); see also, Miller, supra note 8, at 1260–62 (suggesting that no clear principle of sua sponte decision-making permits courts to perform a legislative function, rather than a judicial function, when they deem appropriate, which also permits judges to prevent injustices when they see fit).
opportunity to meet by the production of evidence. 67 The Court, however, decided the case on different grounds and expressly declined to rule on the question on which it had granted certiorari. 68

The Supreme Court came closer to deciding the question in a 1989 case, Neitzke v. Williams, in which the Court had to decide whether the standards for dismissal of a frivolous in forma pauperis complaint under 42 U.S.C. § 1983 and dismissal under Rule 12(b)(6) were the same. 69 In holding they were not, Justice Marshall wrote for the court:

Under Rule 12(b)(6), a plaintiff with an arguable claim is ordinarily accorded notice of a pending motion to dismiss for failure to state a claim and an opportunity to amend the complaint before the motion is ruled upon. These procedures alert him to the legal theory underlying the defendant’s challenge, and enable him meaningfully to respond by opposing the motion to dismiss on legal grounds or by clarifying his factual allegations so as to conform with the requirements of a valid legal cause of action. This adversarial process also crystallizes the pertinent issues and facilitates appellate review of a trial court dismissal by creating a more complete record of the case. By contrast, the sua sponte dismissals permitted by, and frequently employed under, § 1915(d), necessary though they may sometimes be to shield defendants from vexatious lawsuits, involve no such procedural protections. 70

Justice Marshall’s statement that a plaintiff with “an arguable claim” is “ordinarily” provided with notice and an opportunity to respond has taken on great significance in the current circuit split, as some courts interpret this as carving out a set of cases in which plaintiffs do not have “arguable” claims and are accordingly entitled to “extraordinary” treatment. 71 These courts rely on the fact that the Supreme Court in Neitzke expressly left open the question of whether sua sponte dismissals under Rule 12(b)(6) are permitted 72 when it said,

---

67 LeTulle, 308 U.S. at 416.
68 Id. at 419.
70 Id. at 329–30 (footnote omitted) (citations omitted).
71 See, e.g., McKinney v. Okla. Dept. of Human Servs., 925 F.2d 363, 364–65 (10th Cir. 1991) (stating that although a plaintiff should be given notice of a pending motion to dismiss for failure to state a claim, this may be waived when it is “patently obvious that the plaintiff cannot prevail based on the facts alleged”).
72 As stated in McKinney:

The Supreme Court in Neitzke expressly declined to rule on the propriety of sua sponte dismissals under Fed.R.Civ.P. 12(b)(6). However, in Baker, the D.C. Circuit held “that a trial court may dismiss a claim sua sponte without notice ‘where the claimant cannot possibly win relief.’” . . . [W]e agree with and adopt the reasoning of the D.C. Circuit and hold that a sua sponte dismissal under Rule 12(b)(6) is not reversible error when it is “patently
“We have no occasion to pass judgment, however, on the permissible scope, if any, of *sua sponte* dismissals under Rule 12(b)(6).” The Court has not explicitly taken the question up since leaving it open in *Neitzke*.

The Supreme Court came close again in *Day v. McDonough*. In affirming the dismissal of a habeas petition on limitations grounds that the district court had raised *sua sponte*, Justice Ginsburg, writing for the majority in a 5–4 split, noted in passing that “[i]f course, before acting on its own initiative, a court must accord the parties fair notice and an opportunity to present their positions.” The Court, however, did not consider or comment on the circuit split discussed above, and did not consider the question more broadly, outside of the habeas context. And in saying what she did, Justice Ginsburg cited the Second Circuit in *Acosta v. Artuz*, where that court in turn cited its earlier opinion in *Snider* as follows: “*Unless it is unmistakably clear* that the court lacks jurisdiction, or that the complaint lacks merit or is otherwise defective, we believe it is bad practice for a district court to dismiss without affording a plaintiff the opportunity to be heard in opposition.” The use of “unless,” and the Supreme Court’s reliance on it, leaves unanswered the question whether notice and a hearing is required where a complaint is said to be defective on its face. It also raises the question of whether there is a different standard for non-habeas cases, as the Second Circuit’s non-habeas jurisprudence unmistakably requires notice and an opportunity to be heard in all such cases.

Perhaps more importantly, whatever the Supreme Court intended to communicate in *Day*, some circuits have not received the message—nor have district courts, if *Collins v. Merrill*, for example, is anything to go obvious” that the plaintiff could not prevail on the facts alleged. . . .

*Id.* at 365 (citations omitted).

73 *Neitzke*, 490 U.S. at 329 n.8.
75 *Id.* at 210.
76 *Id.* at 208–09.
77 *Id.* at 210.
78 *Acosta v. Artuz*, 221 F.3d 117, 124 (emphasis added) (citing *Snider v. Melindez*, 199 F.3d 108, 113 (2d Cir. 1999)).
79 *See Day*, 547 U.S. at 212 (Scalia, J., dissenting) (“The Civil Rules ‘govern the procedure in the United States district courts in all suits of a civil nature.’ This includes ‘proceedings for habeas corpus,’ but only ‘to the extent that the practice in such proceedings is not set forth in statutes of the United States or the Rules Governing Section 2254 Cases.’ Thus, ‘the Federal Rules of Civil Procedure apply in the context of habeas suits to the extent that they are not inconsistent with the Habeas Corpus Rules,’ and do not contradict or undermine the provisions of the habeas corpus statute”) (citations omitted).
Milani and Smith argue that, in another case, the Supreme Court effectively required notice and an opportunity to be heard prior to an appellate court acting sua sponte, but this is stretching too far what the Court actually said. Milani and Smith argue that “a close reading of the Supreme Court’s opinion in United States National Bank of Oregon v. Independent Insurance Agents of America Inc. suggests that sua sponte decision making is an abuse of judicial discretion.” Their argument is as follows:

As to whether the court of appeals abused its judicial discretion in addressing the issue, the Court pointed out that the court of appeals gave “the parties ample opportunity to address the issue” before making its decision. In view of this opportunity, the Court concluded that the court of appeals’ “decision to consider the issue was certainly no abuse of discretion.”

The Supreme Court’s emphasis on the fact that the parties were given “ample opportunity to address the issue” raised sua sponte by the court of appeals suggests that this fact led the Court to conclude that there was no abuse of judicial discretion. This language suggests that the result would have been different had the court of appeals not given the parties “ample opportunity” to address the sua sponte issue. Arguably then, Independent Insurance is Supreme Court authority for the proposition that an appellate court abuses its judicial discretion when it not only raises an issue sua sponte but decides that issue without giving the parties an opportunity to address it.

If Milani and Smith are saying that these comments in Independent Insurance are an indication of how the Supreme Court might rule on this issue if it was called on to decide it, I agree. (Day perhaps portends the same thing.) If they are saying that it would be right for the Supreme Court to rule this way, I also agree. I do not agree, however, that Independent Insurance is binding Supreme Court authority for the proposition that notice before a sua sponte decision is always required. At least some circuits certainly don’t see it that way. This is why I argue below that the Supreme Court should take a case in which this issue is squarely raised and clearly decide the issue once and for all.

---

82 Id. (footnote omitted).
83 Id. at 289–90 (footnotes omitted).
84 See Day, 547 U.S. at 210.
VI. WHY SUA SPONTE DISMISSALS ARE A BAD IDEA

There is a lot wrong with sua sponte dismissals. They are inconsistent with the adversary system, and change the judge’s role from referee to contestant.\(^{85}\) They can undermine respect for the legal system.\(^{86}\) And they increase the likelihood of errors, leading to unnecessary appeals and a waste of judicial resources.\(^{87}\) But most importantly, they lack the very due process the courts are supposed to safeguard.\(^{88}\)

A. Failure to Provide Due Process

Sua sponte decisions are inconsistent with due process.\(^{89}\) Period. There is no other way to look at it.\(^{90}\) Not only does a plaintiff surprised by a sua sponte dismissal not receive “due” process, she receives no process at all.\(^{91}\) She has no idea her lawsuit is in jeopardy of being dismissed, no idea what the reasons for that dismissal might be, and no opportunity to respond.\(^{92}\) This is the case whether the court’s dismissal decision is right or wrong.\(^{93}\) As Allan Vestal puts it:

> When [issues are] considered sua sponte both parties are taken completely by surprise and the court decides the matter on grounds not urged by either. Neither has had any opportunity to consider the matter, and both are now bound by res judicata grounded on considerations which represent not well reasoned positions for the litigants, but rather only the fortuitous decision of a wayward court.\(^{94}\)

The reference to res judicata here is important. As Milani and

---

\(^{85}\) Milani & Smith, supra note 5, at 272, 278.
\(^{86}\) Id. at 280, 295.
\(^{88}\) Id. at 262–63.
\(^{89}\) Offenkrantz & Lichter, supra note 8, at 132.
\(^{90}\) Id. at 131–33; Milani & Smith, supra note 5, at 262–63.
\(^{91}\) Milani & Smith, supra note 5, at 210 (“Just like the defendant in Nelson, the losing party in an appeal decided sua sponte only learns of the legal theory deemed controlling by the court when judgment is entered and never has an opportunity to rebut the court’s reasoning on whether, or how, that theory should apply.”).
\(^{92}\) For an interesting take on the issue from the perspective of the victims of sua sponte decision-making, see Crook, supra note 8, at 12.
\(^{93}\) Vestal, supra note 8, at 483–84 (1958).
\(^{94}\) Id. at 494.
Smith point out, the res judicata doctrine requires a party or its privy to be a participant in the former proceeding before the court can bind him to the consequences of that proceeding because, according to the Supreme Court, “The opportunity to be heard is an essential requisite of due process of law in judicial proceedings.” If this is the standard applied to former proceedings, how can it not apply to proceedings currently before the court? Lon Fuller once wrote of sua sponte decision-making:

[I]f the grounds for the decision fall completely outside the framework of the argument, making all that was discussed or proved at the hearing irrelevant . . . the adjudicative process has become a sham, for the parties’ participation in the decision has lost all meaning.

The situation is even harder to defend when there is no hearing at all.

B. Undermining Respect for the Legal System

The perception that the courts are regularly failing to provide due process cannot do anything but undermine respect for the legal system. Sir Robert Megarry, in the speech quoted at the beginning of this article, underlined the importance of sending the unsuccessful litigant away feeling as though he has had a fair hearing. Justice Harlan was obviously cognizant of this problem in his dissent in Mapp, when he warned that the Court’s sua sponte decision in that case was “not likely to promote respect . . . for the court’s adjudicatory process.”

This is not a farfetched concern. Offenkrantz and Lichter note that in the Second Circuit’s high-profile decision to “[sua sponte remove] Judge Shira Scheindlin from further proceedings in two stop-and-frisk cases,” an order which left the Judge “completely blindsided,” “newspapers were reporting that appellate courts had carte blanche to raise and decide important issues in a case without ever seeking the input

95 Milani & Smith, supra note 5, at 266 (quoting Postal Tel. Cable Co. v. City of Newport, 247 U.S. 464, 476 (1918)).
97 See Milani & Smith, supra note 5, at 304 (arguing that when an appellate court issues a sua sponte decision, it should grant motions for rehearing as a matter of right for the losing party).
98 Id. at 280.
99 See Megarry, supra note 1.
100 Id. at 410.
of any of the parties to it.” 102

Megarry tells a story of a client of his who had a fatal flaw in his case, but insisted on going ahead anyway. 103 Instead of seizing on the fatal flaw at the outset, the trial judge heard the case all the way through. 104 The client won on his two collateral points, but, as expected, lost on the key issue. 105 Megarry tells the story of what happened next:

The course taken by the judge must have prolonged the hearing by an hour or two. But the effect on the defeated tenant was striking. True, he had lost the last point and the case as a whole; but he had been victorious on the other two points. All that nonsense about the agent’s lack of authority and the letter not having been received in time had been blown away by the judge. It was a pity about the wording of the letter, of course; but he had seen his case being put in full, and none of his grievances had been left unheard or unresolved. 106

This is as it should be. Courts must not, as Megarry puts it, give in to “the temptation of brevity.” 107 Their very legitimacy hangs in the balance. A loss of respect for the courts marks the beginning of the unraveling of the rule of law. This is simply too high of a price to pay for efficiency.

C. Increased Likelihood of Errors

Justice Scalia once remarked that “the refusal to consider arguments not raised is a sound prudential practice.” 108 When a court decides sua sponte, it is deciding without input of the people who know the most about the case—the parties and their counsel. 109 This increases the likelihood that the court will miss some relevant statute, precedent, fact, or argument in making its decision. 110 That is especially so in the adversary system, which relies on the parties to provide the court with relevant information. 111 Judge Frank Easterbrook says, “It is hard

102 Offenkrantz & Lichter, supra note 8, at 113, 115, 125.
103 Megarry, supra note 1, at 410–11.
104 Id. at 411.
105 Id.
106 Id.
107 Megarry, supra note 1, at 410.
109 Shannon, supra, note 8, at 34–35.
110 Id. at 35.
111 “The rule that points not argued will not be considered is more than just a prudential rule of convenience; its observance, at least in the vast majority of cases, distinguishes our adversary system of justice from the inquisitorial one.” United States v. Burke, 504 U.S. 229, 246 (1992) (Scalia, J., concurring).
enough to navigate when the court sticks to questions fully ventilated by counsel.”112 Or as another (anonymous) judge put it, “There’s nothing worse than a lawyer being beaten by an assumption that simply is incorrect and wasn’t raised.”113

There is no doubt that such mistakes happen. Judges are only human, after all. In my own career, I have had at least two lower court decisions wrongly decided on grounds never argued that were easily overturned on appeal. One minute of oral argument or a one-page letter to the judge could have avoided an unnecessary appeal in those cases—if only those judges had asked. Many lawyers share this experience. This is what Justice Butler was getting at in his Erie dissent, when he argued that “[i]t may not justly be assumed that the labor and argument of counsel for the parties would not disclose the right conclusion and aid the Court in the statement of reasons to support it.”114

Milani and Smith give the powerful example of Poyner v. Loftus,115 the case that caused them to write their article.116 A blind plaintiff, who used neither a cane nor a seeing-eye dog, was severely injured when he fell from an elevated walkway because someone had moved a planter that he usually relied on to mark the edge.117 The D.C. Court of Appeals decided against him on a proposition of law, which applied in several states at the time, that a blind person is contributorily negligent if, when injured, he was using neither a dog nor a cane.118 The problem was that this law, which the court raised sua sponte, had been expressly abrogated by statute in the District of Columbia some 25 years earlier.119 The court was apparently unaware of the statute.120 So was the plaintiff, who accordingly did not seek a rehearing or other remedy.121 Not only did the court make a needless mistake, a grave injustice was done.122

113 See MARVELL, supra note 36, at 122 (reporting on a series of anonymous interviews with an unnamed “Midwestern appellate court” about its practice of sua sponte decision-making).
116 Milani & Smith, supra note 5, at 261 n.91.
117 Poyner, 694 A.2d at 70.
118 Milani & Smith, supra note 5, at 260 (citing Poyner v. Loftus, 694 A.2d 69 (D.C. 1997)).
119 Id. at 260–61.
120 Id. at 261.
121 Id. at 261 n.91.
122 See Milani & Smith, supra note 5, at 261.
There are many other examples. Indeed, several of the circuit court cases demonstrating the circuit split on sua sponte dismissals are cases in which, on appeal, a sua sponte dismissal was reversed because the court of appeals disagreed with the district court’s sua sponte reasoning. At least some of those reversals might have been avoided by providing the plaintiff with an opportunity to further explain its position or cite additional authority. We are left to wonder how many other sua sponte dismissals that are clearly wrong are never appealed by pro se litigants, those who are poorly advised, or those who simply don’t have the time or energy to appeal.

D. Stultification of Common Law

The common law was created by judges and juries recognizing legal rights and obligations that were previously unknown. That’s how our legal system works: one party cites a proposition of law, then the other makes an argument to distinguish a case, apply some competing doctrine, or create an exception. And, sometimes, as a result of this process of give-and-take, the law bends to accommodate new circumstances. Even in the statute-dominated world of the modern federal courts, such developments are not unheard of. Judges who cut this process short with a sua sponte dismissal are missing out on this key step in the adversary process. One author puts it this way:

Judges who decide cases without responding to the adversaries’ legal arguments may miss significant new points. Although judges are usually credited for landmark decisions, it is the litigants who usually suggest the

\[\text{References:} \]

123 See, e.g., Offenkrantz & Lichter, supra note 8, at 126–29 (discussing cases that are examples of appellate courts reaching erroneous results, which resulted in confusion and bad law).

124 See Perez v. Ortiz, 849 F.2d 793, 798 (2d Cir. 1988); Street v. Fair, 918 F.2d 269, 272 (1st Cir. 1990); but see Baker v. Dir. U.S. Parole Comm’n, 916 F.2d 725, 727 (D.C. Cir. 1990) (agreeing with district court that sua sponte dismissal was appropriate because plaintiff could not possibly prevail on the facts alleged in the complaint).

125 See Christopher J. Peters, Adjudication as Representation, 97 COLUM. L. REV. 312, 360 (1997) (“Court decisions, we know, can become rules of sorts. What we call ‘the common law method’ means that cases are decided by analogy, by comparison of their facts to the facts of previously decided cases and conformity of their results to the results of those past cases.”).

126 See id. at 361–62.

127 See id. at 363.

128 See, e.g., Planned Parenthood of Southwestern Pennsylvania v. Casey, 505 U.S. 833 (1992) (plurality opinion) (claiming to reaffirm the “essential holding” of Roe v. Wade under the doctrine of stare decisis, but creating a new test for establishing when a state’s interest in a viable fetus begins).
If judges routinely dismiss cases that challenge existing doctrine without hearing any arguments about why that doctrine is wrong, or doesn’t apply, or should be modified, the process of common-law development is frustrated and cut short. As the Supreme Court of Canada said in a leading case about a Canadian counterpart to Rule 12(b)(6):

The law is not static and unchanging. Actions that yesterday were deemed hopeless may tomorrow succeed. Before *Donoghue v. Stevenson* introduced a general duty of care to one’s neighbour premised on foreseeability, few would have predicted that, absent a contractual relationship, a bottling company could be held liable for physical injury and emotional trauma resulting from a snail in a bottle of ginger beer. Before *Hedley Byrne & Co. v. Heller & Partners Ltd.*, a tort action for negligent misstatement would have been regarded as incapable of success. The history of our law reveals that often new developments in the law first surface on motions to strike or similar preliminary motions, like the one at issue in *Donoghue v. Stevenson*. Therefore, on a motion to strike, it is not determinative that the law has not yet recognized the particular claim.

This is not a phenomenon unique to Canada. Justice Marshall said much the same thing in *Neitzke*:

Close questions of federal law . . . have on a number of occasions arisen on motions to dismiss for failure to state a claim, and have been substantial enough to warrant this Court’s granting review, under its certiorari jurisdiction, to resolve them. It can hardly be said that the substantial legal claims raised in these cases were so defective that they should never have been brought at the outset. . . . Indeed, we recently reviewed the dismissal under Rule 12(b)(6) of a complaint based on 42 U.S.C. § 1983 and found by a 9-to-0 vote that it had, in fact, stated a cognizable claim—a powerful illustration that a finding of a failure to state a claim does not invariably mean that the claim is without arguable merit.

---


130 For example, Justice Butler urged the Court in *Erie* to seek the input of counsel before deciding to set aside current law in an attempt to reach a more precise and thorough explanation for why the Court was departing from an older doctrine. *Erie R.R. v. Tompkins*, 304 U.S. 64, 87 (1938) (Butler, J., dissenting).


Just as they need to be open-minded about facts, judges should always be prepared to listen to arguments about why the law is not as it appears to be, or should be something else. Sua sponte dismissals are fundamentally incompatible with this approach to deciding cases.

VII. ARE THERE GOOD REASONS FOR SUA SPONTE DISMISSALS?

The typical reason that courts give for allowing sua sponte dismissals is that providing due process in a hopeless case would make no difference because the case would be dismissed in any event, thus wasting judicial resources. Getting rid of such cases quickly allows the court to spend more time dealing with those cases that do have merit.133

It does not take much thinking to recognize that these arguments are unsound. For example, most criminal defendants are probably guilty—some quite obviously so.134 Would this ever justify a conviction sua sponte without a trial or any opportunity to argue innocence? That would, after all, allow the court to spend more time on the close calls and difficult decisions. Of course, in civil cases, liberty is not at stake, so maybe due process is less important.135 But is that really true? The absence of a fundamental liberty interest might justify a lesser form of process, but surely it cannot justify eliminating due process altogether.136

Another argument for sua sponte dismissal is that if a plaintiff really does have a meritorious argument, it can always seek post-dismissal relief under Rule 60(b) of the Federal Rules of Civil Procedure.137 This would allow the court to quickly weed out hopeless

---

133 See, e.g., Baker v. Dir. U.S. Parole Comm’n, 916 F.2d 725, 726 (D.C. Cir. 1990) (noting that adherence to a standard prohibiting all sua sponte dismissals granted under Rule 12(b)(6) would lead to a waste of judicial resources).

134 See Criminal Cases, BUREAU OF JUSTICE STATISTICS, https://www.bjs.gov/index.cfm?ty=p&t&tid=23 (last visited Oct. 7, 2017) (“More than three-fourths of felony defendants had a prior arrest history, with 69% having multiple prior arrests. . . . About two-thirds of felony defendants were eventually convicted and more than 95% of these convictions occurred through a guilty plea.”).

135 See, e.g., Mathews v. Eldridge, 424 U.S. 319, 334 (1976) (“Due process is flexible and calls for such procedural protections as the particular situation demands.”) (quoting Morrissey v. Brewer, 408 U.S. 471, 481 (1972)).

136 See Offenkrantz & Lichter, supra note 8, at 133 (suggesting that applying the factors derived from Mathews v. Eldridge leads to the conclusion that appellate courts deciding a case sua sponte may permissibly relax due process requirements, but also recognizes that there is a “strong private interest in the result of litigation”).

137 See Milani & Smith, supra note 5, at 263 n.98 (suggesting that “a post-judgment opportunity to be heard may satisfy due process requirements”); Paterno v. Lyons, 334 U.S. 314, 319 (1948) (noting that the petitioner could have challenged any errors by appealing the judgment, withdrawing his guilty plea, or filing a motion in arrest of judgment); Curley v.
cases, but would still allow continuation of the few cases in which the pleadings can be repaired or some truly novel argument made.

But do we really want a legal system in which people receive due process only if they know enough to ask for it? Doesn’t the Constitution guarantee due process to everybody who comes before the court—both those who are procedurally savvy and those who are not? Use of sua sponte dismissal in this way would create two classes of plaintiffs: those who are sophisticated enough to ask for due process and those who are not. While some might argue that differences in treatment are inevitable (and therefore tolerable) where parties have different resources at their disposal and can therefore hire more or less sophisticated lawyers who can spend more or less time on their case, the courts should at least try to give everybody equal treatment when it comes to basic due process. In fact, it is those with the least access to high-powered lawyers that most need a judge who will carefully listen to the substance of their arguments, however unartfully they may be framed. 138

Arguments that a wrongly made sua sponte dismissal can always be cured by an appeal are even less persuasive. First of all, an unnecessary appeal involves an enormous waste of judicial resources—the very type of waste the sua sponte dismissal was supposed to avoid. 139 And in some cases, an appeal is not available, for example, when the court sua sponte remands to a state court after removal. 140 But perhaps most importantly, what kind of message does it send if we tell litigants that they are entitled to due process only if they appeal? Doesn’t that come with a huge risk of creating the perception that the district court does not have time to provide due process, and that parties must go elsewhere if they want it?

Perry, 246 F.3d 1278, 1284 (10th Cir. 2001) (suggesting aggrieved litigant can seek relief pursuant to Rule 59(e) or 60(b) of the Federal Rules of Civil Procedure); Turner v. Kirkwood, 62 F.2d 256, 260 (10th Cir. 1932) (dismissal without notice was proper since the appellant was heard in court several times before the court dismissed the case).

138 Shannon, supra note 8, at 33 (suggesting that the court’s duty to administer justice sometimes requires the court to assist pro se litigants and those who are underrepresented).
139 See, e.g., Perez v. Ortiz, 849 F.2d 793, 797 (2d Cir. 1988).
140 14C CHARLES ALAN WRIGHT ET AL., FEDERAL PRACTICE AND PROCEDURE § 3740 (4th ed. 2009) states:

Although there are statutory exceptions to the no-appeal rule and judge-made exceptions that also have developed . . . in general the Supreme Court has made it clear that so long as the order to remand was based on the grounds set out in Section 1447(c)—1) a lack of subject-matter jurisdiction in the district court and 2) any defect in the removal procedure—the correctness of that decision is unreviewable.
VIII. ALTERNATIVES TO SUA SPONTE DISMISSALS

What can a district court do if it receives a complaint that is clearly ill founded, perhaps even displaying on its face the court’s lack of jurisdiction? The simplest way to proceed would be to do exactly what the court is inclined to do—draft an order and explanatory opinion dismissing the case—but, before filing it, send the draft opinion and order to the plaintiff, giving the plaintiff a short period of time to respond either by amendment or argument. If, as is bound to happen in the vast majority of fatally flawed cases, whatever the plaintiff sends to the court in response does not change the court’s mind, the court can simply file its draft opinion and order. All that has been lost is the price of a stamp and a few days of grace. At least the plaintiff has been given the due process that the Constitution guarantees—and that the courts are supposed to jealously safeguard. And in that minority of cases in which the plaintiff’s arguments or amendments change the court’s mind, error, injustice, and a waste of judicial resources will have been avoided.

Some circuits that require notice before dismissal are more prescriptive about process. In Tingler v. Marshall, for example, the Sixth Circuit adopted a procedurally rigorous approach, saying:

[A] district court faced with a complaint which it believes may be subject to dismissal must: (1) allow service of the complaint upon the defendant; (2) notify all parties of its intent to dismiss the complaint; (3) give the plaintiff a chance to either amend his complaint or respond to the reasons stated by the district court in its notice of intended sua sponte dismissal; (4) give the defendant a chance to respond or file an answer or motions; and (5) if the claim is dismissed, state its reasons for the dismissal.

Of course there are an infinite variety of other procedures that could be adopted. They do not always have to involve high ceremony or

---

141 The comparable step in appellate proceedings that several authors point to as being often—but not always—invoked is to call for further briefing or a rehearing when the court identifies a new issue sua sponte. The quintessential example of this was in Brown v. Board of Education, 347 U.S. 483 (1954), “where the Court asked for, not one, but two sets of supplemental briefs and arguments.” Milani & Smith, supra note 5, at 296; see also Megarry, supra note 1, at 413–14; see also Offenkrantz & Lichter, supra note 8, at 119–121.

142 Tingler v. Marshall, 716 F.2d 1109, 1112 (6th Cir. 1983).

143 Zachary, supra note 37, proposes a framework for dealing with “Problems Posed by the ‘Delusional’ or ‘Wholly Incredible’ Complaint” under the Prison Litigation Reform Act. Notably, he argues that notice and an opportunity to be heard are not preconditions to dismissal. Michelle Lawner argues that district courts should raise sovereign immunity sua sponte when it appears to apply, but proposes that “the court would ask the parties to brief the
significant resources. (The Sixth Circuit recognized this when it relaxed its Tingler requirements for service on the defendant in Morrison v. Tomano, holding that “sua sponte dismissal for failure to state a claim is not necessarily rendered invalid because of lack of service on the defendant or failure to provide the defendant an opportunity to respond.”144) The courts might even develop a form listing the ten most common fatal defects in a pleading, and when a judge finds those problems, the judge could tick the applicable box or boxes, sign the form, and send it to the plaintiff in the form of a show cause order. Any procedure will do, as long as it is fair, gives a plaintiff some notice of the reasons why its claim appears to be fatally flawed, and provides some opportunity to be heard before the court makes a decision.

IX. CONCLUSION: THE SUPREME COURT SHOULD BAN SUA SPONTE DISMISSELS

The Supreme Court should resolve the circuit split on the practice of sua sponte dismissals sooner rather than later. How and why? The “how” is easy: the Court should embrace the position taken by the First, Second, Sixth, and Eleventh Circuits that a failure to provide notice and an opportunity to be heard before a dismissal is reversible error. Nothing less will do. No other position is compatible with the requirement for basic due process. District courts should have maximum flexibility to determine how they can most efficiently provide a litigant who files a hopeless claim with due process before dismissal. They can’t, however, cut out due process altogether.

The “why” is also pretty simple. First, although Day and Independent Insurance provide hints of how the Supreme Court would settle this question, in at least some circuits the hints are not loud enough, so the question needs to be decided clearly and directly—with clear reference to the current circuit split. The Court should explicitly overrule Omar and its progeny.

Second, the district courts are the main point of interaction between the public and the legal system—if they don’t appear to be fair, no court will. The current practice in some circuits of adopting a “no harm, no foul” approach to the failure to provide due process before a sua sponte

144 Morrison v. Tomano, 755 F.2d 515, 517 (6th Cir. 1985).
dismissal undermines the public’s relationship with, and confidence in, the district courts. It is time for the Supreme Court to fix this.

The public must have absolute confidence that, like it or not, when it comes to outcomes, the federal courts at least treat litigants fairly. The district courts need a strong message from the Supreme Court that, busy or not, basic due process is too important a value to compromise in the name of efficiency. Litigants should not have to appeal just to get a fair hearing—or any hearing at all. Sir Robert Megarry was right to remind us that “Justice in full takes time; but often it is time well spent.”

145 Megarry, supra note 1, at 411.
THE PARADOX OF MILITARY TRIBUNALS THAT TRY CIVILIANS

Golan Luzon*

I. INTRODUCTION ................................................................................ 53
II. MILITARY TRIBUNALS UNDER MARTIAL LAW .............................. 56
III. MILITARY TRIBUNALS FOR POPULATIONS UNDER MILITARY OCCUPATION......................................................... 60
IV. MILITARY TRIBUNALS AGAINST CIVILIANS IDENTIFIED AS THE ENEMY ........................................................................ 69
V. SUMMARY .......................................................................................... 76

I. INTRODUCTION

In most countries, criminal matters are tried in civilian criminal courts, which are part of the general judicial system of the State. By contrast, military tribunals, which encompass any military court, litigate particular matters, usually under a special system of laws. Military tribunals operate within the framework of the military forces, and their function is primarily to try soldiers who are subject to military rule for criminal offenses. A military tribunal has the authority to determine the guilt of the armed forces personnel subject to military law and, if the defendant is found guilty, to decide his punishment.

The present article deals with an additional function of military tribunals: subjecting citizens who are not members of the armed forces to military law. Many regimes that protect civil, political, and property

* Golan Luzon is a post-doctoral research fellow in the Faculty of Law at Bar-Ilan University, Israel, and a visiting scholar in the Research Unit Criminal Law and Criminology, University of Leuven, Belgium.

1 EUGENE R. FIDELL, MILITARY JUSTICE: A VERY SHORT INTRODUCTION xx (2016).
2 Id. at xxi, 111.
3 Id. at 27.
("CPP") rights have utilized this function, especially when martial law is declared, but this is not the only context in which military tribunals try citizens who are not ordinarily subject to military law.6 For example, military tribunals are used for the prosecution of prisoners of war for war crimes.7 Historically, the need for military tribunals did not arise because of the abolition of the civilian justice system, but because of the system’s inability to function in emergency situations or when war is raging, making it technically impossible to conduct an orderly legal process.8 Military trials are conducted by officers who are not necessarily jurists, and the procedures are different from those in civil courts.9 For example, the judge may also be the prosecutor, and evidentiary and procedural rules may differ from those found in civilian courts.10

One of the fundamental principles of the rule of law requires that citizens be tried in civilian courts before professional judges, whose
personal independence is guaranteed (as opposed to a military judge, who is subject to the authority of the higher ranking officer and may be
dismissed at any stage of the trial). It is theoretically possible to create
a military judicial framework that meets all the fundamental
requirements of the rule of law, but there is probably no logical reason to
allow civilians to be tried within such a framework.

Unlike civil courts, proceedings in military tribunals are usually
quick, laconic, and characterized by military discipline and unanimity. 
Proceedings tend to be quick because they take place in a hierarchal
system. Military tribunals often require different standards of proof,
are usually not subject to review by judicial civil courts, and allow little
time for gathering evidence. Hearings before military tribunals often
take place in a charged atmosphere, where the principle of necessity
reigns supreme and considerations of public safety, intelligence
gathering, and national security override individual rights.

The jurisdiction of military tribunals to try soldiers, reservists, and
civilian members of the armed forces is based on a written or unwritten

---

11 A judge’s personal independence is still a cause of concern in many countries,
including the United States, primarily because many judges in military tribunals do not have
guaranteed term limits, which allows the judge to be removed at any time. FIDELL, supra note
1, at 69. To this day, the Army and Coast Guard are the only two U.S. military branches with
fixed term limits for judges. Id.

12 This can be attributed to the elaborate and unique structure of the armed services, in
addition to the military’s need to ensure order, set uniform objectives, and achieve military
goals. Id. at 1–2; MINDIA VASHAKMADZE, GUIDEBOOK: UNDERSTANDING MILITARY JUSTICE

13 FIDELL, supra note 1, at 49; VASHAKMADZE, supra note 12, at 10 (“The main
rationale for a separate court system is the unique character of military life, where discipline,
organization and hierarchy play a crucial role. These are fundamental for maintaining the
effectiveness and combat readiness of the armed forces. Cases must be dealt with quickly and
sentences for certain offences can be severe.”).

14 President George W. Bush commented that, given the endangerment of United States
citizens post 9/11, it would be impractical to apply the principles of law and evidence in
civilian criminal cases to military tribunals. MAROUF HASIAN JR., IN THE NAME OF
Although the United States Supreme Court was given jurisdiction over court-martial appeals
in 1984, many courts-martial are still not eligible for the Court’s review. FIDELL, supra note
2, at 5–6. The military justice system also allows a practice known as “pleading for the
contingencies of proof,” in which prosecutors are permitted to charge the same offense in a
multitude of ways. Id. at 76–77.

15 For example, the war against terrorism raised questions over the State’s “necessity” to
circumvent individual freedoms to combat terrorism. HASIAN, supra note 14, at 5; see also
CLINTON L. ROSSITER, CONSTITUTIONAL DICTATORSHIP: CRISIS GOVERNMENT IN THE
MODERN DEMOCRACIES 11 (1948) (explaining that there are instances in which a national
emergency has prompted governments to act in a manner contrary to the State’s law,
constitution, or customs).
agreement under which potential defendants “waive” their right to be tried in a civilian court in certain matters when they agree to become part of an organization called the “military.” The source of authority for military tribunals to try civilians of their country or of a foreign country, however, is not self-evident. Unlike the civilian judicial system, the military judicial system was not designed to be a comprehensive legal mechanism that provides solutions for settling disputes and striking a balance between various considerations, such as deterrence and rehabilitation. Rather, the main purpose of the military system is to enforce the security policy of the country at a given time.

The present article distinguishes between three types of military tribunals and examines the source of their jurisdiction to civilians and their justifications. In the process, it uncovers that each type of military tribunal is riddled with paradox, which manifests in time if it remains in existence long enough, leading to inherent self-contradiction that negates its own raison d’être. In Part II, I discuss the military tribunals operated by states to try civilian populations under martial law in times of emergency and crisis. In Part III, I describe military tribunals that states operate in territories occupied during war, within the framework of a military government. Part IV discusses a new type of military tribunal established in the U.S. to handle the problem of unlawful combatants. The article concludes by showing how, with the injection of the element of time, all types discussed end in paradox that nullifies their justification.

II. MILITARY TRIBUNALS UNDER MARTIAL LAW

To understand the source military tribunals use to establish jurisdiction over civilians, it is necessary to distinguish between a military government and martial law. A military government can take many forms, including military occupation, martial law, military

---

16 See Robinson O. Everett, Persons Who Can Be Tried by Court-Martial, 5 J. PUB. L. 148–49 (1956) (stating that if an individual has taken an oath of allegiance, he will have subjected himself to military jurisdiction; although if a person does not take the oath, he may still be subject to military jurisdiction if he has continued with the life of a soldier).

17 Stephanie Simmons, When Restoration to Duty and Full Rehabilitation is Not a Concern: An Evaluation of the United States Armed Forces, 34 NEW ENG. J. ON CRIM. & CIV. CONFINEMENT 105 (2008).

18 Id. (“The purpose of the military system of law is to promote justice, assist in maintaining good order and discipline in the U.S. Armed Forces, and promote efficiency and effectiveness in the military establishment, thereby strengthening the national security of the United States.”); FIDELL, supra note 1, at 2.
dictatorship, and stratocracy. The institution known as martial law is both an extension of military government and a classic institution of common law constitutional dictatorship that exists in regimes that protect CPP rights. The circumstances that allow the imposition of martial law are usually rebellion or invasion.

Under martial law, military authority is imposed on the sovereign territory of the State and on its civilian population during an external invasion or internal rebellion. The guiding principle of martial law is that of “necessity.” In other words, the circumstances justify taking all measures that authorities consider necessary to repel an invasion or suppress a rebellion. Because of the difficult circumstances, means considered illegal in peacetime become legitimate. For example, under certain circumstances, measures that violate civil or property rights do not constitute an obstacle to authorities operating under martial law if they are essential for maintaining public order and security.

In practice, martial law appoints a high-ranking military officer as the military governor or head of government, removing all authority from the executive, legislative, and judicial branches of the previous government. On occasion, it is the government itself that declares martial law in order to impose its rule on the public. Such action can be taken after a coup, during popular protest, to repress political
opposition, or to prevent insurrections.

The U.S. Supreme Court has already addressed the issue of the authority of martial law in the Ex Parte Milligan affair, which occurred during the Civil War, and outlined clear criteria for the imposition of martial law on the civilian population. The Court opined that, according to the Constitution, checks and balances curb the powers of the President, Congress, and military commanders and, even in times of emergency, civil liberties should not be abrogated and citizens should not be subject to a single branch of government’s sole discretion to impose martial law. The U.S. also resorted to martial law during World War II. Immediately after the bombing of Pearl Harbor, the governor of Hawaii suspended habeas corpus, declared martial law, and transferred his powers to the military commander until the danger of invasion passed.

Under martial law, rules regarding what are considered legal and illegal are not defined in advance. The commander who imposes his authority has broad discretion to act according to changing circumstances and conditions. When order is restored, however, those who assumed exceptional powers should be prepared to prove in court that the prevailing conditions justified the imposition of martial law.

Martial law is associated with the operation of military tribunals that try civilians. Its imposition involves procedures such as curfews, the


See id. (discussing that as the conditions of violence that spark martial law “vary in intensity,” so does the severity of the measures adopted in order to preserve public safety). See id. at 143; see also Duncan, 327 U.S. at 307.
suspension of civil law and civil rights, abolishing or suspending habeas corpus, and applying military law or military justice to civilians.\(^{39}\) 

Citizens who challenge martial law can be brought before a military tribunal.\(^{40}\)

The conceptual basis of military tribunals, as it emerged from common law, is that the regular courts do not stop operating by virtue of some declaration, but because the conditions prevailing in their jurisdiction do not allow them to meet, hold hearings, and ensure that their verdicts are enforced.\(^{41}\) The military rule established to replace the civilian one serves as an administrative structure, whose function is to restore public order.\(^{42}\) Under some circumstances, government authorities can punish and carry out decrees and decisions without a trial, but instead, they establish special courts for purposes of punishment and enforcement.\(^{43}\) In this way, they ensure a degree of institutionalization and protection from arbitrariness.\(^{44}\)

In recent decades, martial law was imposed for long periods of time in several countries that protect CPP rights, such as Israel and Taiwan, during which military tribunals for the civilian population undermined the basic legal rights of citizens.\(^{45}\) We justify the jurisdiction of military tribunals for trying civilians because of the crisis and risk of danger to public safety; to overcome the crisis, we temporarily surrender our basic

\(^{39}\) See Rossiter, supra note 15, at 145 (“The government may wield arbitrary powers of police to allay disorder, arrest and detain without trial all citizens taking part in this disorder and even punish them (in other words, suspend the writ of habeas corpus), institute searches and seizures without warrant, forbid public assemblies, set curfew hours, suppress all freedom of expression, institute courts-martial for the summary trial of crimes perpetrated in the course of this regime . . . .”) (footnotes omitted).

\(^{40}\) Id. A prominent example of this occurred during the U.S. Civil War, where President Lincoln subjected those who supported the Confederacy to martial law; those citizens who challenged the martial law were subjected to military tribunals. Ex Parte Milligan, 70 U.S. 2, 15–16 (1886).

\(^{41}\) See Rossiter, supra note 15, at 147 (“[U]nder martial law [civil courts] are not superseded by the proclamation, but by the fact that conditions are so disturbed that they cannot sit.”).

\(^{42}\) Id. at 148.

\(^{43}\) Id.

\(^{44}\) See id. (discussing that the special court is intended to allow “some sort of institutional form”). But cf. Rossiter, supra note 15, at 148 (these special courts have been criticized as being “quite illusory” rather than ensuring a degree of institutionalization or protection from arbitrariness).

legal rights for the sake of coping with the crisis. 46 The assumption is that military tribunals are more efficient in solving crises because they need not contend with the restrictions imposed by civil law. 47 These restrictions are intended to protect the rights of citizens when business is as usual. 48 It is reasonable to say that the civil legal system that protects the rights of citizens is based on the very existence of the regime that enables such a legal system, and the moment the regime is in danger, by definition, the legal system is liable to collapse together with the regime. This is what justifies the temporary imposition of martial law. If martial law is extended indefinitely—or beyond what is necessary to restore order—this justification is no longer adequate. The objective of protecting the freedoms guaranteed by the legal system of the endangered regime, which martial law was intended to save, is scuttled by the long duration of martial law itself, which ipso facto takes away these freedoms. There is no good solution to this inherent contradiction between the objectives and results of long-term martial law. The protracted Emergency Law in Egypt, which was in effect for many years, serves as a good example of martial law, although by a different name. 49

III. MILITARY TRIBUNALS FOR POPULATIONS UNDER MILITARY OCCUPATION

Military occupation is a State’s effective temporary control over a territory that is not under its formal sovereignty and is against the will of the residents. 50 Under military occupation, the controlling power does

46 See Rossiter, supra note 15, at 142–43.
47 See id. at 9 (“The institution of martial rule is a recognition that there are times in the lives of all communities when crisis has so completely disrupted the normal workings of government that the military is the only power remaining that can restore public order and secure the execution of the laws.”).
48 See W.A. Graham, Martial Law in California, 31 CAL. L. REV. 6, 10 (1942) (“If [martial law’s] necessary operation shall interfere with business as usual, business must understand that but for such interference, no businesses may exist except the business of the enemy.”).
49 See Egyptian Emergency Laws, HARV. DIVINITY SCH. RELIGIOUS LITERACY PROJECT, https://rlp.hds.harvard.edu/faq/egyptian-emergency-law (last visited Aug. 3, 2017) (stating that the initial emergency law was enacted in 1981, but was extended until it was dissolved after the Arab Spring in 2011).
50 Naz K. Modirzadeh, Dark Sides of Convergence: A Pro-Civilian Critique of the Extraterritorial Application of Human Rights Law in Armed Conflict, 86 INT’L STUD. SER. US NAVAL WAR COL. 349, 365 (2010); Military Occupation, Black’s Law Dictionary (10th ed. 2014); see also Marc Cogen, Democracies and the Shock of War: The Law as a
not grant the civil rights its own citizens enjoy to the population under its control because the military government rules the occupied territory, which may be characterized as supervising the occupied territory.\textsuperscript{51} There is no obligation to announce an official start of the military government and no need for a declaration on the part of the commander of the occupying forces about instituting the military government.\textsuperscript{52} The military government is the result of facts on the ground: the earlier sovereign has been defeated and deposed, and control is now in the hands of the rival military.\textsuperscript{53} An announcement, however, is useful in informing all residents of the occupied area about the laws and regulations guiding the occupier in the exercise of its power.\textsuperscript{54}

The rules of military government are enshrined in various international agreements, particularly the Hague Convention of 1907\textsuperscript{55} and the Fourth Geneva Convention of 1949.\textsuperscript{56} Other rules are derived from practices in each country. Therefore, the legal source of authority and of the power of the State to establish a military tribunal for citizens in an occupied territory is public international law, which is part of the laws of war.\textsuperscript{57} The obligation of the occupying State toward the international community is based on the customary international law and on applicable international treaties to which the State is party.\textsuperscript{58}

Article 42 of the Hague Convention states that a “[t]erritory is considered occupied when it is actually placed under the authority of the hostile army. The occupation extends only to the territory where such authority has been established and can be exercised.”\textsuperscript{59} Article 43 of the Hague Convention defines the authority of the military government as follows:

The authority of the legitimate power having in fact passed into the hands of the occupant, the latter shall take all the measures in his power to restore, and ensure, as far as possible, public order and safety, while respecting, unless

\textsuperscript{52} Id. at 220.
\textsuperscript{53} Id. at 213.
\textsuperscript{54} Id. at 220.
\textsuperscript{55} See generally Hague Convention (IV) Respecting the Laws and Customs of War on Land, Oct. 18, 1907, 36 Stat. 2277 [hereinafter Hague Convention].
\textsuperscript{57} Greenspan, supra note 51, at 5.
\textsuperscript{58} Id. at 4–5.
\textsuperscript{59} Hague Convention, supra note 55, at art. 42.
absolutely prevented, the laws in force in the country. 60

These provisions of the Hague Convention revolve around two main issues: ensuring the legitimate security interests of the occupier and meeting the needs of the civilian population in the occupied territory. The regulations create a certain balance between the two issues: in some matters the emphasis is on military necessity and in others the emphasis is on the needs of the civilian population. The laws of war usually create a delicate balance between two separate interests in military necessity and humanitarian considerations. 61 In both military and civilian matters, the fundamental premise is that the military commander does not inherit the rights and status of the defeated regime and is not sovereign in the occupied territory. 62

Article 43 of the Hague Convention refers to the authority of the military government to restore and ensure “public order and safety.” 63 This involves both the security needs of the military government itself and the needs of the civilian population under its control. 64 Article 43, though not explicitly stated, may implicitly grant the military government the ability to establish military tribunals for trying the population of the occupied territory when “absolutely prevented” from respecting “the laws in force in the country,” but it also imposes a duty to restore and ensure public order and safety. 65 This authority is twofold: first, restoration of public order wherever it had been disrupted and, second, ensuring the continued maintenance of public order and safety.

In sum, according to the Hague Convention, the powers of the defeated regime are suspended and, by virtue of public international law, administrative and governmental authority in the occupied area is granted to the military commander, which implicitly includes the authority to establish tribunals for trying the local population. 66 From the legal point of view, these powers are inherently temporary, as is the

60 Id. at art. 43.
64 Hague Convention, supra note 55, at art. 43.
65 Hague Convention, supra note 55, at art. 43.
66 The duties of the occupying power include the duty to ensure law and order and public safety. Id. In conjunction with international law, such as Article 66 of the Fourth Geneva Convention, an element on ensuring law and order could include the formation of military courts or tribunals. Geneva Convention, supra note 56, at art. 66.
occupation of territory.\textsuperscript{67}

Article 64 of the Fourth Geneva Convention talks about observing the laws and making changes in the occupied territory:

The penal laws of the occupied territory shall remain in force, with the exception that they may be repealed or suspended by the Occupying Power in cases where they constitute a threat to its security or an obstacle to the application of the present Convention. Subject to the latter consideration and to the necessity for ensuring the effective administration of justice, the tribunals of the occupied territory shall continue to function in respect of all offences covered by the said laws.\textsuperscript{68}

Similar to Article 43 of the Hague Convention, Article 64 of the Geneva Convention demands that the military government preserve the penal law and the existing courts in the occupied territory, but it also allows the military government to establish its own tribunals and to change or suspend the existing laws, if it is necessary for maintaining security and in order to implement the Convention.\textsuperscript{69}

Article 66 of the Fourth Geneva Convention explicitly refers to trying the population of the occupied territory by military tribunals:

In case of a breach of the penal provisions promulgated by it in virtue of the second paragraph of Article 64, the Occupying Power may hand over the accused to its properly constituted, non-political military courts, on condition that the said courts sit in the occupied country. Courts of appeal shall preferably sit in the occupied country.\textsuperscript{70}

Thus, when read together, the articles of the Hague and Geneva Conventions grant the occupying power the authority to institute a military government and military tribunals for civilians within the occupied territory. But the duration of the military government that operates the military tribunals is expected to be limited, and Article 6 of the Fourth Geneva Convention explicitly limits the amount of time that the military government can be in effect:

The present Convention shall apply from the outset of any conflict or


\textsuperscript{68} Geneva Convention, \textit{supra} note 56, at art. 64.

\textsuperscript{69} \textit{Id}.

\textsuperscript{70} \textit{Id.} at art. 66.
occupation mentioned in Article 2. In the territory of Parties to the conflict, the
application of the present Convention shall cease on the general close of
military operations. In the case of occupied territory, the application of the
present Convention shall cease one year after the general close of military
operations. . . .

When comparing military tribunals under martial law and military
government, there is yet another parameter to consider: the fact that
citizens standing trial under the military government are, for the most
part, not citizens of the occupying state. Their interest in military
tribunals is minimal and, at times, such tribunals are entirely contrary to
their interests because of the severe harm the tribunals cause to their
rights. Article 67 of the Fourth Geneva Convention seeks to reduce the
harm:

The courts shall apply only those provisions of law which were applicable
prior to the offence, and which are in accordance with general principles of
law, in particular the principle that the penalty shall be proportionate to the
offence. They shall take into consideration the fact that the accused is not a
national of the Occupying Power.

Therefore, it is required of military tribunals to take into account the fact
that the defendant is not a citizen of the occupying power. This
requirement may be regarded as yet another balancing move, although it
does not clarify how the military tribunals should accommodate
defendants who are not citizens of the state.

As noted at the outset, it is not desirable to have military tribunals
judge civilians; there is almost no disagreement about the fact that
military tribunals, by virtue of their very existence, violate civil liberties.
Under martial law, when the State tries its own citizens in military
tribunals, the balance is between violation of citizens’ rights in the short
term and restoration of order and security. Such tribunals are expected
to operate for a limited time only. Under military government in an

71 Id. at art. 6.
72 See GREENSPAN, supra note 51, at 254–56.
74 Geneva Convention, supra note 56, at art. 67.
75 See ROSSITER, supra note 15, at 11.
occupied territory, where military tribunals are instituted for civilians, it is of special significance to delimit the jurisdiction of these tribunals while taking into account the length of time during which they operate and the fact that the tribunals are not operated by citizens of the state.

The key question underlying the jurisdiction of military tribunals to try civilians who are not citizens of the State concerns the scope of the authority military tribunals receive to ensure public order and safety. Does the military government have the same authority as the regular government that imposes martial law, or does the fact that it is a military government restricts its options? The provisions of international law indicate that the scope of the authority of the military government is restricted by two main parameters, which both involve a dimension of time: the obligation of the military government to provide good governance that cares for the local population in all areas of life and the limitations that the military government faces for being a temporary administration rather than a permanent government or sovereign.

The first parameter tends to expand the authority and obligations of the military government, to the point of creating an analogy with regular government. The second parameter tends to narrow its authority and obligations by creating a significant difference between the scope of authority of the military government and those of regular government. The proper scope of the jurisdiction of military tribunals under a military government appears to be the product of the cumulative and offsetting effect of these two parameters. The first parameter determines the optimal boundary of the authority and the second parameter restricts it.

The first parameter, which determines the scope of the jurisdiction of military tribunals, is concerned with good governance that cares for the local population. To understand the meaning of this parameter, it is necessary to examine the authority and obligations of good governance in general. In determining the scope of the authority of the military government, in accordance with Article 43 of the Hague Convention regarding the preservation of “public order and safety,” the distinction (discussing the fact that a declaration of martial law is only made out of necessity, therefore, trials of civilians by military tribunals may only occur as long as there remains a “reasonably sound factual basis” for the use of the military tribunal).

77 See, e.g., JULIUS STONE, LEGAL CONTROLS OF INTERNATIONAL CONFLICT: A TREATISE ON THE DYNAMICS OF DISPUTES- AND WAR-LAW 727 (1954) (stating that the Hague Regulations sought to negate the status of the occupying power with the status of the Sovereign; however, this negation does not in itself delimit the powers of the occupant).

78 See Hague Convention, supra note 55, at art. 43; see also Geneva Convention, supra note 56, at art. 70.
between short-term military government (akin to martial law) and long-term military government must be taken into account.  

Whether considering it from the perspective of the military, of the population, or a combination of the two, the element of time affects the scope of the authority. Naturally, in the case of a short military occupation, military and security needs are paramount, similar to martial law. In the case of a long-term military occupation, however, the needs of the local population receive greater emphasis. For example, trying the population in the occupied territory in the military courts of the occupying power, which may be inappropriate and unjustified under long-lived military government, may be appropriate and justified under short-term military government for the purpose of restoring order.

Because the Hague and Geneva regulations were enacted with a view toward short-term military government, many of the questions that arise in the daily life of long-term military occupation are not properly addressed by these regulations. Nevertheless, the distinction between types of military rule based on the duration of occupation can serve as an appropriate policy consideration whenever such policy needs to be developed within the framework of the regulations themselves. Article 43 of the Hague Convention and Article 6 of the Fourth Geneva Convention are good examples. Public order and safety must be considered alongside the factor of time. Although the regulations were drafted with respect to short-term military government, they cannot prevent changes or developments in the scope of authority of long-term military government in operating military tribunals for civilians in the occupied territory.

The second parameter that affects the framework of the authority of the military government has to do with the nature of the military rule. This rule does not derive its authority from the choice of the residents of the region, but from the laws of war. It is a temporary rule by its very nature.  

79 See Doris A. Graher, The Development of the Law of Belligerent Occupation 1863–1914, at 113 (1949) (discussing Edgar Loening’s use of the length of occupation to assess the extent to which the occupant must take the population’s needs into account).
80 Dinstein, supra note 61, at 287.
81 See Oppenheim, supra note 62, at 177.
83 See id.
84 Greenspan, supra note 51, at 212.
nature, even if this temporariness ends up lasting a long time. It follows from the very essence of this temporariness that certain powers given to regular government are not given to military government. Therefore, military government cannot impose significant, permanent changes to legal institutions within the area it occupies, barring exceptional cases, such as when the existing institutions oppose, by their nature, elementary concepts of justice and morality.

The two parameters above—good governance over time and a rule of short duration—delineate the framework within which the authority of the military government to operate military tribunals takes place. It seems, however, that several trends affect the proper balance between these two parameters. Note that these trends are manifest in places where military or security considerations do not apply and where the only consideration taken into account is the well-being of the local population.

The life of a population does not stand still, but is in a constant state of motion that includes development, growth, and change. A military government cannot ignore this. It is not entitled to freeze life. If the military occupation lasts a long time before peace is achieved, it is the occupier’s obligation toward the civilian population to change the rules because the needs of society change over time, and the law must meet the changing needs. A military government, which was once proper in employing the use of military tribunals at the beginning of the occupation, is inherently no longer fulfilling its obligation if it continues its use of military tribunals over time.

The authority of the military government extends to all the necessary measures needed to ensure the rights of the occupied population, including the civil rights of a person in court. This

---


86 See, e.g., WEILL, supra note 82, at 21 (“As a matter of principle, Article 43 limits the power of the occupier, whose government is of temporary nature.”).

87 See Hague Convention, supra note 55, at art. 43; see also WEILL, supra note 82, at 21 (“[Article 43] imposes a general obligation on the occupying power to respect, unless absolutely prevented, the law that was in force prior to the occupation, thus preventing the occupying power from extending its own legal system over the occupied territories and ‘from acting as a sovereign legislator.’”) (emphasis added); see also, GREENSPAN, supra note 51, at 224 (discussing that it would be impractical to continue to enforce a government institution’s policy that runs contrary to protecting human rights if the reason for occupation is to liberate the people from such an institution).

88 Geneva Convention, supra note 56, art. 71–78 (detailing the rules that must be followed in the proceedings of military tribunals under military government rule in order to
approach requires that the military government cancel or reduce the activity of its military tribunals against the civilian population and direct the citizens to regular civil courts, in compliance with the first parameter above.\textsuperscript{89} This approach shows concern for the local population, as required of good governance over time, but does it not ignore the second parameter by changing the temporary character of military government, and does it not blur the distinction between military and regular government? Does the second parameter, concerning the temporary nature military government, rule out the actions required by the first?

It is difficult to provide an unequivocal answer to this question by examining the text of the Hague and Geneva conventions. The nature of military government in an occupied territory is temporary and its main function, taking into account the needs of war and security, is to act as best it can to maintain public order and security.\textsuperscript{90} Its temporary nature affects the scope of the authority of the military government in carrying out far-reaching changes to legal institutions.\textsuperscript{91} Essential long-term investments, which are likely to lead to permanent changes and have a lifetime beyond the period of military rule, may be allowed if they are necessary for the benefit of the local population, so long as they do not bring about a fundamental change in the basic institutions of the region.\textsuperscript{92}

It is reasonable to argue that reducing or abolishing the military tribunals and directing the population to civil courts fulfills the military government’s obligation to provide for the welfare of the population of the region. The very operation of the military tribunals against the population of the region for a long period of time already has a far-reaching effect on the situation in the region. Therefore, the justification of the operation of military tribunals against an occupied population must take into account the violation of the legal rights of the population, the duration of the military occupation, and the fact that the population is a foreign one and not under the sovereignty of the occupier.

The justification of military tribunals under military government is riddled with paradox. In the case of imposing martial rule, it is assumed

\begin{footnotesize}
\begin{enumerate}
\item Id. at art. 64.
\item See Hague Convention, supra note 55, at art. 43.
\item \textbf{David Kretzmer, The Occupation of Justice: The Supreme Court of Israel and the Occupied Territories 67} (2002).
\end{enumerate}
\end{footnotesize}
that the public’s interest in resolving the crisis and consolidating the legitimate regime outweighs the temporary harm to civil liberties; however, no such assumption exists where an occupying power is imposing military law on an occupied population. An occupied population is by and large not interested in the maintenance and consolidation of the occupying power. Admittedly, the use of military courts in occupied territory may be justified by the provisions of international law dealing with the authority of the military government and by the obligation of the military government to act in the interest of the local population in all areas of life. In the case of extended military government, however, it is necessary to take into account the ongoing violation of the legal rights of the occupied population and, in the case of long occupation, the occupied must be tried in civilian courts. This solution is no less problematic because, according to the provisions of international law, alongside the military government’s duty to act in the interest of the local population, there are also restrictions imposed on the military government, which is not a permanent government or sovereign, but a temporary administration. Although a transition from military courts to civil courts serves to protect the basic legal rights of the local population, it also exhibits signs of annexation of the territory in practice. Annexation is inconsistent with the provisions of international law. Thus, over time, the two goals of military courts trying civilians under military government become inherently inconsistent with one another.

IV. MILITARY TRIBUNALS AGAINST CIVILIANS IDENTIFIED AS THE ENEMY

When operating against enemy forces, military tribunals work outside the breadth of both conventional criminal and civil

93 Geneva Convention, supra note 56, at arts. 64, 66.
94 Hague Convention, supra note 55, at art. 43.
95 See id.
96 See, e.g., id. at arts. 43–56 (providing, in article 43, that the military government has a duty to act in the interest of the local population and enumerating, in articles 44–56, various restrictions on military government authority).
97 An example of a military government that has gone past the point of avoiding this paradox can be seen in the Israeli military government rule over the West Bank, which is the longest running military government in existence today. The 50-Years War, WASH. POST (Jun. 1, 2017), https://www.washingtonpost.com/graphics/opinions/israel-settlements/?utm_term=.a26c56eb1506.
proceedings.98 These tribunals usually operate as an inquisitorial arm of the military authorities; military officers prosecute and sentence the accused.99 Throughout history, the defendants were considered members of an enemy army.100 Over the years, however, several states have used military tribunals against civilians who are not members of an enemy army and, at times, even against citizens of the state itself.101

In this section, I discuss the operation of tribunals referred to as “military commissions,” which exist in the United States. Under the U.S. Constitution, military tribunals have no jurisdiction over persons who are not members of the military in some form and who are alleged to have violated the criminal or civil law.102 Nevertheless, military tribunals have been established in the U.S. to try people who are not associated with the military of any particular country, but are considered combatants who operated in violation of the laws of war.103 The Guantanamo military commissions are one such example.104

The activities of these tribunals were preceded by a Presidential declaration, an American Bar Association statement, and the U.S. Department of Defense establishment of the tribunals.105 These tribunals are considered problematic because they violate basic legal rights and have been criticized in several court rulings in the last decade. In Rumsfeld v. Padilla, a U.S. citizen was arrested and declared an enemy combatant who presents a serious threat to national security.106 The U.S. Court of Appeals for the Second Circuit ruled that the executive branch cannot imprison American citizens in military detention centers without

---

100 See, e.g., Ex Parte Quirin, 317 U.S. 1, 26–27 (1942) (holding that the president has the authority to establish military courts for enemy trials, and specifically, that the members of the German armed forces were subject to U.S. military trial).
102 See Ex Parte Milligan, 71 U.S. 2, 128 (1866) (“Marital rule can never exist where the courts are open, and in the proper and unobstructed exercise of their jurisdiction.”).
the approval of Congress.\textsuperscript{107} After a series of appeals, the Second Circuit ruled that criminal conspiracy charges must be filed against the accused in a civilian court, not in a military tribunal.\textsuperscript{108} In \textit{Rasul v. Bush}, the Supreme Court ruled that the accused had habeas corpus rights, despite the fact that the detainee was not a U.S. citizen, because he was not a citizen of a country that was at war with the U.S.\textsuperscript{109} In \textit{Hamdi v. Rumsfeld}, a U.S. citizen was arrested as a Taliban combatant on the battlefield in Afghanistan and, as a result, was declared an unlawful enemy combatant during a national emergency.\textsuperscript{110} The case reached the Supreme Court where the majority opinion challenged the President’s authority to deny a detainee’s basic legal rights and individual civil liberties, such as habeas corpus and due process.\textsuperscript{111} Nevertheless, five members of the Court did not deny the jurisdiction of military tribunals to try civilians and agreed that U.S. citizens can be held as enemy combatants.\textsuperscript{112} In \textit{Hamdan v. Rumsfeld}, the Supreme Court ruled that President Bush’s attempt to establish military tribunals to judge civilians was a violation of the United States Code of Military Justice (“UCMJ”) and the Fourth Geneva Convention of 1949.\textsuperscript{113} The Court disqualified the military tribunals, asking Congress and the President to reconsider their authority.\textsuperscript{114}

Following the Supreme Court ruling, Congress passed the Military Commissions Act of 2006 (“2006 MCA”), which has affected the nature of the U.S. military tribunals.\textsuperscript{115} The stated purpose of the act was “[t]o authorize trial by military commission for violations of the law of war, and for other purposes.”\textsuperscript{116}

The main effect of the law was to grant jurisdiction to military

\begin{itemize}
  \item Padilla v. Rumsfeld, 352 F.3d 695, 698 (2d Cir. 2003), rev’d on other grounds, 542 U.S. 426 (2004).
  \item Id. at 724.
  \item Id. at 536–37; see Jenny S. Martinez, \textit{Hamdi v. Rumsfeld}, 124 S.Ct. 2633, United States Supreme Court, 98 Am. J. Int’l L. 782, 783–84 (2004).
  \item Hamdi, 542 U.S. at 519.
  \item Hamdan v. Rumsfeld, 548 U.S. 557, 560 (2006) (“The military commission at issue lacks the power to proceed because its structure and procedures violate both the UCMJ and the four Geneva Conventions signed in 1949.”).
  \item Id. at 635.
\end{itemize}
tribunals to arrest and try those declared as unlawful enemy combatants. The law extended the jurisdiction of the military commissions established to try cases of violations of the law or of the laws of war. The act conferred the jurisdiction of military commissions “solely to aliens who have engaged in hostilities against the United States or who have purposefully and materially supported hostilities against us.”

The law sparked controversy because it authorized the President to determine which individuals to designate as “unlawful enemy combatants,” making it possible to bring them under the jurisdiction of the military commissions and deny them habeas corpus rights. Indeed, the law was harshly criticized when, in *Boumediene v. Bush*, the Supreme Court determined that foreign detainees held by the U.S., including those in the detention camp at Guantanamo Bay, have habeas corpus rights according to the U.S. Constitution, and the 2006 MCA had illegally suspended that right.

The sharp criticisms against the military commissions charged that the commissions violated individual rights, violated the right to due process, and breached sections of the Fourth Geneva Convention. Despite these charges, the amendment to the law under the Obama administration, the Military Commissions Act of 2009 (“2009 MCA”), did not make decisive changes, leaving military commissions in place. Congress granted jurisdiction to the military commissions to try civilians on 32 counts, including pillaging, taking hostages, torture, mutilation, rape, conspiracy, and providing material support for terrorism.

In the interval between the legislative changes, Hamdan appealed his conviction in the United States Court of Military Commission

Hamdan claimed, *inter alia*, that the military commission had tried him without authority and that the charge—financial support for terrorism—was not a violation of international law under the laws of war. He also argued that the 2006 MCA violated the Constitution by bringing foreign citizens to trial in a military tribunal. The court held that “Congress exercised authority derived from the Constitution to define and punish offenses against the law of nations by codifying an existing law of war violation into a clear and comprehensively defined offense of providing material support to terrorism.” The court reasoned that crimes equivalent to the offense of providing material support for terrorism have long been tried by military commissions. The court concluded that “Congress had a rational basis for the disparate treatment,” and that such treatment does not violate the Equal Protection Clause of the Constitution.

The 2009 MCA somewhat softened the severe violation of individual rights, and authorized the Secretary of Defense to enact easements to allow the defendants to be present at all hearings, to examine all evidence brought against them, to bring witnesses to testify on their behalf, and to cross-examine government witnesses. Currently, in the U.S., judges of special military tribunals are required to have the same qualifications as judges in courts martial. Military tribunals automatically assign military advisors, who meet the standards of advocacy in regular military court proceeding, to defendants. Defendants can also choose to be represented by a civilian attorney. The jury is composed of officers on active duty. In many ways, the legal procedure in military tribunals is identical to that in civilian courts, and, in the case of a conviction, there are three instances of appeal.

Note, however, that many of the problems afflicting military
commissions persist from the perspective of the evolving approach of the international community. For example, the appropriate use of military commissions to try Guantanamo prisoners accused of committing war crimes is still the subject of controversy and confusion. Under a military commission’s rules of evidence, detainees are unable to meaningfully contest the factual basis of their detention. There are no legal standards for assigning individual detainees to criminal prosecution, military commissions, or indefinite detention without charge.

There is now less emphasis on U.S. citizenship to receive protection under the Constitution and the courts. Some are even concerned with the abandonment of the rule of law. In an address on the future of military commissions on May 23, 2013, President Obama pledged to close down Guantanamo Bay, but at the same time defended the “military justice system” as a place to “bring terrorists to justice” in the war on terrorism. Unexpectedly, the Obama Administration decided to bring new cases, rather than shutting down the military commissions.

An examination of the U.S. military tribunals in the past fifteen years has provided invaluable understanding into how states that espouse human rights justify the jurisdiction of military tribunals. The


144 Id.

145 See, e.g., Justin Walker, The Execution of the Innocent in Military Tribunals: Problems from the Past and Solutions for the Future, 119 W. VA. L. REV. 1, 45 (2016) (suggesting the importance of understanding the history of military tribunals, and in particular, the imposition of U.S. military commissions over the past fifteen years, as a means of understanding how best to proceed with military tribunals in the future).
jurisdiction of these courts is based on the legitimacy of regimes that protect CPP rights to defend themselves against terrorist attacks.\footnote{William Michael Jr., Military Tribunals are an Appropriate Procedure, 59 BENCH & B. MINN. 20, 22 (2002) (stating that the jurisdiction of military courts to punish, \emph{inter alia}, enemy belligerents of war has been long established).} The assumption is that protecting public security justifies the operation of military tribunals against suspected supporters of terrorism as long as there is a danger to public safety.\footnote{See Byard Q. Clemmons, The Case for Military Tribunals, FED. LAW., May 2012, at 27 (advocating for the use of military tribunals to try suspected terrorists).} According to the U.S. experience, it is difficult to pinpoint the “duration” of such a period and it is difficult to know who meets the definition of an unlawful enemy.\footnote{See generally, George C. Harris, Terrorism, War, and Justice: The Concept of the unlawful Enemy Combatant, 26 LOY. L.A. INT’L & COMP. L. REV. 31 (2003) (comparing U.S. cases of \emph{Hamdi} and \emph{Padilla}, along with historical precedent, and discussing that the concept of who can be classified as an enemy combatant comes down to evidence or lack thereof).} When the main argument for operating these tribunals is necessity,\footnote{See, \emph{e.g.}, Williams & Benjamin, supra note 119, at 609 (arguing that military commissions are a product of necessity).} it is important to distinguish between actual and declared necessities. Regimes often tend to forget this, and worse, tend to use necessity to fabricate simulated crises.\footnote{See, \emph{e.g.}, Eric. K. Yamamoto, \textit{White (House) Lies: Why the Public Must Compel the Courts to Hold the President Accountable for National SecurityAbuses}, 68 L. & CONTEMP. PROBS. 285, 287–88 (2005) (arguing that the U.S. Supreme Court irrationally sided with the Justice and War Department’s argument of “necessity” to intern Japanese Americans during World War II, even though the military knew that there was no necessity for incarcerating a single race of citizens).} In recent years, other countries have found themselves in a situation similar to that of the U.S. and have needed to establish military tribunals for civilians.\footnote{See, \emph{e.g.}, Eric. K. Yamamoto, \textit{White (House) Lies: Why the Public Must Compel the Courts to Hold the President Accountable for National SecurityAbuses}, 68 L. & CONTEMP. PROBS. 285, 287–88 (2005) (arguing that the U.S. Supreme Court irrationally sided with the Justice and War Department’s argument of “necessity” to intern Japanese Americans during World War II, even though the military knew that there was no necessity for incarcerating a single race of citizens).} These countries are debating similar questions: Are we at war? Is it appropriate to try enemy combatants, away from the battlefield, in a military tribunal rather than a civilian court? And if so, what would be sufficient for constitutional due process? How can we ensure that military tribunals protect the civil rights of the defendant while protecting our national security?

Countries that wish to grant legitimacy and authority to military tribunals to try civilians might claim that they are at war in a time when wars are not always fought on a battlefield and often involve clandestine
and covert operations. At the same time, it is difficult to declare war these days because many governments find it difficult to define the entities that they are fighting. In the current climate of international terrorism that endangers many citizens worldwide, it may be impossible to grant enemy combatants all the rights provided by civil courts. But most countries that protect CPP rights do not want to sacrifice these rights, which have become part of their essence.

The paradox of trying civilians away from the battlefield in military courts is inescapable. In the absence of occupation or of an internal state of emergency, there is no justification for trying civilians in military tribunals because the civil courts have shown themselves quite capable of dealing with the most heinous crimes. The only justification for such action is to relax the requirements of due process and of habeas corpus, which the U.S. Supreme Court has clearly rejected. Even if such actions are justified in a war-like emergency, they cannot be upheld in a protracted arrangement in the absence of a crisis. If, however, legislation or precedent reintroduces due process and habeas corpus into the military tribunals, even through the back door, the very justification for these tribunals disappears.

V. SUMMARY

I have distinguished between three types of military tribunals that try civilian populations. Common to all three types is that in all cases there is violation of human rights (such as the right to due process, habeas corpus, etc.). There are differences, however, between the populations being tried by these tribunals and the grounds for the jurisdiction of the tribunals, especially once the element of time becomes a factor.

Martial law imposed by a country on its sovereign territory and its civilian population in times of crisis is expected to be lifted once the

---


155 See U.S. Const. art. I, § 9, cl. 2.
crisis is resolved and citizens are to resume standing trial in civilian courts. It is assumed that the long-term interest of the citizens in the resolution of the crisis overrides the temporary violation of their liberty during the period when military tribunals are being operated for a limited time. When all citizens of the state suffer the violation of their civil rights, it is possible to justify the broad powers of military tribunals over time, provided that the crisis situation persists and requires the continued imposition of martial law, but not beyond that.

The case of a military government operating military tribunals in an occupied territory is more complex. The occupied population is not interested in maintaining the regime of the occupier. The justification of military courts under military occupation, based on international law, has led to the paradox between the obligation of the military government to provide good governance that cares for the local population in all areas of life and the limitations that the military government faces for being a temporary administration rather than a permanent sovereign. The former requires transitioning from military tribunals to civilian courts as a function of the duration of the military government, whereas the latter warns against the annexation of the occupied territory.

Finally, throughout history, enemy combatants were considered prisoners of war and tried if they committed war crimes. Recently, new military commissions have been established to try, among others, nationals of countries that are not enemies of the states or citizens of their own countries who pose a threat because of their support of terrorist activities. The only justification, however, for trying these individuals in military tribunals is to deny some of their civil rights, which the U.S. Supreme Court has ruled against, forcing the tribunals to reinstate these rights to a considerable degree, and thereby making void the reason for their existence in the first place.

In all cases, military tribunals trying civilians appear to be destroying to some degree the rights they set out to protect, undermining to the same degree their own legitimacy. This is especially apparent when the use of military tribunals for trying civilians becomes a practice that can no longer be considered temporary. It is not entirely clear at which point in time this transition occurs. It may not be possible to indicate that at time $P$ the jurisdiction of military tribunals to try civilians is justified, but at time $Q$, sometime later, it no longer is. But it is possible to point to the beginning of the process, at time $A$, when it is justified, and to a much later time $Y$, when it clearly is no longer justified.
A SOUND TAXONOMY OF REMEDIES

Tim Kaye*

I. INTRODUCTION .................................................................................... 80
   A. The Identification Problem .............................................................. 81
   B. A Lack of Taxonomy ........................................................................ 83
   C. The Menu ......................................................................................... 85

II. THE BASIS OF A TAXONOMY OF REMEDIES ...................................... 87
   A. Remedy-by-Remedy or Wrong-by-Wrong? ..................................... 87
   B. Wrongs and “Not-Wrongs” .............................................................. 88
   C. The Taxonomical Task ..................................................................... 90
   D. Historical Origins ............................................................................ 91
   E. The Problems of Classifying by Goals .......................................... 93
      1. Overlapping Categories ............................................................... 93
      2. Omission of Remedies ................................................................. 94
      3. Identical Remedies for Different Goals ....................................... 96
      4. Different Remedies for Identical Goals ....................................... 98

III. A RIGHTS-BASED TAXONOMY .......................................................... 99
   A. John Austin .................................................................................... 99
   B. Power-Conferring Rules .................................................................. 100
   C. Primary Rights .............................................................................. 101
   D. Interests .......................................................................................... 103
   E. Secondary Rights ........................................................................... 104
   F. Comparing Primary and Secondary Rights .................................... 108
   G. Tertiary Rights ............................................................................... 111

IV. REPLICATIVE REMEDIES ................................................................. 112
   A. Property Law .................................................................................. 114
      1. Replevin ...................................................................................... 114
      2. Detinue ......................................................................................... 116
      3. Ejectment (or Quiet Title) ............................................................. 117
      4. Self-Help Remedies ..................................................................... 119

* I should like to thank Ann Piccard for her comments and suggestions on an earlier draft. I should also like to thank Stetson University College of Law for a grant that facilitated the research and writing of this paper.
I. INTRODUCTION

In 1988, in the case of Bowen v. Massachusetts, the United States Supreme Court faced what initially appeared to be the utterly banal task of explaining the meaning of the remedy of “money damages.”¹ Yet the Court split 6–3 in deciding the issue—and that split was not along ideological lines.²

The majority concluded that the State of Massachusetts was entitled
to receive reimbursement from the federal government for expenditures incurred under Medicaid.\footnote{See Bowen, 487 U.S. at 909–10.} The Court reached this conclusion because it correctly held that Massachusetts was not making a claim for “money damages,” which would have been barred by section 702 of the Administrative Procedure Act (as amended in 1976).\footnote{Id. at 910.} On the basis that “money damages” are a common law remedy, however, the majority then went on to conclude—erroneously—that the remedy sought by Massachusetts must be equitable.

The dissent, written by Justice Scalia, correctly saw through that conclusion.\footnote{Id. at 917 (Scalia, J., dissenting).} If the remedy sought had indeed been equitable, it would mean that the U.S. Claims Court would be “out of business,” because it has jurisdiction only over common law claims.\footnote{Id. at 919.} The U.S. Claims Court would no longer, for example, have jurisdiction over a seller’s action for the price, despite the fact that it routinely deals with such claims.\footnote{Bowen, 487 U.S. at 920–21 (Scalia, J., dissenting).} Yet the dissent also drew an erroneous conclusion; namely that, because such a ramification was obviously wrong, the remedy sought by the State of Massachusetts must be “money damages” at common law and, therefore, not recoverable.\footnote{Id. at 927.}

Thankfully, Massachusetts got its money,\footnote{Id. at 909.} which was the correct result. But it is alarming—to say the least—that none of the nine Justices on the Supreme Court correctly identified the nature of the remedy that the State sought. It is even more alarming because the remedy in question was nothing more complex than debt.\footnote{Id. at 882–84, 884 n.2.}

\textit{A. The Identification Problem}

Unfortunately, understanding of the nature of remedies and—even more importantly—of the implications of different remedies, has deteriorated even further since Bowen. Cases alleging patent infringement, for example, are rife with such errors, with plaintiffs routinely claiming the non-existent remedy of “an accounting for
damages."\textsuperscript{11} This appears to be a confusion between the remedies of damages and an accounting for profits.\textsuperscript{12} Yet the appropriate remedy is actually neither of the two. It is, in fact, an account for money had and received.\textsuperscript{13}

This inability to classify a remedy correctly is akin to a botanist being unable to categorize a specific plant. If the plant is exotic or very rare, that might be understandable. But, even in such cases, the botanist’s training in taxonomy would stand him or her in very good stead. When faced with a plant that she does not recognize, the competent botanist understands that she must check for certain characteristics so as to locate the plant within an appropriate classification. This means that, as long as an exhaustive list of plant taxonomies has already been developed, it will be possible to categorize the specific plant under investigation. This will, in turn, provide the basis for narrowing down the range of potential identities until the correct one is established (or it is established that the plant is a species hitherto unknown).

Remedies to recover debt, however, are neither rare nor exotic. On the contrary, they are probably the most common of all remedies addressed by the courts.\textsuperscript{14} So, if the nine Justices of the highest court in the nation cannot recognize or appropriately categorize a debt, something fundamental is clearly awry.

To be clear, I am \textit{not} suggesting that the Justices in \textit{Bowen} would have been unable to recognize a simple contractual debt. What I \textit{am} saying is that they were apparently unable to recognize a debt outside of its regular, everyday context. But this is when the lawyer’s ability to identify the correct remedy matters most. We scarcely need someone with a law degree and many years’ legal practice or judicial experience to tell us that Joe owes Jim $5,000 when, in the knowledge that it has not been paid, we can simply look at a document spelling out that Joe was meant to pay Jim precisely that amount last year. What we \textit{do} need is for

\begin{enumerate}
\item See, e.g., Mark A. Lemley, \textit{The Ongoing Confusion over Ongoing Royalties}, 76 MO. L. REV. 695, 698–99 (2011) (discussing how courts have power in equity to order accountings, which “[s]trictly speaking, . . . are not damages”).
\item See \textit{infra} Part IV(C)(2).
\end{enumerate}
lawyers searching for a remedy in unfamiliar circumstances to know how to approach such an inquiry. In short, the law needs an equivalent of the botanist’s method.

In fact, lawyers should have a major advantage over botanists. There are hundreds of thousands of plant species, and we have not yet identified them all. Lawyers probably have to deal with fewer than a hundred different types of remedies, and—at least in principle—they should all already have been identified. And yet we evidently struggle to identify both what they are and when they apply. Professor David Partlett apparently believes that “[t]he common law adheres cohesively through its disciplined practice of deliberate, practical reasoning,” and so has no need of a thorough taxonomy. Cases like *Bowen v. Massachusetts*, however, demonstrate that this is a hopelessly optimistic and misguided view. All reasoning needs to be anchored somewhere. Judges and practitioners need recognized categorizations to act as the premises from which reasoned arguments develop. Otherwise, as *Bowen* shows, we will all simply be talking past one another.

**B. A Lack of Taxonomy**

The lack of a meaningful taxonomy of remedies is strange for many reasons. After all, taxonomy has been called “the world’s oldest profession” because humans are probably hard-wired to think by means of classifying things. Even the most rudimentary understanding of substantive law demands a basic grasp of fundamental taxonomies. Criminal law makes no sense if we cannot distinguish a felony from a misdemeanor, or crimes that require proof of mens rea from those that

---

15 “[I]t is estimated that the total number of plants is of the order of 400,000 species.” Plant Species Numbers, BOTANIC GARDENS CONSERVATION INT’L, https://www.bgci.org/policy/1521/ (last visited Oct. 14, 2017).


17 See id. at 2060.

18 See Joel Walker Hedgpeth, Taxonomy: Man’s Oldest Profession, Eleventh Annual University of the Pacific Faculty Research Lecture (May 22, 1961). There is, of course, another occupation that vies for this title.

19 See Emily Sherwin, Legal Taxonomy, 15 LEGAL THEORY 25, 25–26 (2009) (stating that “[t]o some extent, taxonomy inevitably plays a role in legal analysis: to think intelligently about law, one must sort legal rules and decisions into categories and generalize about fields of law.”). See also Peter Birk, Rights, Wrongs, and Remedies, 20 OX. J. LEG. STUD. 1, 3 (2000) (stating that “[t]he law simply could not be understood unless it took care to classify itself ‘methodically’. ”).
do not. Similarly, no one can be considered even slightly competent in the law of torts without understanding the difference between intentional, negligent, and strict liability torts. Yet, not only do we currently lack a meaningful taxonomy of remedies, but some commentators have also even argued that attempting to create one would be futile. According to this view, the law on remedies is simply so unsystematic that each remedy must essentially be viewed on its own terms. The late Peter Birks had a perfect response:

No department of human knowledge ever advanced without attention to taxonomy. A legal system which insists on working in a taxonomic vacuum... will not be able to treat like cases alike and will make many decisions which, on reflection, will be rationally indefensible. It will tumble from the highest branches of ethics to the lowest levels of gut reaction.

Refusing to develop a meaningful taxonomy of remedies leaves us with nothing better than an arbitrary list, like random items on the menu of an eclectic restaurant. At least such a restaurant would typically ensure that the names of each dish are accompanied with a brief description to reduce the chance of a diner making an ill-advised choice. Unfortunately, the law’s list of remedies does not come with an equivalent explanation. As a result, courts’ decisions on the remedy to award are often haphazard, reached through a combination of politics and chance.

This has, in fact, been a recurring pattern throughout the history of the common law. As Sir Thomas Holland once put it: “the old-fashioned English lawyer’s idea of a satisfactory body of law was a chaos with a full index.” What Partlett apparently celebrates as the common law’s “practice of deliberate, practical reasoning,” led in the past to almost total confusion in the laws of contracts and torts, from which the law had

20 See, e.g., Sherwin, supra note 19, at 27 (criticizing a taxonomical approach as “formalistic and pointless”).
22 Birks, supra note 19, at 37.
to be rescued by the importation from civil law countries of taxonomies that we now treat as commonplace and axiomatic. Contract law was rescued in the eighteenth century by the wholesale importation of concepts and categories from French law, as interpreted by Robert Pothier, while tort law was rescued a century later when Oliver Wendell Holmes imported the taxonomy of intentional, negligent, and strict liability torts from the German law of unlawful acts (unerlaubte Handlungen).

The law of remedies is in need of similar rescue. It is a field in which judges and practitioners currently struggle to see the wood for the trees. Moreover, the fact that some commentators believe that the law of remedies is simply too haphazard and unsystematic to be susceptible to taxonomy is actually a clear sign that a meaningful taxonomy is long overdue. It certainly does not mean that such a taxonomy cannot be created. Only two years before he himself announced a new taxonomy of torts, Holmes declared that, because it was such a shambles, “Torts is not a proper subject for a law book.” Clearly, he took such a problem as a challenge to be overcome. The taxonomy of remedies presented here proceeds on a similar footing.

C. The Menu

The rest of this article proceeds in the following manner. Part II considers three possible foundations on which a taxonomy of remedies might be based. Thus remedies might be classified according to:

---


27 “The whole structure of the common law of contracts and sales is based largely on Pothier’s treatises on obligations and sales.” Perillo, supra note 26, at 267 (footnote omitted).


29 See, e.g., Michael Tilbury, Remedies and the Classification of Obligations, in THE LAW OF OBLIGATIONS: CONNECTIONS AND BOUNDARIES 30 (Andrew Robertson et al. eds., 2004) (stating that a taxonomy of remedies has “its limitations” and that because “[l]aw is solely the product of the human mind . . . [i]t is subject to constant revision, reinterpretation and reinvention”).


31 See Grey, supra note 26, at 144–50 (discussing Holmes’s contributions in creating a taxonomy of tort law).
(1) The type of wrong necessitating the remedy;
(2) The various goals that each remedy is designed to achieve; or
(3) The differing rights that each remedy is designed to vindicate.

It then explains why any attempt to construct a taxonomy of remedies based on either (1) or (2) is bound to fail. A taxonomy based on wrongs fails because many remedies are not predicated on a wrong at all. A taxonomy based on goals fails because the goals are not mutually exclusive, and such exclusivity is a fundamental requirement of any taxonomy. Goal-based taxonomies also fail because they cannot explain why the law, in practice, provides identical remedies in the apparent pursuit of different goals, as well as different remedies in the apparent pursuit of identical goals. In sum, theories based on these grounds simply do not match the law. The only sound basis for a taxonomy of remedies is, therefore, a taxonomy of rights.

Part III explores this taxonomy of rights. It begins with John Austin, who attempted such a taxonomy around a century ago, but whose work has largely been ignored since significant flaws were famously exposed by H.L.A. Hart in the middle of the twentieth century.32 But, properly understood, Hart’s work does not so much undermine Austin’s work as enhance it. Drawing on more recent work by Peter Birks33 and Birks’s former doctoral student, Rafal Zakrzewski,34 but modifying it significantly in order to eradicate inaccuracies and unhelpful metaphysics, I develop a taxonomy of primary, secondary, and tertiary rights.

Parts IV, V, and VI then develop a complementary taxonomy of replicative, substitutionary, and transformative remedies. These are not mere theoretical constructs. On the contrary, these parts draw heavily on decided cases both to illustrate the defining characteristics of the different types of remedies and to show how they work in practice. Part IV shows how replicative remedies vindicate primary rights. They do not require proof of either harm or wrongdoing, and may be asserted against the whole world. Eleven different remedies, from four different fields of law, are explored in detail to show how the theory of replicative remedies works—and is easily applied—in practice. Part V takes a
similar approach to substitutionary remedies. These exist to vindicate secondary rights, but, unlike replicative remedies, they become available only if a legally-protected interest has been lost or damaged because of someone else’s wrongdoing. Substitutionary remedies may thus be asserted only against wrongdoers and those vicariously responsible, while the measure of the remedy awarded is defined by the loss or harm caused. Part VI then briefly discusses transformative remedies, such as divorce, bankruptcy protection, naturalization, and political asylum. These are not based on prior rights at all, but seek instead to create an entirely new status for the party concerned.

Part VII returns to the case of Bowen v. Massachusetts to show how a proper understanding of the taxonomy of remedies can turn an apparently difficult case into one that is really quite straightforward. The paper concludes with Part VIII, where I argue that it is time for the taxonomy to be both formally adopted by the courts and taught within law schools as part of the remedies curriculum.

II. THE BASIS OF A TAXONOMY OF REMEDIES

A. Remedy-by-Remedy or Wrong-by-Wrong?

It has been said that a course on remedies must follow one of two patterns: it must proceed either “remedy-by-remedy or wrong-by-wrong.”

35 This assertion implies that we have a choice of classifying remedies according to the type of wrong which they are intended to put right, or according to particular characteristics of the remedies themselves. In other words, any meaningful taxonomy of the law of remedies must apparently rest on the basis of one or the other.

Unfortunately, however, this is a false dilemma. Organizing a taxonomy of remedies “wrong-by-wrong” involves a fundamental error. Despite Douglas Laycock’s bald assertion that a “remedy is anything a court can do for a litigant who has been wronged or is about to be wronged,”

36 many remedies are not, in fact, predicated on a wrong at all.

To be fair to Laycock, he was simply repeating an error first made by the English legal philosopher, John Austin, over a century ago.

37

37 1 JOHN AUSTIN, LECTURES ON JURISPRUDENCE, OR THE PHILOSOPHY OF POSITIVE
Austin did, however, notice one problem with his assertion, involving the case of a person who mistakenly receives money. There is no doubt that the recipient then has a duty to repay that money if asked for its return within a reasonable period of time—but on what basis? As Austin himself recognized, “the possessor is not guilty of a wrong,” so how can the rightful owner of the money claim a right to its return?

Austin’s own answer was that the person in mistaken possession of the money is “under a quasi-contract to restore” it. But this explains nothing. How did this supposed quasi-contract come into existence? (If it means nothing more than that the law deems it to be a contractual obligation, then it means nothing at all.) Yet Austin did not flinch. Apparently oblivious to the fact that he had failed to identify any wrong, which would provide the legal justification for recovery of the overpayment, he simply affirmed his belief that “every right of action arises from a wrong.” Clearly, this will not do.

B. Wrongs and “Not-Wrongs”

The mistaken receipt of money is far from the only circumstance in which someone may be entitled to a remedy even though no wrong has been committed. Other examples include rescission of a contract for innocent misrepresentation, reformation of a contract because of a mutual mistake, and a government’s pursuit of the collection of sales or income tax. In none of these cases has the person subject to the remedy done anything wrong. Indeed, in the case of sales tax, she must have made a sale, which would generally count as a good thing. And yet the more sales she makes, the more tax the state is entitled to collect.

But how can the law justify the availability of a remedy when there has been no wrongdoing? To address this apparent conundrum, it is important to realize that the common law did not originally speak of “remedies” at all. Instead, a plaintiff simply sought a court order, known

38 Id.
39 Id.
40 Id.
41 AUSTIN, supra note 37, at 766.
43 See, e.g., CAL. CIV. CODE § 3399.
44 See Birks, supra note 19, at 28 (stating “[t]axable events are not wrongs.”).
45 See id.
(because it was in writing) as a “writ.” 46 The language of “remedies” actually originates from the medical field. 47 While it is true that a medical remedy is sought when something is “wrong,” such as in the case of illness, disease, or trauma, these ailments are often suffered without any predicate human agency, let alone human wrongdoing. When a person catches the flu, for example, it does not mean that it is another’s fault.

When imported into the discourse of the law, however, the word “remedy” somehow came to bear the connotation that some sort of prior human misdeed must have occurred. 48 But that connotation is as false in the legal context as it is in the medical. 49 While a legal remedy may be obtained where something has, to put it colloquially, “gone wrong,” this does not necessarily mean that anyone has “done wrong.” 50 Contrary to both Austin’s assertion “that every right of action arises from a wrong” 51 and Laycock’s assertion that a “remedy is anything a court can do for a litigant who has been wronged or is about to be wronged,” 52 remedies sometimes vindicate rights without any predicate wrongdoing at all. 53

Peter Birks coined the term “not-wrongs” to refer to the many instances where no wrong has been committed but a party is nevertheless entitled to a remedy. 54 He gave an example of a not-wrong in the following terms:

There is no conception of wrong which will reach the receipt of a mistaken payment. . . . A shop gives me change as though for a £20 note. I had paid actually with a ten. I stuff the change in my pocket, unaware of anything amiss. As I walk out the shop assistant realizes and calls out. There is no . . . wrong at the moment that knowledge of the mistake supervenes. The law is, and rightly, that I was unjustly enriched when I received the extra change. The

47 See Birks, supra note 19, at 9 & n.32.
48 According to the theory of legal autopoiesis, such a change in meaning is inevitable when a concept is taken from one field and applied in another. See Neil Lyons, Autopoiesis: Evolution, Assimilation, and Causation of Normative Closure, in LAW, JUSTICE, AND MISCOMMUNICATIONS: ESSAYS IN APPLIED LEGAL PHILOSOPHY 80–83 (Tim Kaye ed., 2011).
49 See Birks, supra note 19, at 25.
50 Id.
51 AUSTIN, supra note 37, at 766.
52 LAYCOCK, supra note 36, at 1.
53 See Birks, supra note 19, at 25.
54 Id.
shop’s right to restitution was born from that unequivocal not-wrong.55

No one has committed any breach of the law when the shop assistant makes too much change and the purchaser, without checking, puts it in his or her pocket. The shop, however, is still entitled to a remedy because the purchaser has been unjustly enriched. Indeed, Birks observed the existence of three categories of events that create not-wrongs and which, therefore, trigger the availability of a legal remedy:

(1) manifestations of consent, such as making a contract (providing the basis for an order of specific performance or for the collection of the price charged),56

(2) unjust enrichment;57 and

(3) miscellaneous other events, including a salvor’s right to a reward58 a government’s ability to demand the payment of tax,59 and certain proprietary rights.60

Sensibly, he recognized in the third category of not-wrongs the existence of many other occasions where a legal remedy may be available as a consequence of a not-wrong.61 No-fault divorce laws, for example, may be added to Birks’s list.62 In every such instance, the law makes a remedy available, even in the absence of any act or omission that the law considers wrongful.63

C. The Taxonomical Task

Because many remedies are not predicated on prior wrongdoing, it is clear that we cannot proceed to generate a meaningful taxonomy of remedies by proceeding “wrong-by-wrong.” Such an approach would inevitably omit a significant number of remedies. Yet we cannot proceed “remedy-by-remedy” too literally, for that would not generate a

55 Id. at 28.
56 Id. at 27.
57 Birks, supra note 19, at 27–28.
58 Id. at 28.
59 Id.
60 Id. at 28–29.
61 Birks, supra note 19, at 28 (“The third category of not-wrong is difficult to handle. Nobody has been brave enough to enumerate its members.”).
62 See, e.g., Katherine M. Staley, No-Fault Divorce, 59 WOMEN LAW. J. 54, 57 (1973) (acknowledging that no-fault divorce changes the basic conception of bringing suit as an injured party seeking to recover from a wrongdoer).
63 See, e.g., id. at 54 (commenting that marriages frequently cease even though neither partner is guilty of a matrimonial offense).
taxonomy so much as a simple list. We need instead—as with botanical taxonomies—to engage in the following tasks:

1. Identify the specific characteristics of each remedy; and
2. Establish which characteristics are mutually exclusive—no taxonomy can function if something can be placed in two different categories.64

D. Historical Origins

Perhaps the best-known way to classify every remedy according to mutually-exclusive criteria involves identifying whether it originates from common law or from equity.65 For many years, this was indeed the most fundamental criterion for classification, since it determined whether a court had jurisdiction to hear a case. Even today, it remains true that statutory remedies must typically be classified as having either common law or equitable antecedents.66 In Curtis v. Loether, for example, the Supreme Court held that a jury trial was required for an action under Title VIII of the Civil Rights Act of 1968 because it “sounds basically in tort—the statute merely defines a new legal duty, and authorizes the courts to compensate a plaintiff for the injury caused by the defendant’s wrongful breach.”67

The historical origins of a remedy, including a statutory remedy, continue to affect both the nature of the trial and the circumstances in which the remedy in question may be ordered.68 Nevertheless, the almost universal69 merger of these jurisdictions70 has drastically reduced

---

64 See, e.g., Sherwin, supra note 19, at 33–34 (stating that, in structuring a formal legal taxonomy, “legal categories must not overlap”).

65 See, e.g., Samuel L. Bray, The System of Equitable Remedies, 63 UCLA L. REV. 530, 541 (2016) (noting that state and federal courts continue to classify remedies as either legal or equitable).

66 Id. at 544.


68 See id. (discussing legal and equitable rights and the implications for the right to a trial by jury, as guaranteed by the Seventh Amendment).


70 Following the English law implementation of the Judicature Act of 1873 merging the jurisdictions, the Federal Rules of Civil Procedure of 1938 instituted such a merged system of pleading law and equity claims in the U.S. See W. Hamilton Bryson, The Merger of Common-Law and Equity Pleading in Virginia, 41 U. RICH. L. REV. 77, 77–78 (2006) (“The merged system of pleading has always been used in Texas. It was adopted in New York by the Field Code of 1848. It has been used in England and Wales since the Judicature Act of 1873, in the
the significance of this distinction. Moreover, cases like *Bowen v. Massachusetts* demonstrate that treating the common law-equity distinction as the primary basis for a taxonomy of remedies is apt to cause more confusion than it resolves. This is because such an approach encourages judges to assume that the determination of whether a remedy is available turns simply on whether it originated in the common law or in equity. In reality, however, the issue at stake is much more likely to be determined not according to which side of the common law-equity division the remedy falls, but rather where within each such category it lies.

This is not to say that the historical origins of a remedy play no role in the law today. It is clear that they do, primarily in placing riders or restrictions on the availability of certain remedies. But riders and restrictions can come into play only at a second or third order of significance. The function of remedies, therefore, needs to be identified first.

Unfortunately, functionality can be perceived in two very different ways. On the one hand, it can be defined teleologically, meaning that the law uses remedies as a means of striving for certain policy goals or objectives in specific cases. On the other, functionality can be defined deontologically, to give effect to specific rights (irrespective of where that might lead in a particular case). It is obviously important to resolve this ambiguity.

dependent courts of the United States since the Federal Rules of Civil Procedure of 1938, and now in Virginia as of 2006."


See Nora J. Pasman-Green & Alexis Derrossett, *Twenty Years After Bowen v. Massachusetts—Damages or Restitution: When does It Still Matter? When Should It?*, 69 LA. L. REV. 749, 749–55 (2009) ("Both the majority and the minority positions in *Bowen* equated the characterization of the claim as specific monetary relief with being an equitable claim. Both opinions ignored a more apt characterization of the reimbursement sought as a legal claim for restitution that is neither specific relief nor equitable.").

See id. at 763 (noting that the legacy of *Bowen* has resulted in courts struggling to classify the remedy sought as being either compensatory damages or restitution).

See Bray, supra note 65, at 544–50 (highlighting some of the consequences of classifying a remedy as legal or equitable).

Laycock stipulates that remedies can be classified by function or form. See LAYCOCK, supra note 36, at 3.


*Id.* at 21–24.
E. The Problems of Classifying by Goals

1. Overlapping Categories

The idea of classifying by goals has been, by far, the more popular approach taken over the last fifty years or so.\(^{78}\) Unfortunately, as we shall now go on to see, it simply cannot measure up to the challenge. It is, in fact, one of the reasons that we currently lack a meaningful taxonomy of remedies.

My colleague, Marco Jimenez, has argued that we should classify remedies according to whether they are restorative, retributive, coercive, or protective.\(^{79}\) Unfortunately, as Rafal Zakrzewski has previously explained, such a classification is simply not up to the task.\(^{80}\) The problem is that:

\[\text{[S]pecific enforcement, coercion, or declaration cannot be said to be a purpose, objective, or goal of a right in the same way that compensation, restitution, or punishment are.} \ldots \text{[they] do not represent goals of a similar nature.}\]

\[\ldots \text{The categories of compensation, restitution, and punishment classify remedies} \ldots \text{according to goal. The categories of specific enforcement, coercion, and declaration classify remedies according to features they have when viewed as} \text{court orders. By focusing on different types of features and hence criteria, overlapping categories have been created. These categories cut across each other with the consequence that some remedies belong in two categories.}\]

Damages, for example, may have a punitive, compensatory, or restitutionary goal, but, in any case, they are always coercive.\(^{82}\) After all, damages are awarded by means of a court order.\(^{83}\) Orders are inherently coercive, so all judicially-ordered remedies must be coercive; otherwise, there would have been no litigation.\(^{84}\) According to this scheme, therefore, damages must always appear simultaneously in at least two

\(^{78}\) See Tilbury, supra note 29, at 28.
\(^{80}\) ZAKRZEWSKI, supra note 34, at 75.
\(^{81}\) Id. at 70, 75.
\(^{82}\) Id. at 74.
\(^{83}\) Id. at 17
\(^{84}\) See, e.g., ZAKRZEWSKI, supra note 34, at 72–73 (suggesting that coercive orders encompass specific performance, injunctive relief, compensation, restitution, and punishment).
categories. As we have already noted, no taxonomical project can succeed if categories overlap.

In reality, Jimenez’s classifications do not really form a taxonomy at all, but more a set of Weberian ideal-types. In other words, they are abstractions that approximate reality in order to highlight potentially interesting features or tendencies. Ideal-types certainly have their place, but that is as models for the purposes of academic theory. Indeed, as Max Weber himself explained, “it is probably seldom if ever that a real phenomenon can be found which corresponds exactly to one of these ideally constructed pure types.” They are thus, by definition, an inappropriate basis for practical application by judges and practitioners.

2. Omission of Remedies

Douglas Laycock has proposed a slightly different scheme, into which he claims remedies “fit rather easily”. Specifically, his categories are as follows:

1. Compensatory remedies
2. Preventive remedies
   a. Coercive remedies
   b. Declaratory remedies
3. Restitutionary remedies

---

85 Id. at 75.
86 Id. (“So far as a taxonomy includes two or more criteria, the combination must be hierarchical. In other words, mixed classifications must be avoided.”). See also Sherwin, supra note 19, at 33–34 (commenting that Birks insisted “that legal categories must not overlap”).
87 An ideal type is formed by the one-sided accentuation of one or more points of view and by the synthesis of a great many diffuse, discrete, more or less present and occasionally absent concrete individual phenomena, which are arranged according to those one-sidedly emphasized viewpoints into a unified analytical construct.
89 Id. at 223–24.
92 LAYCOCK, supra note 36, at 2.
4. Punitive remedies
5. Ancillary remedies

The two-level hierarchy in Laycock’s scheme has the virtue of relegating the categories of coercion and declaration to secondary-level significance, and so avoids the “overlap trap” of attempting to compare remedial goals with the properties of court orders. But Laycock undoes this good work by including a new category of “ancillary remedies” which are not so much remedies as methods for the enforcement of remedies. As a result, we find ourselves once again in a situation of being asked to compare things that are incommensurate. Laycock has resolved an issue with coercive and declaratory remedies only to reproduce it with ancillary remedies.

But far bigger problems exist with both Laycock’s and Jimenez’s schemes. One major failing is that neither can cope with a whole range of remedies that Zakrzewski has called “transformative remedies.” These remedies, such as divorce, bankruptcy protection, naturalization, and the granting of political asylum, resolve a situation by placing the party seeking assistance in a wholly new position. By definition, such remedies do not seek to restore something from the past, nor do they seek to punish or compensate for past conduct. They might sometimes have the effect of preventing something from occurring, but—even with bankruptcy protection—the primary goal of such remedies is to enable a fresh start. Transformative remedies are almost entirely forward-looking remedies.

Because their schemes focus on the past, neither Laycock nor Jimenez address such remedies. Yet a purported taxonomy that fails to account for such a wide range of remedies is clearly unsustainable. Unfortunately, the failings of these attempts at taxonomy do not end there.

---

92 Id. at 2–3.
93 Id. at 4.
94 ZAKRZEWSKI, supra note 34, at 203; see also infra Part VI.
95 ZAKRZEWSKI, supra note 34, at 203–04.
96 Rather, these remedies “transform or modify the legal relations between the claimant and defendant which existed before trial.” Id. at 203.
97 A transformative remedy provides a fresh start because it alters the legal relationship between the parties before the order was instituted. Id.
98 Id. at 203.
3. Identical Remedies for Different Goals

Perhaps the most significant flaw in the goals-based approach to a taxonomy of remedies is its singular lack of explanatory power. In the interests of time and space, one example should suffice, taken from the law of torts. According to the type of case involved, the law of torts may be said to pursue one or more of any of the following goals:

1. General Deterrence
2. Specific Deterrence
3. Economic Efficiency
4. Distributive Justice
5. Boundary Maintenance
6. Denunciation and Stigmatization
7. Appeasement or Vengeance
8. Retribution
9. Compensation
10. Corrective Justice

---

99 See Kaye, supra note 28, at 76.
102 See, e.g., Guido Calabresi, The Costs of Accidents: A Legal and Economic Analysis (1970) (suggesting efficiency as a tertiary goal of “accident cost reduction”); Landes & Posner, supra note 100, at 13 (stating “the economic function of tort law is to optimize rather than minimize the number of accidents”).
105 See, e.g., Michael A. Mogill, Teaching Law Day: A Senior Moment, 1 Stetson J. Advoc. & L. 34, 35–36, 38 (2014), https://www2.stetson.edu/advocacy-journal/teaching-law-day-a-senior-moment/ (presenting a case study focused on the “Hot Coffee” case, in which a McDonald’s customer was awarded $160,000 for the injuries she had suffered after spilling hot McDonald’s coffee on herself, and the public’s reaction to the lawsuit).
Redress of Social Grievances

And yet the law of torts typically employs just one remedy to achieve all these goals. That remedy is compensatory damages, and such damages are calculated in the same way in every instance, regardless of the particular goal the law is apparently keen to pursue in the instant case.

There is simply no way that a goals-based approach to taxonomy can explain this. A goals-based approach would mean that each goal would be reflected in the plaintiff’s obtaining a remedy tailored precisely to that goal. This could happen either through the award of different remedies to satisfy different goals or by using different means of calculating compensatory damages. But the only time the law adopts such an approach is when punitive damages, qui tam, or treble damages are awarded in cases where retribution is appropriate. Otherwise, the same remedy of compensatory damages is awarded, calculated in the same way, irrespective of the goal(s) that the law wishes to emphasize in any given case.

We could, of course, try to fit each of the above objectives into Laycock’s more restricted list of legal goals. Then, for example, both general and specific deterrence might be included within the category of “preventive remedies,” while compensation and corrective justice could be placed within that of “compensatory remedies.” But that would achieve little because, according to the goals-based approach to taxonomy, the nature of the remedy would still be expected to change from category to category.
this occurs in practice is when the law awards punitive, treble, or *qui tam* damages.\textsuperscript{117} Even then, such awards must be calculated by reference to the amount of compensatory damages awarded.\textsuperscript{118} This is not punishment in the sense of either reflecting the degree of culpability or deterring similar conduct in the future.\textsuperscript{119} In fact, it is more significantly related to the harm done. So far as Laycock’s other categories are concerned, moreover, the same remedy of compensatory damages is still awarded and calculated on the same basis every time.

In other words, the only way to fit compensatory damages into a goals-based approach to taxonomy would be to assign this remedy to every category! But that is the epitome of self-contradiction. As we have already noted more than once, one of the bedrocks of taxonomy is that the categories involved must be mutually exclusive.

4. Different Remedies for Identical Goals

As if this were not bad enough, the goals-based approach is also unable to explain why remedies with apparently the *same* goal work so differently. Why, for example, is the measure of damages in contracts different from that in torts, when both are intended to compensate the victim? According to the goal-based approach, if the same remedy (compensatory damages) with the same objective is involved, then the result should be the same. Otherwise, by definition, the objectives would be missed. And yet compensatory damages in contracts are, in fact, calculated on a very different basis from compensatory damages in torts.\textsuperscript{120}

The goals-based approach cannot explain this conundrum at all. The glib response that “[d]amages in tort are designed to make the plaintiff whole”\textsuperscript{121} simply does not work, because contract law also seeks to make the innocent party “whole.”\textsuperscript{122} (It hardly seeks to make

---

\textsuperscript{117} See supra note 114.

\textsuperscript{118} On the relationship between harm caused and the amount of punitive damages awarded, see *Gore*, 517 U.S. at 580-83 and 31 U.S.C. §3730(d) (2006).


\textsuperscript{121} *Id.*

\textsuperscript{122} See, e.g., CAL. CIV. CODE § 3300 (“For the breach of an obligation arising from contract, the measure of damages... is the amount which will compensate the party
her “half.”)

The true answer is that there are different conceptions of what constitutes “wholeness” at work here. But this, of course, invites the question of where those different conceptions come from—to which the appropriate response is that there are different rights at stake in contracts from those in torts. This response points toward the only viable basis for a taxonomy of remedies. For, while it has been said that “[t]he nature of the right giving rise to a remedy can affect the remedy,”\(^{123}\) this is to state the truth in a very half-hearted manner. In fact, *the nature of the right involved inevitably dictates the type of remedy available.*

### III. A RIGHTS-BASED TAXONOMY

A taxonomy of remedies that is based on rights posits (a) that individuals have certain rights in certain contexts and (b) that the role of legal remedies is to vindicate those rights, provided that some other prerequisites have also been satisfied.\(^ {124}\) It follows, therefore, that, in order to classify remedies, it is first necessary to classify rights.

This is a task that was first attempted by John Austin in the late nineteenth century.\(^ {125}\) Austin’s work has been largely overlooked for over half a century because of the famously devastating criticism published by H.L.A. Hart in the 1960s.\(^ {126}\) But Austin’s reputation has recently been enjoying a degree of rehabilitation,\(^ {127}\) and it is my contention that combining some of the stronger points of Austin’s work with the improvements provided by Hart enables us to generate the sound taxonomy of rights that we need in order to develop a useful taxonomy of remedies.

**A. John Austin**

The foundation of Austin’s approach to rights and remedies lay in the following assertion:

\(^{124}\) See *AUSTIN, supra* note 37, at 43–44.
\(^{125}\) *Id.* at 43–45.
\(^{126}\) See *HART, supra* note 32, at 18, 20.
Rights and duties not arising from delicts [i.e., legal wrongs], may be
distinguished from rights and duties which are consequences of delicts, by the
name of primary (or principal). Rights and duties arising from delicts, may be
distinguished from rights and duties which are not consequences of delicts, by
the name of sanctioning (or secondary).\textsuperscript{128}

To put this in more modern terminology, a primary right is a right created without anyone behaving wrongfully, while a secondary right is a right created as a result of someone else’s wrongful conduct.

Inheritance is an example of a primary right.\textsuperscript{129} Except in cases where the decedent was the victim of a wrongful death, inheritance involves the acquisition of rights by the beneficiaries of the decedent’s estate without any wrongdoing.\textsuperscript{130} The same is true of rights created by a contract. Unless brought about by misrepresentation or duress, the formation of the contract that creates such rights involves no wrongdoing whatsoever.\textsuperscript{131}

Secondary rights, on the other hand, are created by another person’s wrongful conduct.\textsuperscript{132} A tort is the obvious example; breach of contract is another. Both are legal wrongs because they arise when someone breaches a legal obligation. In such cases, the innocent party must first have had certain primary—typically proprietary or contractual—rights, which the defendant was obliged not to infringe.\textsuperscript{133} Where infringement does occur, the victim obtains new, secondary rights in addition to the primary rights that she had before, and which entitle her to take action against (i.e., to sanction) the wrongdoer.\textsuperscript{134}

B. Power-Conferring Rules

Austin made a fundamental mistake, however, in insisting that only

\textsuperscript{128} Austin, supra note 37, at 43.

\textsuperscript{129} See, e.g., In re Estate of Dito, 130 Cal. Rptr. 3d 279, 287 (2011) (appellant’s right to receive share of estate as a surviving spouse was a primary right).

\textsuperscript{130} Inherit, Black’s Law Dictionary (10th ed. 2014); see, e.g., Freuler v. Helvering, 291 U.S. 35, 45 (1934) (recognizing beneficiary’s acquisition of rights as property rights).

\textsuperscript{131} See Austin, supra note 37, at 769 (noting that duties from contracts and quasi-contracts are created from their own existence, while breach of contract is not).

\textsuperscript{132} Carlos Manuel Vázquez, Treaty-Based Rights and Remedies of Individuals, 92 Colum. L. Rev. 1082, 1089–90 (1992).

\textsuperscript{133} Id. (“Secondary rules are ‘in a sense parasitic upon’ primary rules. . . . A secondary or remedial right is thus a right to obtain a remedy or sanction upon the violation of a primary right.”) (footnote omitted).

\textsuperscript{134} Id.
secondary rights are legally enforceable.\footnote{AUSTIN, supra note 37, at 766.} As we saw earlier, “not-wrongs” can also trigger legally-enforceable rights.\footnote{See supra Part II.} In one of the most famous law books of all time,\footnote{See Martin Krygier, The Concept of Law and Social Theory, 2 OXFORD J. LEGAL STUD. 155, 155 (1982) (describing H.L.A. Hart’s influential book, The Concept of Law).} H.L.A. Hart explained that Austin’s error lay in his view that all laws impose obligations.\footnote{HART, supra note 32, at 6–8, 27–28; AUSTIN, supra note 37, at 767.} But while some laws do indeed work this way, others impose no obligations at all. These laws, instead, \textit{confer powers}.\footnote{HART, supra note 32, at 27–28.}

Hart himself provided three examples. One was of a contract; another of a trust; the third was of marriage.\footnote{Id. at 28.} The law does not order anyone to get married or set up a trust, yet there plainly are legal rules governing how both activities may be accomplished.\footnote{Id.} And, while part of the law of contracts may be conceived as issuing a command not to breach a contract, the law issues no command to form a contract in the first place. It simply sets out the rules so as to “provide individuals with \textit{facilities} for realizing their wishes, by conferring legal powers upon them to create . . . conditions, [and] structures of rights and duties.”\footnote{Id. at 27.}

This is an important insight because it enables us to identify the origins of primary rights, which Austin had hardly considered at all. In fact, primary rights are precisely what are generated through the utilization of power-conferring laws.

\textbf{C. Primary Rights}

In contracts, for example, the parties create new rights and obligations for themselves by taking advantage of the doctrines of offer, acceptance, and consideration (or through the alternative doctrine of promissory estoppel).\footnote{Tim Kaye, Torts as Relational Contracts, in SHAKING THE WORLD GENTLY: A WEBFESTSCHRIFT IN HONOR OF PROFESSOR ROBERT DALE BICKEL 271, 295 (Tim Kaye ed. 2013), http://www2.stetson.edu/bickel/Webfestschrift/files/assets/basic-html/index.html?page295.} On the other hand, primary rights regarding property may be acquired not only by contract but also through laws governing inheritance, prescription, or adverse possession; equity
enables the acquisition of primary rights through the creation of a trust, while statutes enable the creation of primary rights by granting powers, such as to marry or collect tax.\textsuperscript{144}

The law neither obliges nor commands anyone to do any of these things. It simply provides the power to do so. When individuals take advantage of such powers, they generate primary rights that did not previously exist. Sometimes, as in the case of a contract, the parties will generate new primary rights both for themselves and for others. In other cases, as with births, gifts, trusts, activities subject to tax, or inheritance, those obtaining primary rights might not be among those utilizing the appropriate powers. But once a person obtains primary rights, the law both recognizes them and provides for them to be vindicated by a particular category of remedies.\textsuperscript{145}

The vindication of primary rights is precisely what is involved when a person seeks a remedy for a "not-wrong."\textsuperscript{146} In a case where a store assistant makes too much change, which is then unwittingly pocketed by the shopper, three sets of primary rights are implicated. The first set comprises the store’s proprietary, possessory rights to its merchandise and the cash in its register,\textsuperscript{147} which it will have obtained through various power-conferring rules, especially those involved in the formation of contracts. The second set comprises the shopper’s possessory rights to the money with which she is proposing to pay for the merchandise. The third set comprises the primary rights created when the sale itself is made. This enables the transfer of (a) the rights of ownership and possession over the merchandise from seller to buyer, and (b) the rights of ownership and possession over the purchase price from buyer to seller.

So what of the surplus change? As this approach makes clear, the rights of ownership and possession over that sum of money remain throughout with the store. The surplus change, in other words, belongs to the store. The primary rights to it were never at any stage transferred

\textsuperscript{144} See ROBERT STEVENS, TORTS AND RIGHTS 7, 10 (2007) ("Property rights can be acquired in a number of ways (gift, sale, inheritance, accession, etc.) . . . Promises which are intended to have legal effect and are supported by consideration are legally binding. However, contract is not the only way of voluntarily creating rights. Bare promises in deeds, promissory estoppel, and express trusts are examples of ways of voluntarily creating rights outside of contract.").

\textsuperscript{145} HART, supra note 32, at 92 (describing Hart’s so-called “rule of recognition”).

\textsuperscript{146} See Birks, supra note 19, at 25.

\textsuperscript{147} It might also have rights of ownership over both, but that would depend on its precise business arrangements and makes no difference to the point being made here.
to the buyer. One way of describing the position where the shopper has unwittingly obtained surplus change is to say that she has been unjustly enriched at the expense of the store. But that is unnecessarily opaque. The real point is that the store retains a primary right to the surplus change, and so is entitled to have it returned. If necessary, court action could, therefore, be taken to recover the sum involved.

Primary rights have the important distinction of being more than merely personal rights.\textsuperscript{148} Irrespective of how they are created, they are “real” rights in the sense that they attach to the property or sum of money involved, and so are enforceable against the world.\textsuperscript{149} Thus, if I purchase a car that is then stolen from me and sold to an innocent third party, I have the right to demand the return of the car from the innocent third party. Although the mechanism by which I obtained the primary right of ownership was contractual, that right itself is “real.” I can thus demand the return of the car from whomever happens currently to be in possession of it simply because it is mine.\textsuperscript{150}

**D. Interests**

Primary rights thus fit the definition of a right provided by the First Restatement of Property as “a legally enforceable claim of one person against another, that the other shall do a given act or shall not do a given act.”\textsuperscript{151} But legally recognized rights are not exclusively primary. In fact, the law also recognizes secondary and tertiary rights. The next question to address, therefore, is where do these other rights originate?

The answer to that question lies in the law’s acknowledgment of a wide variety of concepts and components that go well beyond rights. We have already seen, for example, that the very act of creating primary rights involves the use of something other than a right, namely a power, which the First Restatement of Property defines as “an ability on the part of a person to produce a change in a given legal relation by doing or not

\textsuperscript{148} See AUSTIN, supra note 37, at 46–47.

\textsuperscript{149} Id.

\textsuperscript{150} See Vicars v. Atl. Disc. Co., 140 S.E.2d 667, 673 (Va. 1965) (stating that the defendant, who purchased a previously stolen vehicle, cannot be considered a bona fide purchaser); see also infra Part IV(1)(ii).

\textsuperscript{151} RESTATEMENT (FIRST) OF PROPERTY § 1 (AM. LAW INST. 1936). This definition of a right has been adopted by Dennison v. N.D. Dep’t of Human Servs., 640 N.W.2d 447, 453 (N.D. 2002).
doing a given act.”152 The law recognizes other elements too, such as “privileges” and “immunities”153 or, indeed, “the object of any human desire.”154 We refer to all these elements collectively as “interests.”155

Primary rights are stronger than other interests because they demand universal recognition.156 Other interests might not necessarily be recognized by the law at all, and those interests that are recognized in one field or context might not be recognized in another.157 Thus an interest, other than a primary right, that is recognized as worthy of legal protection within torts might not be recognized as such in contracts. Unlike primary rights, therefore, this type of interest cannot be protected from the world, but only from specific individuals.158

Spouses, for example, have an interest in each other’s income. Neither has a primary property or contractual right to the other’s income, but the lack of a primary right does not mean that this interest never has legal protection. On the contrary, if one spouse is killed through the wrongful actions of a third party, the surviving partner may then be able to bring a wrongful death suit against the tortfeasor because that interest has been harmed.159 Moreover, if spouses divorce or otherwise become estranged, the interest in each other’s income may form the basis for a court order for the payment of alimony.160

E. Secondary Rights

Where the law does recognize an interest, it is protected by a secondary right.161 There are several characteristics that differentiate a secondary from a primary right. As is already apparent, secondary rights are not rights against the whole world. In other words, they are not “real” rights, but purely personal rights. This is because of a second
differentiating factor. As Austin noted, secondary rights exist only to sanction wrongdoing.\textsuperscript{162} They provide no right of redress in the case of a not-wrong because there is then nothing to sanction. Secondary rights may only be exercised against (a) a wrongdoer or (b) someone who is vicariously liable for that wrongdoing.\textsuperscript{163}

Constitutional law, property law, and the law of restitution have no means of creating secondary rights. While they all boast power-conferring rules, none provides a mechanism for redressing wrongs.\textsuperscript{164} Instead, they typically rely on the law of torts to provide meaningful secondary rights. Conversely, the law of torts deals only with secondary rights.\textsuperscript{165} It is, by definition, concerned only with legal wrongs without any concept of “not-wrongs.” It also has no mechanism for creating primary rights, for which it relies instead on the powers conferred within constitutional, property, and contract law.\textsuperscript{166}

In fact, as Figure 1 illustrates, only three branches of civil law deal with wrongs: torts, contracts, and trusts. A secondary right must, therefore, be vindicated by pleading one of the following actions: a tort, a breach of contract, or a breach of trust. As even a cursory knowledge of those actions will reveal, however, each of them requires the presence of an additional prerequisite before a secondary right can be vindicated. Unlike a primary right, which can demand vindication by its very infringement, vindication of a secondary right also requires proof that the wrongdoing caused loss of (or harm to) the interest that the right is intended to support.\textsuperscript{167}

Thus a claim in negligence will not succeed merely because the plaintiff can prove that the defendant breached his or her duty.\textsuperscript{168} Such a claim only establishes that the plaintiff had an interest that was infringed

\textsuperscript{162} AUSTIN, supra note 37, at 43. See supra Part III(A).
\textsuperscript{163} AUSTIN, supra note 37, at 44–45.
\textsuperscript{164} See, e.g., John C. P. Goldberg, The Constitutional Status of Tort Law: Due Process and the Right to a Law for the Redress of Wrongs, 115 YALE L. J. 524 (2005) (arguing that although the Constitution does not provide specific tort causes of action, there should be an individual right to seek redress of private wrongs through tort law reform).
\textsuperscript{166} See Kaye, supra note 143, at 293, 298–301.
\textsuperscript{167} See AUSTIN, supra note 37, at 763.
\textsuperscript{168} RESTATEMENT (SECOND) OF TORTS § 281 (AM. LAW INST. 1965) (stating the elements that must be proven to prevail on a negligence claim).
by a wrong. The plaintiff must also prove that the breach caused harm to that interest. As a former Director of the American Law Institute once wrote, when used in torts, a “breach of such a duty does not make the actor liable. . . . Thus, whether [the actor] is liable or not depends upon whether his breach of his duty results in injury to someone to whom his duty is owing. . . ." So, as I have said elsewhere, “the causation of harm is not by itself the basis for a successful claim, but neither is a violation of a right where no harm has been done.”

The same is true in an action for breach of contract or breach of trust. The breach itself is insufficient as a basis for vindicating the plaintiff’s secondary right. She must, in addition, prove that the breach in question either caused damage to her interest or else caused that interest to be lost altogether.

169 Id.
170 Id.
171 William Draper Lewis, Director’s Note to Restatement (First) of Torts § 1 (AM. LAW INST. 1934) (1933).
174 RESTATEMENT (THIRD) OF TRUSTS § 93 cmt. a (AM. LAW INST. 2012) ("[A] trustee is personally liable . . . for losses attributable to a breach of trust . . . .").
<table>
<thead>
<tr>
<th>BRANCH OF LAW</th>
<th>PRIMARY (&quot;NOT-WRONG&quot;)</th>
<th>SECONDARY (WRONG)</th>
<th>TERTIARY (COURT ORDER)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTITUTIONAL</td>
<td>Constitutional Rights</td>
<td></td>
<td>Judgment</td>
</tr>
<tr>
<td></td>
<td>Natural/Inalienable</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rights</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONTRACTS</td>
<td>Offer AND Acceptance</td>
<td>Breach of</td>
<td>Judgment</td>
</tr>
<tr>
<td></td>
<td>AND Consideration OR</td>
<td>Contract</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Promissory Estoppel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRUSTS</td>
<td>Trust</td>
<td>Breach of Trust</td>
<td>Judgment</td>
</tr>
<tr>
<td>PROPERTY</td>
<td>Inheritance Contract</td>
<td></td>
<td>Judgment</td>
</tr>
<tr>
<td></td>
<td>Prescription</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adverse Possession</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RESTITUTION</td>
<td>Unjust Enrichment</td>
<td></td>
<td>Judgment</td>
</tr>
<tr>
<td>TORTS</td>
<td></td>
<td>Tort</td>
<td>Judgment</td>
</tr>
<tr>
<td>MISCELLANEOUS</td>
<td>Marriage Divorce</td>
<td></td>
<td>Judgment</td>
</tr>
<tr>
<td></td>
<td>Custody</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bankruptcy</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Protection</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Salvor’s Reward</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tax</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 1: Origins of Rights*
F. Comparing Primary and Secondary Rights

The notion of “duty” that correlates with a “right,” therefore, differs according to the type of right involved. A primary right imposes a duty on everyone else not to infringe that duty by any means, including not-wrongs.\(^\text{175}\) A secondary right, on the other hand, imposes a duty only on a certain person or class of persons to refrain from wrongfully causing harm to the particular interest relating to the right.\(^\text{176}\)

One other point bears highlighting. It has already been noted that the concept of “interests” includes, but is not exhausted by, rights. It has also been noted that interests can originate and be recognized in different fields of law. What might not yet be so clear, however, is that one of the ways that interests arise is through the creation of primary rights. In fact, the creation of a primary right automatically generates a corresponding interest at the same time.

When I purchased my car, for example, I obtained both primary rights of ownership and possession over it and an interest in it. If the need arises, therefore, I may be able to assert both primary and secondary rights. If my car were to be stolen and then sold to an innocent third party, I would be able to assert my primary right to the car against the third party. Even though she would have engaged in a not-wrong and had no other relationship with me, my primary right in the car remains good against the whole world, and I am entitled to have the car returned to me because it is mine. But I could also assert my secondary right against the thief (by bringing an action in the tort of conversion), because his theft would have caused harm to my interest in the car.\(^\text{177}\) The nature of the remedy available against the thief would, however, differ from that against the person currently in possession of the car, because the person currently in possession would have performed a not-wrong, while the thief would have committed a wrong. This is precisely why it is so important to develop a sound taxonomy of remedies.

Interests created by primary rights thus have a special degree of protection. Naturally, they are protected by the primary rights themselves. But that protection is reinforced by secondary rights. Interests created by means other than primary rights do not enjoy this

\(^{175}\) See HART, supra note 32, at 78–79.

\(^{176}\) AUSTIN, supra note 37, at 761.

\(^{177}\) I cannot, of course, obtain double recovery for the same loss, and so I might not be able to claim both remedies at the same time. See infra Part IV(3)(ii).
special status. If they are recognized by the law at all, they are supported only by secondary rights. The rights to “Life, Liberty and the pursuit of Happiness” are examples of such rights. Although the Declaration of Independence calls them “unalienable,” that is a matter of political rhetoric. As a matter of law, a person can legitimately forfeit his or her right to life through someone else’s exercise of reasonable self-defense, a judicially-ordered sentence of death, or an act of war. The right to life is therefore plainly not a primary right that may be asserted against the whole world. It is instead a secondary right. When the interest to which the right to life relates is lost or harmed as a result of the wrongdoing of someone else—as in the case of someone whose life expectancy is shortened by an accident caused by a third party’s negligence—the victim may seek to have that right vindicated in the law of torts.

Similarly, no one can assert the right to liberty as a primary right. It is not good against the world, and it cannot be vindicated where another person has engaged only in a not-wrong. This is why, for example, American citizens cannot successfully claim that the right to liberty makes them immune from an obligation to pay tax. But the right to liberty is certainly a secondary right, which supports a person’s

---

178 See, e.g., Joseph William Singer, *The Legal Rights Debate in Analytical Jurisprudence from Bentham to Hohfeld*, 1982 WIS. L. REV. 975, 988 (1982) (“It is not true that merely because one has the legal liberty to do an act that others have legal duties not to interfere with the permitted act.”).

179 *The Declaration of Independence* para. 2 (U.S. 1776).

180 Id.

181 See, e.g., Joel Feinberg, *Voluntary Euthanasia and the Inalienable Right to Life*, 7 PHIL. & PUB. AFF. 93, 98 (1978) (arguing that believing in the right of life is hypocritical when one supports war, capital punishment, or self-defense).


183 See, e.g., Feinberg, *supra* note 181, at 98 (grouping liberty in property rights with the right to life, which is not a primary right).

“fundamental liberty interest.” It may be vindicated in circumstances where a person’s mobility, for example, has been wrongfully diminished, or where someone is falsely imprisoned.

The rights to life and liberty provide excellent examples of how the law may only recognize interests other than primary rights in certain contexts. Recognition of the right to the pursuit of happiness is even more limited. Although this secondary right is acknowledged by the law of torts, the circumstances in which it may be enforced are highly circumscribed.

Finally, it is worth noting that cases sometimes arise where one party exercises a primary right against another, while the latter has an enforceable secondary right against the former. Sims v. Century Kiest Apartments provides an example of such a case. Sims was a tenant of Century Kiest Apartments under a week to week tenancy for approximately nine years. He had, however, repeatedly complained about the state of the premises, and had reported violations of various housing, building and health codes to city authorities. In retaliation, and despite Sims’s rent being fully paid, Century Kiest terminated the tenancy with ten days’ notice. Sims sought damages for wrongful eviction, to which Century Kiest replied that it had a right to choose whom to have on its premises and could, therefore, lawfully evict Sims if it so chose.

---


186 See, e.g., Bernier v. Boston Edison Co., 403 N.E.2d 391, 398 (Mass. 1980) (sustaining the jury verdict, which found negligent the design of light pole that struck plaintiff during a car accident).


188 See, e.g., Thing v. La Chusa, 771 P.2d 814, 829 (Cal. 1989) (stating that there is limited ability to recover in negligent infliction of emotional distress, even though close relatives experience a great deal of suffering after losing a loved one).


190 Id. at 527.

191 Id.

192 Id.

193 Sims, 567 S.W.2d, at 528.
The Texas Court of Appeals effectively held that both parties were correct. In our terms, Century Kiest had primary rights of ownership and possession over the premises. While it gave up the right of possession on a weekly basis, it could choose at the end of each week not to continue to do so. Its vindication of this right by terminating Sims’s tenancy and evicting him was not, therefore, something with which the court would interfere. Nevertheless, Sims’s tenancy had not only granted him a primary right to possession from week to week, but also a broader interest. His eviction for illegitimate reasons had been a wrong that caused him to lose that interest, and so he was entitled to pursue a remedy to vindicate that right by means of a claim for damages.

G. Tertiary Rights

As Figure 1 suggests, there is also a third class of rights. This comprises rights that are created by the order of a court. They entitle the prevailing party to enforce such an order. Tertiary rights derive from the fact that, by definition, the prevailing party must have had a prior right or interest that merited judicial protection. “To take the contractual example, on the primary level are the rights born of the contract; on the secondary level are the . . . rights born of the breach; and at the tertiary level is the right born of the judgment itself . . .”

Yet tertiary rights are distinct from the rights from which they derive. Once a court order is entered, the tertiary rights that it creates subsume the prior rights, so that the latter cease to be independently enforceable. What happens to a primary right to a debt, for example, was explained by Judge Cardamone in the Second Circuit Court of Appeals: “The general rule under New York and federal law is that a debt created by contract merges with a judgment entered on that contract, so that the contract debt is extinguished and only the judgment debt survives.”

This means that, if a simple debt has been confirmed in a judgment, the rate of interest on the debt will change from the day on which the

194 Id. at 532.
195 See Birks, supra note 19, at 30.
196 Id.
197 Id.
198 See Birks, supra note 19, at 30.
199 Westinghouse Credit Corp. v. D’Urso, 371 F.3d 96, 102 (2d Cir. 2004).
judgment takes effect. There will, accordingly, be a period for which “pre-judgment interest” remains payable (reflecting the original, primary right), followed by a period for which “post-judgment interest” is added every day until the debt is satisfied (reflecting the new, tertiary right). Even if the parties themselves had stipulated their own post-judgment rate, the issuance of the court order will mean that the new tertiary right supersedes the primary right from which it derives.

IV. REPLICATIVE REMEDIES

Having established our taxonomy of rights, it is now possible to construct a meaningful taxonomy of remedies. On the basis that there are three categories of rights—primary, secondary, and tertiary—it should come as no surprise that there are also three categories of remedies. They are as follows:

1. Replicative Remedies
2. Substitutionary Remedies
3. Transformative Remedies

Figure 2 identifies the appropriate classification for each of the most important civil remedies. (While the terms, “liquidated damages” and “quantum meruit,” have been placed within two different classifications, this is because, as will be explained below, the same label is typically used by the courts to describe two different remedies.) Figure 2 also subdivides each class according to whether the remedy in question is of common law, equitable, or statutory origin and (although little turns on this point) whether the remedy involves the payment of money or not.

---

201 "When a meaningful judgment [is] entered has significance . . . because that is when post-judgment interest begins to run and pre-judgment interest ceases to apply." Id. at 98.
202 The labels are taken from Zakrzewski, supra note 38, at 79, 81.
203 See infra Parts IV(C)(1) & V(B).
<table>
<thead>
<tr>
<th>Type of Remedy</th>
<th>Reparative</th>
<th>Substitutionary</th>
<th>Transformative</th>
<th>Monetary</th>
<th>Non-Monetary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Statute</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Equity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Common Law</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Taxonomy of Remedies
Replicative remedies exist to vindicate primary rights.\textsuperscript{204} We call them “replicative” because they are designed to replicate the very right that they vindicate. As was noted in Part III and illustrated in Figure 1, primary rights can be created in several different fields of law. It follows, therefore, that these same branches of law—constitutional law, contracts, equity, property, restitution, and a miscellaneous group of statutes—each boast one or more replicative remedy to vindicate the primary rights within their respective domains.

It might be thought that there would be little purpose in having a remedy that does nothing more than replicate a prior right, but the fact that there is a whole debt collection industry should be enough to make clear that such a view is misplaced. The point of a replicative remedy lies in the fact that the creation of primary rights enjoys the law’s backing.\textsuperscript{205} If those primary rights are not satisfied, the plaintiff will then be entitled to take action in the courts, if necessary.\textsuperscript{206} This is often enough to encourage the defendant to comply without further ado.

Lack of space precludes discussion of every replicative remedy. But, in order to show how a sound taxonomy aids the pursuit of clarity and consistency in the law, the rest of this Part addresses some of the more important replicative remedies.

\textit{A. Property Law}

\textit{1. Replevin}

Among other things, property law determines who has the right to possession of specific chattels.\textsuperscript{207} Rights of possession may come into existence through contract, gift, or inheritance,\textsuperscript{208} and without any predicate wrongdoing, and so are clearly types of primary rights.\textsuperscript{209} This means that they must be vindicated by one or more replicative remedies, which will enable the person entitled to possession to obtain precisely

\textsuperscript{204} See Partlett, supra note 16, at 2057–58, fig.5.1.
\textsuperscript{205} See \textsc{Austin}, supra note 37, at 767.
\textsuperscript{206} Id.
\textsuperscript{207} Law of property, BLACK’S LAW DICTIONARY (10th ed. 2014).
\textsuperscript{209} See Birks, supra note 19, at 13 (“Primary rights describe a person’s initial legal entitlements.”) (quoting Kit Barker, Rescuing Remedialism in Unjust Enrichment Law: Why Remedies are Right, 57 CAMBRIDGE L. J. 301, 319 (1998)).
that.\footnote{See Stephen A. Smith, Rights and Remedies: A Complex Relationship, in TAKING RIGHTS SERIOUSLY 47 (Robert J. Sharpe & Kent Roach eds., 2009), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2006626.} In fact, the law recognizes two such replicative remedies: detinue and replevin.\footnote{LAYCOCK, supra note 36, at 481.} Nowadays, the only significant difference between them is that detinue is available when property has been \textit{retained} by someone else, whereas replevin is the appropriate remedy when property has been \textit{taken}.\footnote{Maggio v. Zeitz, 333 U.S. 56, 63 (1948).}

In \textit{Menzel v. List}, the plaintiff and her husband sought to recover a painting by Marc Chagall that was left in their apartment in Belgium when they fled from the invading Nazis in 1941.\footnote{Menzel v. List, 267 N.Y.S.2d 804, 806 (N.Y. Sup. Ct. 1966), modified on other grounds 279 N.Y.S.2d 608 (N.Y. App. Div. 1967), modification rev’d 246 N.E.2d 742 (N.Y. 1969).} The Menzels later settled in the United States.\footnote{Id. at 807.} After the war had ended, they searched for the painting without success until 1962, when it was discovered in the possession of the defendant, List.\footnote{Id.} He refused to return it on the grounds that he had purchased the painting in good faith from a well-known art gallery.\footnote{Id.}

List had clearly done nothing wrong; he had simply taken as true the word of a respected dealer. Yet Mrs. Menzel was the only person with a primary right to possession. While Mrs. Menzel and her husband acquired the painting through purchase (i.e., by means of the powers conferred by the law of contracts),\footnote{The Menzels had purchased the painting in 1932 from the Galerie Georges Giroux in Brussels for 3800 Belgian francs (about the equivalent of $150 at the time). Menzel, 267 N.Y.S.2d at 807–08.} it would be of no consequence if they had alternatively acquired it by gift or inheritance (i.e., by means of the powers conferred by the law of property). What matters is that Mrs. Menzel had acquired her husband’s share through inheritance when he died in 1960.\footnote{Id. at 807.} As the person with the primary right to possession of the painting, she was clearly entitled to vindicate that right through the remedy of replevin against whomever happened to be currently in possession. Accordingly, List was ordered to return the painting to Mrs.
2. Detinue

An analogous situation arose in the case of *D.T. Vicars v. Atlantic Discount Co.* Since the case involved retention of the property by a bailee rather than a physical taking, it implicated not replevin, but detinue.

In *Vicars*, Mrs. Powell purchased a Chevrolet Corvette from Albany Motors with Atlantic Discount Company, Inc., financing the purchase. While Mrs. Powell acquired the primary right of possession of the Corvette, the sale agreement stipulated that ownership was to be transferred to Atlantic until the balance of the loan was paid off. In case of default, Atlantic would be entitled to immediate possession. Two months later, the car broke down and Mrs. Powell left it with a mechanic, Oscar Parker, to repair and deliver back to her. He never did. Without the car, Mrs. Powell defaulted on the payments. Unknown to her or Atlantic, the car had, in fact, been sold from dealer to dealer until purchased by D.T. Vicars. After discovering the car’s location, Atlantic then sought the remedy of detinue so as to recover possession of the vehicle from Vicars.

In other words, by utilizing powers conferred by the law of contracts, the agreement between Mrs. Powell, Albany Motors, and Atlantic Discount created the following primary rights:

1. Albany became entitled to immediate payment of the purchase price;
2. Mrs. Powell became entitled to immediate possession of the Corvette; and
3. Atlantic became the owner of the Corvette, with the right to repayment of the loan (plus interest), or to immediate

---

219 *Id.* at 820.
221 *Id.* at 669.
222 *Id.*
223 *Id.*
224 *Vicars*, 140 S.E.2d at 669.
225 *Id.*
226 *Id.*
227 *Id.*
228 *Vicars*, 140 S.E.2d at 669–70.
229 *Id.* at 669.
possession of the car in the event that Mrs. Powell defaulted.  

Both Albany’s and Mrs. Powell’s primary rights had been satisfied in accordance with the terms of the contract, and thus required no remedy to vindicate them. Atlantic, however, had received neither the payments due nor possession of the vehicle. The contract did not permit Atlantic—i.e., had not granted it a primary right—to obtain the outstanding balance from Mrs. Powell. Instead, Atlantic, who already held title, was entitled simply to possession of the car. That meant seeking a remedy against the person currently in possession, namely Vicars.

Here, in other words, is another classic instance of a replicative remedy being sought for a “not-wrong.” Vicars had done nothing other than enter into an apparently legitimate contract to purchase the car from another dealer, who had also done nothing wrong. Yet Atlantic obtained the primary right of possession to the vehicle from the moment that Mrs. Powell defaulted, and no subsequent sales could alter that fact. As we all know, nemo dat quod non habet: no-one can pass on what he does not have. Therefore, no one other than Atlantic could pass on the right to immediate possession of the Corvette. As the entity with that right, Atlantic could seek to have it vindicated against whomever was currently in actual possession.

3. Ejectment (or Quiet Title)

Ejectment (sometimes known as quiet title), is another replicative remedy recognized by the law of property. Its history, as Justice Marshall once noted, has been such that:

[E]jectment eventually developed into the primary means of trying either the

---

230 Id.
231 Id. at 671–73.
232 Vicars, 140 S.E.2d at 672.
233 Birks, supra note 19, at 25 (stating that upon a cause of action to address a not-wrong, the court is asked to recognize a primary right, not to remedy a wrong).
235 See, e.g., Idaho v. Coeur d’Alene Tribe of Idaho, 521 U.S. 261, 266, 275 (1997) (plurality opinion) (plaintiff sought declaratory judgment to affirm right to quiet title and exclusive use and occupancy of land); Glaski v. Bank of America, 160 Cal. Rptr. 3d 449, 456 (Cal. Ct. App. 2013) (plaintiff asserted cause of action for quiet title due to broken chain of title, which was the result of a defective loan transfer).
title to or the right to possession of real property.

In particular, ejectment became the principal means employed by landlords to evict tenants for overstaying the terms of their leases, nonpayment of rent, or other breach of lease covenants.\footnote{Pernell v. Southall Realty, 416 U.S. 363, 373 (1974) (footnotes omitted).}

In other words, ejectment is a replicative remedy that vindicates the primary rights of ownership and/or possession of real property. It thus plays a role analogous to that of detinue and replevin in relation to personal property.

Tenants who overstay their leases, or who fail to pay rent, or who breach the terms of their lease in some other way, are all perpetrators of some form of legal wrong.\footnote{See, e.g., Willison v. Watkins, 28 U.S. 43, 49 (1830) (rejecting tenant’s adverse possession claim because he was a wrongdoer who overstayed his lease).}

So too are trespassers with no prior relationship with the owner.\footnote{See Schwartz v. Heyden Chem. Corp., 188 N.E.2d 142, 144 (N.Y. 1963).} As a result, one may think the law provides a remedy so as to vindicate not primary rights, but secondary rights. After all, secondary rights are created by a legal wrong.\footnote{Birks, supra note 19, at 12.} The fact that a plaintiff happens to enjoy secondary rights, however, does not mean that she simultaneously loses pre-existing primary rights. In fact, when wrongdoing occurs that violates her primary rights, the plaintiff will then potentially be able to vindicate both primary and secondary rights, provided that they do not lead to double recovery for the same harm.\footnote{See, e.g., Landgraf v. USI Film Products, 511 U.S. 244, 253 (1994) (discussing the prevention of awarding duplicate compensatory damages).}

Thus an owner of realty may choose to vindicate his or her primary rights by means of ejectment, irrespective of whether she chooses also to attempt to vindicate her secondary rights.\footnote{Metromedia Co. v. W.C.B.M. Maryland, 610 A.2d 791, 792 (Md. 1992) (holding that an officer of a corporation who is unlawfully occupying property may be held liable for damages in an ejectment action); Mack v. Fennell, 171 A.2d 844, 845 (1961) (stating that plaintiff is entitled to treat an ejected tenant as either an occupant and recover rent or as a trespasser and recover damages, he can seek recovery under either claim—but not both); Slater v. Shell Oil Co., 137 P.2d 713, 715 (Cal. Dist. Ct. App. 1943) (stating that, although the plaintiff could have sought ejectment and damages originally, “where . . . a party elects to sue for damages past and prospective he is deemed to have waived the invasion and consented to the occupancy of the land.”).} Indeed, there may be little point in attempting to vindicate a secondary right against an impecunious tenant or trespasser because such a remedy will typically demand financial compensation, and the defendant might simply lack the
means to pay. In such circumstances, it makes sense for the plaintiff just to utilize the replicative remedy available.

4. Self-Help Remedies

One interesting feature common to the remedies of replevin, detinue, and ejectment is that they can be avoided if the injured party just takes the matter into his or her own hands. 242 Personally repossessioning the painting in Menzel, or the car in Vicars, or locking out an overstaying tenant, are all lawful means of regaining possession of the property at stake, so long as no breach of the peace or other excessive behavior is involved. 243 Such repossessions are examples of what may be called “self-help remedies,” and are remarkably commonplace. Abating a private nuisance—which, in the case of noise pollution, might literally achieve quiet title—is another example of a remedy that may be effected either with or without court approval or assistance. 245 In some states, this is even set out expressly in legislation. In Montana, for example, statute provides that: “A person injured by a private nuisance may abate it by removing or, if necessary, destroying the thing that constitutes the nuisance without committing a breach of the peace or doing unnecessary injury.”

Strangely, however, some commentators have argued that any


246 MONT. CODE ANN. § 27–30–302 (2014). Such abatement might, for example, involve what would otherwise be a trespass onto another person’s premises in order to prevent a flood of water over the abator’s own property. See State v. Moffett, 1 Greene 247, 249 (Iowa, 1848).
taxonomy of remedies should exclude self-help remedies. Laycock, for example, has asserted that “[a] remedy is anything a court can do for a litigant who has been wronged or is about to be wronged,”247 while Zakrzewski has argued that, until a court order has been issued, it makes no sense to talk of a “remedy.”248

These baffling assertions create an unnecessary disconnect from the general understanding of what constitutes a remedy. To most people, the point of a remedy is to resolve a problem. If the law empowers someone to do so without resorting to judicial proceedings, it seems quite arbitrary to refuse to dignify such a power with the term, “remedy.” In other words, “although [self-help remedies] are not judicial, they are legal.”249 The absurdity of suggesting otherwise is immediately evident upon asking, “If the label ‘remedy’ is out of bounds, what should we call such a power?”

In any event, the taxonomy of remedies proposed here can easily cope with self-help remedies. This is because, as the examples of replevin, detinue, ejectment, and abatement amply illustrate, every instance of a self-help remedy involves nothing other than a vindication of a primary right.250 Self-help remedies, in other words, are simply replicative remedies that can be exercised without prior judicial approval.

B. Contracts

1. Debt

Debt is another replicative remedy that may be obtained by self-help. Yet it has been strangely overlooked in casebooks and classes on remedies. The leading casebook on remedies, for example, makes no mention of debt.251 Perhaps this omission goes some way to explaining why the Supreme Court got in such a muddle in Bowen v. Massachusetts.

An action for debt, however, is one of the very oldest remedies

247 LAYCOCK, supra note 36, at 1 (emphasis added).
248 ZAKRZEWSKI, supra note 34, at 44–48.
250 Even Zakrzewski maintains that self-help remedies “are often secondary (or remedial) rights as they often arise from breaches of primary duties.” ZAKRZEWSKI, supra note 34.
251 LAYCOCK, supra note 36.
known to the common law, with a history stretching all the way back to the twelfth century. In fact, the remedy of detinue is an offshoot of the much older remedy of debt. It therefore pre-dates the development of equity by over 300 years. Equity did not appear until the late fourteenth to early fifteenth century. W. Hamilton Bryson, Equity and Equitable Remedies, in ENCYCLOPEDIA OF THE AMERICAN JUDICIAL SYSTEM 545 (Robert J. Janosik ed., 1987).

Debt is a common law remedy that existed long before equity was even a twinkle in the Lord Chancellor’s eye. Debt also pre-dates the law of contracts by around 600 years. Although I have placed it here under the rubric of the law of contracts (because that is the source of most debts nowadays), it remains true that many debts arise out of relationships that have nothing to do with the law of contracts. As John William Smith wrote in the middle of the nineteenth century:

If the record create a debt, that is, render a sum certain payable by the one party to the other, an action of debt will lie to enforce payment . . . . The action of debt lies in every case where there is a liquidated pecuniary duty from one person to another . . . .

Fifteen years later, the U.S. Supreme Court observed that:

Debt lies whenever a sum certain is due to the plaintiff, or a sum which can readily be reduced to a certainty—a sum requiring no future valuation to settle its amount. It is not necessarily founded upon contract. It is immaterial in what manner the obligation was incurred, or by what it is evidenced, if the sum owing is capable of being definitely ascertained.

Edward Jenks subsequently explained that, ever since the early days of the common law, “a Writ of Debt could only be issued for a specific
sum—the ‘very debt itself’ was to be recovered. The action could not be employed to recover ‘unliquidated damages’; because there was no jury to assess them.”

Indeed, the law originally took this to its logical conclusion, even though that would nowadays seem absurd:

[T]he original notion of the Writ of Debt was as “real” as anything could possibly be. Even when the subject of the action was a sum of money . . . the mind of the framer of the writ [was] evidently bent on getting back the specific coins lent.

So far as consumer debt is concerned, the modern definition of a debt is contained in the Fair Debt Collection Practices Act (“FDCPA”):

The term “debt” means any obligation or alleged obligation of a consumer to pay money arising out of a transaction in which the money, property, insurance, or services which are the subject of the transaction are primarily for personal, family, or household purposes, whether or not such obligation has been reduced to judgment.

It is significant that there is now a statutory definition of a debt: actions for debt are among the most commonly utilized remedies known to the courts. Indeed, the remedy of debt is now so ubiquitous that it has spawned a whole debt collection industry.

Today, a debt typically arises when two parties have agreed that one will pay the other for some sort of service. (Utilities are a common example.) When that service has been provided (or when the agreed date for payment arrives) the payment becomes due. At that point, the service provider enjoys a primary right to the payment because of the utilization of legal powers to create a legally-binding

---

260 The term “real” is used here to refer to a right that attaches to property, rather than a personal right. Id. at 58.
261 Id.
263 See Levontin, supra note 14; see also THOMAS ATKINS STREET, THE THEORY AND DEVELOPMENT OF COMMON-LAW ACTIONS 127 (Beard Books 1999) (1906) (noting that actions for debt are among the broadest and diverse remedies sought).
264 See, e.g., Debt, BLACK’S LAW DICTIONARY (10th ed. 2014) (defining debt as “[l]iability on a claim; a specific sum of money due by agreement or otherwise . . . [or a] nonmonetary thing that one person owes another, such as goods or services”).
obligation. If the customer pays the fee as agreed, then the provider’s primary right is satisfied. If, however, the customer fails to pay, then the provider is entitled to avail itself of the remedy of debt. In other words, it can demand that the agreed sum be paid.

Several features of the remedy of debt should now be apparent:

(1) It is a common law, and not an equitable, remedy.

(2) It involves a fixed or determinate sum of money, because it is the “very debt itself” that is to be recovered. It needs no determination by a jury and is, therefore, quite distinct from damages, where the sum of money involved is “at large” until determined by a jury.

(3) It arises from the use of powers (typically, but not exclusively, those conferred by the law of contracts) to generate a primary right to a specified sum of money.

(4) Because it vindicates a primary right, the remedy of debt must be replicative.

(5) As a replicative remedy, debt is susceptible to self-help.

In case it is contended that the last point has not yet been established, it is worth recalling the provision of the FDCPA previously quoted. It ends with the phrase, “whether or not such obligation has been reduced to judgment.” As this implies, a debt may be lawfully collected even without recourse to the courts. Indeed, the purpose of the FDCPA is to regulate that very process, for almost the whole debt collection industry has grown up with the objective of recovering debt as a self-help remedy.

266 Id. at 516.
267 Id. at 511.
268 As Holmes once noted, “debt for the detention of money was the twin brother of the action brought for wrongfully withholding any other kind of chattel.” HOLMES, supra note 255, at 252.
269 Compare Debt, BLACK’S LAW DICTIONARY (10th ed. 2014) (“a specific sum of money due by agreement or otherwise”) with Damages, BLACK’S LAW DICTIONARY (10th ed. 2014) (“Money claimed by, or ordered to be paid to, a person as compensation for loss or injury”).
270 See Birks, supra note 19, at 31.
271 See supra Part IV(A)(4).
272 See supra note 262.
2. Action for the Price

The remedy known as “action for the price” is a form of debt that arises from a contract for the sale of goods.274 It was first formalized in England in § 49 of the Sale of Goods Act 1893,275 and is now set out, for the purposes of American law, in § 2–709 of the Uniform Commercial Code (“UCC”).

An important facet of § 2–709 is how it expressly distinguishes an action for the price from an action for damages. Section 2–709(3) provides:

After the buyer has wrongfully rejected or revoked acceptance of the goods or has failed to make a payment due or has repudiated (Section 2–610), a seller who is held not entitled to the price under this section shall nevertheless be awarded damages for non-acceptance under the preceding section.276

In other words, a creditor may take advantage of the replicative remedy of an action for the price to recover the debt. But, if unable to exercise that remedy, the creditor may nevertheless vindicate her secondary right with a claim for damages. Debt and an action for the price are, therefore, quite different from the remedy of damages. Both the former are replicative remedies. Damages, however, are a substitutionary remedy.

3. Specific Performance

Unlike the remedies that have been discussed so far, specific performance is an equitable remedy.277 But this has no effect on the remedy’s replicative nature. On the contrary, there could scarcely be a more replicative remedy, for the objective of specific performance is precisely to give effect to—i.e., to replicate—the terms of the contract to which it relates.278

The pattern by which a replicative remedy becomes available should, by now, be somewhat familiar. In the case of specific performance, it starts with the power-conferring doctrines of contract law.279 These create a number of primary rights for each party to the

275 Sale of Goods Act 1893, 56 & 57 Vict. c. 71, § 49 (Eng.).
276 U.C.C. § 2–709(3) (AM. LAW INST. & UNIF. LAW COMM’N. 2010).
277 Laycock, supra note 36, at 6.
278 Id. at 5.
279 See Helge Dedek, From Norms to Facts: The Realization of Rights in Common and
contract, including the right of one party to the performance of some work or service or sale from the other.\textsuperscript{280} If that is not forthcoming, specific performance may lie (subject to equitable principles)\textsuperscript{281} to enable the plaintiff to vindicate her primary rights by giving her the very performance she expected.\textsuperscript{282} The point, as § 360 of the Second Restatement of Contracts explains, is to provide precisely what was expected because there is no adequate substitute.\textsuperscript{283} Thus specific performance replicates the primary right to performance of the contractual obligation.

It is easy to see why equity created the remedy of specific performance. Without it, the common law had created a one-sided set of remedies, where a seller could pursue the very debt owed, but a buyer had no ability to obtain the very performance due. Specific performance thus leveled the playing field to some extent. Nevertheless, because of its equitable origins, it puts the power to determine the most appropriate remedy in the hands of the court rather than those of the plaintiff.\textsuperscript{284} By contrast, as we saw in \textit{Menzel} and \textit{Vicars}, it is the plaintiff who makes that choice when the replicative remedy concerned originated at common law.\textsuperscript{285}

\textbf{4. Reformation}

Reformation of a written instrument is another equitable replicative remedy, albeit one exercised much less frequently.\textsuperscript{286} “The purpose of reformation by a court of equity is to make an erroneous instrument express correctly the real agreement between the parties.”\textsuperscript{287} In such circumstances, the written document apparently denies the primary rights that the prior exercise of contractual powers was intended to

\textit{Civil Private Law}, 56 McGill L. J. 77, 109 (2010) ("Specific performance is an instance of diverging reactions to the breach or non-performance (or mal-performance) of a contract.").

\textsuperscript{280} \textit{Id.} at 110.

\textsuperscript{281} See, \textit{e.g.}, Van Wagner Advert. Corp. v. S. & M. Enter., 67 N.Y.2d 186, 195 (1986) (denying specific performance because doing so would be disproportionately harm the defendant).

\textsuperscript{282} RESTATEMENT (SECOND) OF CONTRACTS § 358 (AM. LAW INST. 1981).

\textsuperscript{283} \textit{Id.} at § 360, cmt. c.

\textsuperscript{284} \textit{Id.} at § 357.

\textsuperscript{285} \textit{See supra} Parts IV(A)(1)–(2).


\textsuperscript{287} Manning Lumber Co. v. Voget, 216 P.2d 674, 680 (Or. 1950).
create. Reforming or rectifying the document so that it accurately records the parties’ primary rights is clearly a replicative act.

Reformation of a document may be obtained in either of two circumstances. One is where there has been a mutual mistake. The other is where the reason for the document’s inaccuracy is “the fraud of one of the parties and mistake of the other.” The latter might suggest that wrongdoing must be involved, so that the relevant right will be secondary, which would make the appropriate remedy not replicative but substitutionary. But that would be a misunderstanding.

Once the contract is concluded, the parties obtain certain primary rights. They may then vindicate those rights, in preference to vindicating any secondary rights that they might also acquire, as was the case in Menzel and Vicars. The fact that reformation is an equitable remedy does not alter this analysis. Reformation of a written instrument upholds “the very thing” agreed and is thus a replicative remedy. The significance of the remedy’s equitable origins is simply that, as already noted, the choice of remedy is made not by the plaintiff but by the court.

C. Restitution

1. Quantum Meruit

Remedies according to the law of restitution work in a manner highly analogous to rescission of a contract, in that they recognize and replicate the primary rights of the parties that existed before a specific event. The example of the store whose assistant makes too much change is a case in point. Before the mistake occurred, the surplus was the property of the store. At no time did the shopper and the store engage

---

288 LAYCOCK, supra note 36, at 485.
289 Sheppard v. Koch, 27 S.W.2d 389, 391 (Ky. 1930) (plaintiff sought reformation of a deed because she did not intend to convey title of a roadway); Matthews v. Whitethorn, 77 N.E. 89 (Ill. 1906) (plaintiff sought reformation to correct mutual mistake in description on deed); Archer v. McClure, 81 S.E. 1081, 1083 (N.C. 1914) (plaintiff sought reformation of bond to correct name of intended obligee).
292 See supra Parts IV(A)(1)–(2).
293 LAYCOCK, supra note 36, at 485.
294 See id. at 4.
295 See Birks, supra note 19, at 28.
in any conduct involving a transfer of title to that surplus, nor did either of them commit any legal wrong. The surplus therefore remained the property of the store throughout, and the store can therefore seek a replicative remedy to recover that very surplus in order to replicate its own pre-existing primary right.

The remedy available in such circumstances is typically known as either restitution or quantum meruit (although, for historical reasons, the latter phrase is also often used to describe a common law, substitutionary remedy) and is available because the defendant has been unjustly enriched by the acquisition of money or property belonging to the plaintiff. It is, however, an equitable remedy, and so the assessment of injustice must also take account of the impact that the award of the remedy would have on the defendant. In the case of the shopper with too much change, the store would, for example, be expected to act expeditiously if it wanted to recover the surplus change. An unreasonable delay—known in equity as “laches”—would bar the availability of the remedy.

2. Account for Money Had and Received

Although it is often assumed that the law of restitution is a field of law concerned solely with equitable remedies, the common law recognized the remedy of an account for money had and received centuries before equity developed remedies for unjust enrichment. In fact, the concept of money had and received derives from the ancient writ of indebitatus assumpsit, which translates as “being indebted, he undertook.” As this suggests, the remedy becomes available when a payment has previously been made under the misapprehension that a

---

296 See, e.g., Lai Ling Cheng v. Modansky, 539 N.E.2d 570, 572 (N.Y. 1989) (discussing the remedy of quantum meruit to compensate attorneys for work performed on a client’s case).

297 See Paramount Film Distr. v. New York, 285 N.E.2d 695, 698 (N.Y. 1972) (“Generally, courts will look to see if a benefit has been conferred on the defendant under mistake of fact or law, [or] if the benefit still remains with the defendant . . .”).

298 Id.

299 See Twin States Realty Co. v. Kilpatrick, 26 So.2d 356, 358 (Miss. 1946) (noting that a plaintiff seeking equitable relief for violation of a deed restriction must do so within a reasonable period of time or else laches may bar right to relief).


301 Assumpsit, BLACK’S LAW DICTIONARY (10th ed. 2014).
debt was due. Such a misapprehension might involve mistakenly paying off a debt that did not exist or that has already been paid, making a payment on a contract that has since been rescinded, or paying off a debt by the wrong payer, or to the wrong payee. In any such circumstances, the remedy of an account for money had and received is available to the person whose money it rightfully is, which—in the case of a debt paid to the wrong person—will not be the original payer. The remedy is, as its origins imply, a form of action for debt, where the person erroneously in possession of the money is treated as indebted to its rightful owner.

As in the case of the shopper who does not notice that she has given too much change, no wrongdoing is involved in these cases, and so there is no question of a substitutionary remedy being available. Yet the person in possession of the money is mistakenly in possession: she has no primary right to it. The remedy thus enables the person who does have the primary right of ownership to replicate that right by recovering the money involved.

302 See, e.g., Blue Cross Health Servs., Inc. v. Sauer, 800 S.W.2d 72, 77 (Mo. Ct. App. 1990) (holding that Blue Cross was entitled to restitution for payments mistakenly made for medical bills).
303 See, e.g., Gulf Life Ins. Co. v. Folsom, 349 S.E.2d 368, 369 (Ga. 1986) (holding that a plaintiff can recover payments mistakenly made in case where mistake was plaintiff’s own fault and defendant received two payments from insurance policies as a result of a computer error).
306 See, e.g., Dunham v. Portfolio Recovery Assoc., LLC, 663 F. 3d 997 (8th Cir. 2011).
307 See, e.g., Amoco, 946 S.W.2d at 163.
308 See, e.g., Bank of the Republic v. Millard, 77 U.S. 152, 158 (1870) (plaintiff had the right to recover for funds taken out of bank account when bank cashed a forged check).
309 See, e.g., Amoco, 946 S.W.2d at 165.
310 Roberts v. Ely, 20 N.E. 606 (N.Y. 1889) (holding that plaintiffs may recover portion of insurance proceeds paid to a company that was in possession of their tea); Parsa v. New York, 474 N.E.2d 235, 237–38 (1984) (describing action for account for money had and
The nature of money had and received as a replicative remedy was highlighted by a scenario discussed in many old law reports. If Derek cuts down Patricia’s timber and sells it, Patricia could seek replevin to recover the timber that is rightfully hers. (She could also seek a substitutionary remedy to vindicate her secondary rights because of Derek’s tort.) If that timber is sold, Menzel v. List tells us that Patricia could then instead seek replevin against the third party purchaser in order to recover the timber. But her remedies against Derek would then have changed.

While the substitutionary remedy for the tort would remain available, replevin would no longer be available because Derek would no longer be in possession of the timber. Instead, Patricia could now avail herself of the replicative remedy of money had and received so as to claim the purchase price. This is because the timber remains her property throughout, and so the purchaser has effectively paid the price to the wrong person. In other words, it is Patricia who has the primary right to the price of the timber, because she had the primary right to the timber for which the money was exchanged.

This also means that, if Patricia recovers the price from the tortfeasor, the purchaser will then have obtained primary rights to the timber. Patricia cannot claim primary rights to both the timber and the received and holding that claimant failed to establish that the State was withholding money to which he was entitled.

311 “When the money has been received by the defendant in consequence of some tortious act to the plaintiff’s property, as when he cut down the plaintiff’s timber and sold it, the plaintiff may waive the tort and sue in assumpsit for money had and received.” JOHN BOUVIER, A LAW DICTIONARY, ADAPTED TO THE CONSTITUTION AND LAWS OF THE UNITED STATES OF AMERICA AND THE SEVERAL STATES OF THE AMERICAN UNION; WITH REFERENCES TO THE CIVIL AND OTHER SYSTEMS OF FOREIGN LAW 165 (5th ed. 1855) (citing Eastwick v. Hugg, 1 Dall. 222 (Pa. 1787); Hanna v. Pegg, 1 Blackf. 181 (Ind. 1822); Jones v. Hoar, 5 Pick. 285 (Mass. 1827); Pritchard v. Ford, 1 J.J. Marsh. 543 (Ky. 1829); Whitwell v. Vincent, 4 Pick. 452 (Mass. 1827); Gilmore v. Wilbur, 12 Pick. 120 (Mass. 1831); Bank of North America v. McCall, 4 Binn. 374 (Pa. 1812) (Yeates, J., concurring); Willet v. Willet, 3 Watts 277 (Pa. 1834); King v. McDaniel, 4 Call 451 (Va. 1799)) [Eds. note: Bouvier incorrectly cited “1 Dall. 122” for the quoted proposition; the correct citation is “1 Dall. 222” as stated in the parenthetical.]. See also William A. Keener, Waiver of Tort, 6 HARV. L. REV. 223, 229 (1892) (“When the defendant has converted the plaintiff’s property, and in the act of conversion, or thereafter, sells the same, the plaintiff may waive the tort and sue in assumpsit, using the count for money had and received to recover the proceeds of the sale.”).

312 See infra Part V.

purchase price. Once one is obtained, the other is extinguished, although her secondary rights against Derek will remain unaffected.

In short, as far as the common law is concerned, the price for the timber that Derek obtains is analogous to owing a debt to Patricia. But the analogy only goes so far, because such a debt can arise only once he has actually received the money. An account for money had and received is—like debt itself—a replicative remedy and so cannot be created by Derek’s wrongful act. Instead, Patricia obtains the right to the money only when an innocent buyer mistakenly pays Derek for the timber and so utilizes power-conferring rules to create a contract. As with births, gifts, trusts, taxation, and inheritance, this is an example of a person obtaining primary rights—and the ability to have them vindicated by replicative remedies—as a consequence of others utilizing appropriate legal powers.

D. Trusts

1. Accounting for Profits

An accounting for profits looks superficially similar to an account for money had and received. Indeed, the two remedies are often confused because profits often constitute the “money” involved when pursuing the remedy of an account for money had and received. Yet, while both are replicative remedies, the two are actually quite distinct. In particular, an accounting for profits—unlike an account for money had and received—is an equitable remedy. It applies when a person makes money on a transaction in circumstances where it rightfully belongs to someone else. Such cases typically involve either (a) an agency relationship, where the profits belong not to the agent but to the principal, or (b) a fiduciary duty, such as that of a trustee, where the profits belong to the trust itself (for the benefit of its beneficiaries) rather than to the trustee. In such cases, the person who has earned the

---

314 Amoco Prod. Co. v. Smith, 946 S.W.2d 162, 164 (Tex. Ct. App. 1997) (“This action is not premised on wrongdoing, but . . . inquires whether the defendant has received money which rightfully belongs to another.”) (citing Greer v. White Oak State Bank, 673 S.W.2d 326, 329 (Tex. Ct. App. 1984)).
316 ZAKRZEWSKI, supra note 34, at 186.
318 See Joel Eichengrun, Remedying the Remedy of Accounting, 60 IND. L.J. 463, 468
money effectively holds it in trust for the principal or beneficiary.\textsuperscript{319} Unlike an account for money had and received, an accounting for profits has no requirement that the payer of the money was mistaken as to the amount to be paid or the identity of the correct payee.\textsuperscript{320} In other words, there is nothing erroneous about the transaction in which money changes hands, which is why the common law traditionally saw no need for a remedy in such cases.\textsuperscript{321} It is just that the status of the payee puts him in a position where he is obliged in equity not to make money on his own behalf but to account for it to the principal or beneficiary.\textsuperscript{322} Of course, the fact that this is an equitable remedy means that such a principal or beneficiary must act quickly and fairly to obtain it, otherwise it will be barred by laches, estoppel, or some other equitable defense.\textsuperscript{323}

From the perspective of the plaintiff, therefore, an accounting for profits is usually a significantly inferior remedy in comparison to either an account for money had and received or an award of damages, because the latter remedies are available at common law and so are not as easily denied.\textsuperscript{324} Moreover, as an equitable remedy, an accounting for profits involves a bench trial rather than a jury trial.\textsuperscript{325} These facts make it rather perplexing that, for decades, many plaintiffs have claimed “an accounting for damages” from the courts.\textsuperscript{326} While the Supreme Court in \textit{Dairy Queen, Inc. v. Wood} rightly held that “the constitutional right to trial by jury cannot be made to depend upon the choice of words used in the pleadings,”\textsuperscript{327} it still seems to be an unnecessary hostage to fortune for a plaintiff’s attorney to include language redolent of a more easily resisted equitable remedy when, as in \textit{Dairy Queen} itself, she is entitled

\textsuperscript{109} ZAKRZEWSKI, supra note 34, at 186.

\textsuperscript{109} See, e.g., Babbit Elecs., Inc. v. Dynascan Corp., 38 F.3d 1161, 1182 (11th Cir. 1994) (accounting for profits is proper “[w]here the defendant’s infringement is deliberate and willful . . . under a theory of unjust enrichment” even when the defendant and plaintiff are not in direct competition).

\textsuperscript{169} See, e.g., Eichengrun, supra note 318, at 464 (noting that there was no remedy for feudal landowners to recover from a bailiff without bringing an action in debt, detinue, replevin, or contract).

\textsuperscript{169} ZAKRZEWSKI, supra note 34, at 186.


\textsuperscript{109} See supra note 300, Figure 2.

\textsuperscript{169} Eichengrun, supra note 321, at 470–71.

\textsuperscript{190} See, e.g., Dairy Queen, Inc. v. Wood, 369 U.S. 469, 475 (1962).

\textsuperscript{180} Id. at 477–78.
to maintain a claim for either debt or damages (or both) at common law.328

E. Advantages

It should now be possible to identify the distinctive characteristics of replicative remedies, in addition to the fact that they replicate the very rights that they vindicate.329 These characteristics provide both advantages and limitations.

Clarity and certainty are among the chief advantages. As cases like *Vicars* and *Menzel* illustrate,330 replicative remedies are concerned with restoring to the rightful owner or possessor the “very thing” or the “very debt itself” to which the plaintiff claims a primary right. Replicative remedies, in other words, enable plaintiffs to obtain what is theirs. It follows that plaintiffs are not required to prove that they have suffered any harm or loss. If the money or property is theirs, then that is all that matters. For instance, if someone were to remove a piece of junk from my house, I would be entitled to have it returned to me, even if the theft actually saved me from paying to have someone else remove it. If the thief subsequently sold the item, I would be entitled to claim the money he had and received—simply because it would be mine and not his. (I could not claim both the money and the junk, for that would be double recovery. But I could choose which of those claims to pursue.)

Moreover, since there is no question of the plaintiff having to prove harm or loss, it logically follows that issues of foreseeability of harm or loss are quite irrelevant. Similarly, there is no obligation on the plaintiff to mitigate her loss, because there is simply no loss to mitigate in the first place.331 The remedy to which successful plaintiffs are entitled is the provision of what is theirs, nothing more and nothing less. Accordingly, no question of indeterminacy of loss can arise: the remedy at issue is about as finite and predictable as it is possible for the law to countenance.

These characteristics make adjudication over replicative remedies highly amenable to summary judgment, because they leave (almost) nothing for a jury to do. A debt is typically memorialized in a written

328 Id. at 476–77.
329 See supra Part IV.
330 See supra Part IV(A)(1)–(2).
document, and so too is ownership or possession of real property or of significant items of personal property (like a car or painting). There is then no dispute of fact. This provides the basis for the development of the debt collection industry. It is easy to establish the existence of a debt, and there is no need to prove that it was foreseeable or has been mitigated. This also explains why plaintiffs sometimes choose to pursue replicative remedies even when a wrong has been committed: they are often easier, faster, and less expensive to obtain than a substitutionary remedy.

F. Limitations

The fact that replicative remedies are designed to replicate the primary rights that they vindicate also means, however, that they suffer from several limitations. Most important is this: if the property over which the plaintiff has a primary right of ownership or possession is lost or destroyed, then the right is inevitably lost too. This leaves nothing for any replicative remedy to vindicate. The law will not order an impossibility. Of course, if a court believes that the defendant is fraudulently denying the existence or knowledge of the whereabouts of the property in question, then it can sanction the defendant via proceedings for perjury or contempt. But if the property is genuinely lost or destroyed, then the “very thing” that is the subject of any replicative remedy is also lost, and there is nothing for a replicative remedy to vindicate.

Replicative remedies are thus very tightly defined. Because they are concerned only with the thing or debt itself, they do not provide compensation for harm or losses resulting from a plaintiff’s deprivation of that item or money. (That is the role of substitutionary remedies.) Thus reformation of a contract does not permit the contract to be written more favorably to the plaintiff, specific performance cannot require

---

332 Benjamin E. Hermalin et al., Contract Law, in 1 HANDBOOK OF LAW AND ECONOMICS 51 (A. Mitchell Polinsky & Steven Shavell eds., 2007).
greater performance than that specified in the contract, and actions for debt and for the price do not envision any payments over and above the contract price. Therefore, while the painting that was the subject of *Menzel v. List*\(^{335}\) might have kept its value, or even experienced an appreciation in value greater than the rate of inflation, the Corvette recovered by Atlantic Discount in *D.T. Vicars v. Atlantic Discount Co.*\(^{336}\) would certainly have been worth less than when Mrs. Powell stopped making the finance payments, and yet the replicative remedy of detinue provides no compensation for such loss.

Moreover, replicative remedies are more easily defeated by recourse to equitable defenses than is true of substitutionary remedies.\(^{337}\) A defendant subject to a claim for money had and received or *quantum meruit* may be able to resist successfully on the grounds that he has changed his position to his detriment in reliance on having received the money in good faith, so that the plaintiff is now estopped from seeking recovery.\(^{338}\) Similarly, too great a delay in seeking the remedy may lead to its being barred on equitable grounds even if the relevant statute of limitations has not yet expired.\(^{339}\)

There is a very good reason for such limitations on replicative remedies. What triggers the plaintiff into seeking such a remedy cannot, by definition, ever be a wrong because wrongs trigger secondary rights and the availability of substitutionary remedies.\(^{340}\) They have nothing to do with primary rights and replicative remedies.\(^{341}\) But, if the defendant has not committed a wrong, there can be no justification for applying any sanction beyond what is absolutely necessary to provide redress to the plaintiff.\(^{342}\) As Birks once explained:

When it comes to the law’s response to the facts, there is a crucial difference between a wrong and a not-wrong. The label “wrong” operates as a licence [sic] to the law to mistreat the wrongdoer. The entitlements which the law can

\(^{335}\) See *supra* notes 213–219, and accompanying text.

\(^{336}\) See *supra* notes 220–232, and accompanying text.

\(^{337}\) See e.g., Yorio, *supra*, note 323, at 1201–02 (noting that specific performance claims must overcome a broader array of defenses).

\(^{338}\) Dickerson v. Colgrove, 100 U.S. 578, 580 (1880) (determining whether defendant relied on promise that plaintiff would not claim the farm property in question).

\(^{339}\) Galliher v. Cadwell, 145 U.S. 368, 372 (1892) (discussing laches, giving the defendant a reasonable belief that the plaintiff’s rights are now worthless).

\(^{340}\) Birks, *supra* note 19, at 12.

\(^{341}\) Id. at 13.

\(^{342}\) Id. at 33–34.
accord to the victim of the wrong are, so to say, at large, at least until the range has been narrowed by authority, statutory or otherwise. . . . Those entitlements can be chosen subject only to the constraints imposed by extrinsic considerations, utility, humanity, proportionality, and so on. . . . By contrast, not-wrongs leave very little room for choice, because they offer no general licence [sic] to mistreat the defendant. A claimant who puts in issue an unjust enrichment, such as a mistaken payment, makes no case for, say, making the defendant make good his consequential loss. His facts are weak, they justify nothing but getting that much value back. There are some choices to be made, but only within a very narrow compass. A taxable event is the same. It justifies the demand for the tax, nothing more. 343

V. SUBSTITUTIONARY REMEDIES

Substitutionary remedies are designed to vindicate secondary rights on occasions when the defendant has committed a wrong, thereby damaging one of the victim’s interests. 344 In other words, these remedies provide the sanctions that a court may impose upon the defendant because of his or her wrongful conduct. They do so by supplying the plaintiff with something else (typically money) to substitute for the loss of value to the interest harmed. 345

One point worth making at this juncture is that the ability to sanction the defendant for his or her wrong is one thing, but the degree or size of that sanction is quite another. A plaintiff may have many reasons for wanting to sanction a defendant in any given case, and the courts may permit such sanctioning to occur for a plethora of reasons that might, or might not, coincide with those of the particular plaintiff. But, whatever the motivation behind the desire to sanction a defendant, it merely furnishes a policy justification for the award of a substitutionary remedy. It has no bearing on the level of the sanction imposed, which is always set by either (a) the loss of value to the interest involved, or (b) the cost of repair to the interest involved. 346

As with replicative remedies, it is simply not possible in the space available to discuss every substitutionary remedy. It is worth, however, exploring some of the more important substitutionary remedies to show, once again, how a sound, rights-based taxonomy aids in clarifying what

343 Id. at 33.
345 1 DAN B. DOBBS, LAW OF REMEDIES 305–08 (2d ed. 1993).
346 Id.
has become an unnecessarily murky area of the law.

A. Compensatory Damages in Contracts

It is a truism to say that, because compensatory damages are a substitutionary remedy designed to vindicate secondary rights, and secondary rights exist to sanction a wrong, compensatory damages in contracts must therefore be designed to provide a sanction for a breach of contract. A breach of contract infringes upon the interests at stake in the contract, so the size of the sum awarded as compensatory damages must be the monetary equivalent of the loss in value of those interests.

This loss in value may, however, be mitigated—or even avoided altogether—if the innocent party obtains a replicative remedy—like debt, an action for the price, or specific performance—to vindicate her primary rights. If there is then no resultant loss in the value of the relevant interests, then no compensatory damages will be available. If, however, the replicative remedy merely mitigates the loss in value of the relevant interests, then the monetary equivalent of the loss that remains will still be available as compensatory damages. This is why the replicative remedy of debt is typically accompanied by a payment of interest. Interest is simply an example of the substitutionary remedy of compensatory damages. It is awarded to reflect the fact that, for a period of time, the debtor was responsible for the creditor’s being wrongly deprived of possession of the specified sum of money. Alternatively, a plaintiff may choose to seek only the substitutionary remedy of

347 Id. at 281.
348 DOBBS, supra note 345, at 281–83.
349 See Peter Benson, Contracts as a Transfer of Ownership, 48 WM. & MARY L. REV. 1673, 1674 (2007) (“[A breach may be] remedied through damages or specific performance, both of which aim to put the plaintiff in the position he or she would have been in had the defendant performed as promised”); see also Van Wagner Adv. v. S. & M. Enters., 492 N.E.2d 756, 759–60 (N.Y. 1986) (stating that “specific performance is available, in appropriate circumstances,” but “at some level all property may be interchangeable with money”—the issue coming down to “uncertainty in valuing it.”); see also DOBBS, supra note 345, at 281 (“[T]he plaintiff should be fully indemnified for his loss, but he should not recover any windfall. . . . Courts may use noncompensatory measures or procedures to allow an award . . . . the compensatory purpose may be outweighed by other considerations, which may lead courts to award less than fully compensatory damages.”).
350 The finder of fact makes this determination. LAYCOCK, supra note 36, at 5.
351 See Bruning v. United States, 376 U.S. 358, 360 (1964) (holding that interest on debt should be recoverable in a later action against the debtor personally”).
352 Id.
compensatory damages. Thus a plaintiff in a breach of contract action will often have the option of whether to seek a replicative remedy, a substitutionary remedy, or both. 353

1. The Three Interests

We have already noted that the amount of any award of compensatory damages typically reflects the loss of value in a relevant interest that was caused by the defendant’s wrongdoing. Within the law of contracts, and since the publication of a famous article by Lon Fuller and William Perdue, 354 it has generally been recognized that there are three such interests at stake. 355 As is well known, Fuller and Perdue named them the expectation, reliance, and restitution interests. 356 They asserted that “‘restitution interest’ . . . presents the strongest case for relief,” 357 but the law has never adopted that approach. Instead, the law has always strongly favored the expectation interest, 358 although the courts themselves have failed to provide any meaningful explanation for this preference. The taxonomy of remedies presented here, however, provides that missing explanation.

What distinguishes the law of contracts from other fields of law is that it has both a “formation phase” (embodied in the doctrines of offer, acceptance, and consideration or, in some cases, promissory estoppel) and a “performance phase.” 359 This is unique among the fields of the common law (although it is also a feature of the law of trusts in equity). The formation phase is where the powers conferred by the law of contracts may be utilized to enable the parties to create new primary rights for themselves. 360 The interest created by these rights is the

353 Giron v. Hous. Auth., 393 So. 2d 1267, 1269, 1272 (La. 1981) (stating that plaintiff in breach of contract action is “entitled either to damages or specific enforcement of the contract or to a dissolution of the contract,” apparently at plaintiff’s election).
355 Id. at 53–54; see also RESTATEMENT (SECOND) OF CONTRACTS § 344, cmt. a (AM. LAW INST. 1981).
357 Id. at 56.
359 Kaye, supra note 143, at 293.
360 See Benson, supra note 349, at 1674 (“The contractual rights and duties between the parties are completely specified and determined at formation. Performance adds nothing new but is just the fulfillment of the rights and duties established at contract formation.”).
expectation interest.\textsuperscript{361} It is therefore inevitable that, if the contract is subsequently breached, contract law will give pre-eminence to vindicating the rights that support the expectation interest.\textsuperscript{362}

The other interests—whether restitution or reliance—are not created by the contract itself, and so provide only a “second-best” method of assessing compensatory damages.\textsuperscript{363} In fact, these interests must have existed before the contract was formed.\textsuperscript{364} Whether involving money paid or goods or services rendered in reliance on the contract, the interests will have been acquired through a prior contract, inheritance, gift, or trust. This makes the value of such interests important to assessing compensatory damages only when an appraisal of the value of the lost expectation interest would be too speculative.\textsuperscript{365}

The law’s preferred focus on the value of the expectation interest also explains the significance of what is known as the two rules in \textit{Hadley v. Baxendale}.\textsuperscript{366} Where “special circumstances”\textsuperscript{367} are known to both parties at the time that the contract was made (\textit{Hadley}’s second rule), then they should be taken into account in determining the value of the expectation interest.\textsuperscript{368} But if the party in breach had no reason to know of such circumstances, then the value of the expectation interest should be assessed as if “occurring under ordinary circumstances”\textsuperscript{369} (\textit{Hadley}’s first rule).

\textbf{B. Liquidated Damages}

Liquidated damages are a form of self-help remedy, where,

\begin{footnotesize}
\begin{itemize}
\item\textsuperscript{361} Fuller & Perdue, supra note 354, at 54.
\item\textsuperscript{362} Benson, supra note 349, at 1674–75.
\item\textsuperscript{363} See \textit{Restatement (Third) of Restitution & Unjust Enrichment} § 38, § 38 cmt. a (AM. LAW INST. 2011) (allowing for an award of damages that “restore the plaintiff to the precontractual position, or . . . restore the plaintiff either the cost or the value of the plaintiff’s uncompensated performance. . . . Either approach to damages under [§ 38] yields a second-best, ‘fall-back’ alternative to damages measured by expectancy.”).
\item\textsuperscript{364} Benson, supra note 349, at 1707.
\item\textsuperscript{365} Id. See also \textit{Restatement (Third) of Restitution & Unjust Enrichment} § 38 cmt. a (AM. LAW INST. 2011) (“[Restitution] is distinguished from ordinary contract damages . . . because it permits a plaintiff who cannot prove expectation to recover damages calculated on an alternative basis. . . . [Restitution] offers a remedial alternative in cases where the plaintiff cannot establish expectation damages—either because of difficulties or proof, or because contractual expectancy is provable but negative.”).
\item\textsuperscript{367} Id. at 151.
\item\textsuperscript{368} Id.
\item\textsuperscript{369} Id.
\end{itemize}
\end{footnotesize}
traditionally, the parties specify at the time of contracting the amount of compensation to be paid in the event that either party fails to adhere to the terms of the contract.\textsuperscript{[370]} As a self-help remedy, liquidated damages are subject to the residual oversight of the courts, which will refuse to uphold the payment of such compensation—on the grounds that it amounts to a penalty—if they find that it over-estimates the true value of the loss that the innocent party sustained.\textsuperscript{[371]}

The label is not, however, definitive.\textsuperscript{[372]} What, at first glance, might appear to be a liquidated damages clause may turn out, on closer inspection, to be something rather different. The issue arises because a liquidated damages clause necessarily details a specific sum of money that is to be paid, which is also characteristic of the replicative remedy of debt.\textsuperscript{[373]} (This is why liquidated damages occupies two different locations in Figure 2.) Deciding which is the true characterization of the clause can have a significant bearing on the outcome of a case.

\textit{Lake Ridge Academy v. Carney,}\textsuperscript{[374]} decided in the Supreme Court of Ohio, provides an excellent example of this issue. Carney had reserved a seat for his son in the fourth grade class at Lake Ridge in exchange for a $630 deposit and a promise to pay the balance of the tuition later that year.\textsuperscript{[375]} The contract gave Carney the option to cancel the agreement and withdraw his son without having to pay the rest of the tuition if he did so before August 1st.\textsuperscript{[376]} But he exercised this option too late, and the school demanded the full tuition.\textsuperscript{[377]} In his dissent, Justice Pfeifer (with whom Justice Resnick concurred), argued that the school was seeking to obtain an unenforceable penalty because it was “unquestionable that the parties knew that [the actual] damages would total an amount less than

\textsuperscript{[370]} \textit{Restatement (Second) of Contracts} §356 cmt. a & b (Am. Law Inst. 1981).
\textsuperscript{[371]} \textit{Id.}; \textit{see also} Equitable Lumber Corp. v. IPA Land Dev. Corp., 344 N.E.2d 391, 397 (N.Y. 1976) (holding that, if the amount in the liquidated damages clause “is found to be unreasonably large, then the provision is void as a penalty”).
\textsuperscript{[372]} \textit{See, e.g.,} Truck Rent-A-Ctr. v. Puritan Farms, 361 N.E.2d 1015, 1017, 1019–20 (N.Y. 1977) (holding that despite the plaintiff’s claim that the clause did not represent “a fair estimate of actual damages,” the liquidated damages clause contained in a contract was not considered a penalty and was therefore valid).
\textsuperscript{[373]} \textit{See} Douglas Laycock, \textit{The Death of the Irreparable Injury Rule}, 103 Harv. L. Rev. 687, 698 (1990) (“In suits to collect debts, a fixed sum of money is the specific thing plaintiff lost.”).
\textsuperscript{[374]} Lake Ridge Acad. v. Carney, 613 N.E.2d 183 (Ohio 1993).
\textsuperscript{[375]} \textit{Id.} at 184.
\textsuperscript{[376]} \textit{Id.} at 185.
\textsuperscript{[377]} \textit{Id.}
the full annual tuition.”378

That would seem to be an undeniable statement of legal orthodoxy. Writing for the majority, however, Justice Wright took a different tack. He argued that Carney had entered into a contract that gave him the option either to repudiate the agreement by August 1st or else to pay the full tuition.379 In other words, Justice Wright saw the liquidated damages clause not as providing a substitute for performance of the contract, but as one of two, equally legitimate, ways of performing the contract.380 Whether that characterization was correct is not our concern here. What matters is that the school’s claim necessarily changes the nature of the remedy sought. If payment of the liquidated damages is a valid method of performing the contract, it follows that the school has a primary, and not a secondary, right to that money via the replicative remedy of debt. If the school was owed a debt, then it was clearly entitled to the sum specified because that money was rightfully its own. It is irrelevant whether the money was representative of the school’s true loss, since debt is not concerned with loss, but with the very sum of money itself. So, if the majority’s characterization of the liquidated damages clause was correct, then the school’s claim was justifiably upheld.

C. Compensatory Damages in Torts

While compensatory damages in contracts are primarily designed to substitute for the loss of value to the expectation interest created by the new primary rights,381 the absence of an equivalent “formation phase”382 in the law of torts means that, as Figure 1 illustrates, tort law cannot create new primary rights. Compensatory damages in torts must, therefore, focus on providing a substitute for the loss in value of a quite different kind of interest.383 That interest must either have been previously created in a different field of law (typically property or

379 Id. at 186.
380 Id. at 187 (“As part of the contract in this case, Carney was given the option to cancel. That option expired, by the express terms of the contract, on August 1. As of August 1 any attempt by Carney to exercise his option to cancel—including his letter postmarked August 7—was ineffective.”).
382 Kaye, supra note 143, at 293.
383 Id; see also Kaye, supra note 172, at 945 (“There is no question of seeking ‘substitutive’ damages because no right has been lost for which they can substitute; the plaintiff simply seeks compensation for degree of intrusion or harm suffered.”).
constitutional law) or else be afforded legal recognition only within the field of tort law, such as an interest in consortium with a decedent victim.

It is important to emphasize that compensatory damages in torts do not become payable merely as a sanction for the violation of a secondary right.\(^{384}\) Such an approach would demand that the jury address existential questions such as the value of the rights to "life, liberty, and the pursuit of happiness."\(^{385}\) The common law has always been much more practical than that. It has also been consistent in its approach. Just as compensatory damages in contracts become payable only where a violation of contractual rights has caused loss of (or harm to) an interest protected by the law of contracts,\(^{386}\) compensatory damages in torts become available only if, in addition to the breach of a secondary right, the victim has suffered some loss of (or harm to) an interest protected by the law of torts.\(^{387}\)

In a wrongful death suit, for example, compensatory damages are not sought to compensate the decedent for his or her loss of the right to life.\(^{388}\) If that were true, they would be payable to the decedent’s estate to be passed on by will or the rules of intestacy. In fact, however, the law recognizes the decedent’s dependents as appropriate plaintiffs in a wrongful death suit.\(^{389}\) In other words, such a suit is brought by those who had a financial interest in the decedent’s right to life, and the level of compensation they receive is determined by the amount of financial support that they have lost as a result of the decedent’s demise.\(^{390}\)

\(^{384}\) See Kaye, supra note 172, at 943 (criticizing ROBERT STEVENS, TORTS AND RIGHTS 2 (2007)).

\(^{385}\) Id. at 945–46 ("If damages were, in a significant number of cases, really about compensating the plaintiff for violation of a legally-protected right *per se* rather than about the amount of harm caused thereby, then surely this would be a matter of law to be addressed by the judge rather than a matter of fact for a jury. That would, after all, be the only way of ensuring that each violation led to the same award regardless not only of the identity or status of either party but also of any other circumstances. Yet, in the United States, the assessment of damages is still left to the jury.").

\(^{386}\) Benson, supra note 349, at 1674.

\(^{387}\) RESTATEMENT (SECOND) OF TORTS § 903 cmt. a (AM. LAW INST. 1979).

\(^{388}\) See id. at § 925 cmt. a ("[I]f no action could have been brought by the deceased if still alive, no right of action exists.").

\(^{389}\) Kaye, supra note 172, at 945; see also, e.g., Levy v. Louisiana, 391 U.S. 68, 72 (1968) (considering children—despite being illegitimate—as appropriate plaintiffs because of dependency on their mother).

\(^{390}\) Levy, 391 U.S. at 68 n.1 (quoting the Louisiana statute that was at issue in the case, LA. CIV. CODE ANN. art. 2315 (1967)); see also RESTATEMENT (SECOND) OF TORTS § 925 cmt. a (AM. LAW INST. 1979).
Similarly, while the nature of the interests protected by secondary rights in torts is typically very different from those protected by secondary rights in contracts, the method of assessing the value of those interests is identical. Damages are set either as the value of a lost interest or as the cost of repairing a damaged interest so that it is restored to its previous value.391 Wrongful death suits are examples of the former, as are cases where the victim has suffered a loss of income392 or life expectancy.393 Another example is the total loss of property, whether because of destruction or deemed so by loss adjuster determination.394 Lost business is another.395

Property damage and personal injuries (e.g., a broken limb) that can be successfully treated, are examples of where the cost of repair sets the sum of compensatory damages to be awarded. The cost of repairing damaged property396 or obtaining appropriate medical treatment,397 will then determine the amount involved. For those suffering emotional distress, the cost of mental health care may be used in an analogous fashion to determine the level of harm caused to the victim’s interest in the pursuit of happiness (or emotional wellbeing) and thus set the

391 See, e.g., RESTATEMENT (SECOND) OF TORTS § 929 (AM. LAW INST. 1979) (entitling one whose land has been harmed to damages compensating “the difference between the value of the land before the harm and the value after the harm, or . . . the cost of restoration that has been or may be reasonably incurred”); see also, e.g., Lampi v. Speed, 261 P. 3d 1000, 1004 (Mont. 2011) (discussing remedies for property damage cases).

392 See, e.g., Bigelow v. RKO Radio Pictures, Inc., 327 U.S. 251, 264 (1946) (“[T]he jury may make a just and reasonable estimate of the damage based on relevant data [including lost earnings as a result of the alleged harm], and render its verdict accordingly.”).

393 See, e.g., Alexander v. Scheid, 726 N.E.2d 272, 282–83 (Ind. 2000) (stating that, even though not easily evaluated, in cases where a plaintiff has suffered a decrease in life expectancy, “the jury will have to attach a monetary amount” to the loss, thereby “consider[ing] what value to ascribe to the privilege of living”).


395 See, e.g., J’Aire Corp. v. Gregory, 598 P.2d 60, 65–66 (Cal. 1979) (allowing for recovery of a lost economic advantage, such as a loss of expected business, where there is “foreseeability of the injury” and a “nexus between the defendant’s conduct and the plaintiff’s injury”); see also, e.g., Curd v. Mosaic Fertilizer, 39 So.3d 1216, 1220–22 (Fla. 2010) (holding that commercial fishermen may recover damages in accordance with state law for loss of business resulting from defendant’s negligent discharge of pollutants into a bay).

396 See Fisher, 779 N.E.2d at 182 (explaining that replacement cost and diminution of value are each “a proper way to measure lost property value, the lower of the two figures affording full compensation to the owner”).

397 See, e.g., Wills v. Foster, 892 N.E.2d 1018, 1033 (Ill. 2008) (“All plaintiffs are entitled to seek to recover the full reasonable value of their medical expenses.”).
appropriate sum of compensatory damages. The law of remedies often struggles, however, when the victim suffers pain and suffering. In principle, like physical and psychiatric injuries, the cost of appropriate diagnosis and medication should be capable of measuring the level of harm sustained to the happiness interest. But pain and suffering is something so subjective that, unlike physical and psychiatric injuries, it is often impossible to identify an objectively appropriate course of treatment. This leaves awards of compensatory damages for pain and suffering either as guesses, as amounts fixed by the policy limits in first-party or liability insurance policies, or as a proxy for the payment of a successful plaintiff’s attorneys’ fees. It is surely no coincidence that the so-called “tort reform” movement reserves special ire for awards of damages for pain and suffering. The law has yet to formulate a consistent metric for determining the level of such harm sustained by a victim.

D. “Accounting for Damages”

In Wisconsin Alumni Research Foundation v. Apple, Inc., the plaintiff, Wisconsin Alumni Research Foundation ("WARF"), claimed that Apple’s processors (used in tablets and smartphones) infringed one of WARF’s patents (the “752 patent”), violating 35

---

399 Personal Injury Protection (“PIP”) insurance is a common example. See Jeffrey O’Connell et al., A Federal Bill, with Commentary, to Allow Choice in Auto Insurance, 7 CONN. INS. L.J. 511, 513 (2001).
400 See Nielsen v. Spencer, 196 P. 3d 616, 626 (Utah Ct. App. 2008) (stating that “the jury may have used the amount of prior attorney fees as a rough proxy for the amount of emotional distress or mental pain suffered”).
401 Victor E. Schwartz & Leah Lorber, Twisting the Purpose of Pain and Suffering Awards: Turning Compensation into Punishment, 54 S.C. L. REV. 47, 49 (2002) (raising concern about a plaintiff’s attorney manipulating the system to increase pain and suffering awards); Paul V. Niemeyer, Awards for Pain and Suffering: The Irrational Centerpiece of Our Tort System, 90 VA. L. REV. 1401, 1401 (2004) (arguing that there is no predictable basis for measuring damages for pain and suffering).
403 WARF is the designated patent management organization of the University of Wisconsin-Madison, with a mission to support research and scholarship at the University. About Us, WIS. ALUMNI RES. FOUND., http://www.warf.org/about-us/about-us.cmsx (last visited Nov. 7, 2017).
U.S.C. § 271. On that basis, WARF claimed several remedies, including “[a]n accounting for damages resulting from [Apple]’s infringement of the ’752 patent.” Yet claiming “an accounting for damages” makes literally no sense. In fact, such a pleading seems to confuse three different remedies: 

1. An accounting for profits;
2. An account for money had and received; and
3. Damages.

While an accounting for profits and an account for money had and received are both replicative remedies, damages is a substitutionary remedy. In addition, both damages and an account for money had and received are common law remedies, whereas an accounting for profits is an equitable remedy. So what did the WARF attorneys really mean? The first clue lies in the fact that the complaint ends by “request[ing] a trial by jury on all issues.” Clearly, therefore, they did not mean to ask for an equitable remedy, for that would mean a bench trial. This rules out the possibility that they were seeking an accounting for profits. This would make sense, since Apple acted neither as WARF’s agent nor as its trustee when it exploited the relevant technology.

The question that remains, then, is whether WARF was seeking damages or money had and received. Since they are both common law remedies, a trial by jury would be appropriate in either case, and this did indeed take place. Yet choosing between these remedies is crucial in determining the size of the monetary award that the court should order. A claim for damages is substitutionary, and would mean assessing the value of WARF’s interest in its patent and then working out how much of that value had been lost as a result of Apple’s conduct. It appears, however, that WARF made no attempt to monetize the technology that it had patented. Nor had it any concrete plans to do so. Left to its own devices, then, it seems WARF would have made nothing of its patent, which means that its interest was effectively worthless. A claim for

---

407 See LAYCOCK, supra note 36, at 5.
408 See supra Parts IV(C)(2) & IV(D)(2).
409 Pl.’s Compl., supra note 406, at ¶ 28.
410 Id. at ¶ 21.
damages would therefore result in an award of nothing more than a nominal sum.

An account for money had and received, however, would lead to a very different result. In fact, the situation would be similar to the hypothetical discussed earlier in which Derek tortiously cut down Patricia’s timber and then sold it. We established then that Patricia would be able to claim the purchase price received by Derek because it was her timber. Similarly, any money that Apple makes from its use of WARF’s patented technology rightfully belongs to WARF, who could then obtain “that very sum” via the remedy of an account for money had and received. So, far from demanding “an accounting for damages,” WARF’s attorneys should really have claimed an account for money had and received.

At trial, the jury awarded WARF $234 million, which the court styled “damages.” From the foregoing analysis, it seems likely that the award itself is sound, even if the label applied to the sum awarded is not. Apple has appealed, so it remains to be seen what the Federal Circuit will make of the faulty pleadings. But even if the result turns out correctly in this case, it seems unlikely that such mistakes can continue without the occurrence of serious injustice at some stage. A sound understanding of the taxonomy of remedies cannot come soon enough.

E. Advantages and Limitations

The main advantage of substitutionary remedies over replicative remedies is that they cannot be lost. It is always possible, as a matter of principle, to vindicate a secondary right, even if the primary right has been lost, because a secondary right is a right to sanction a wrongdoer (or whomever now stands in the wrongdoer’s place). Whether the interest at stake has now been lost or destroyed makes no difference to that. Such wrongdoing also provides the plaintiff with, in Birks’s words, greater license to mistreat the defendant. In other words, the jury is given a large measure of discretion in deciding how to value the

---

412 WARF argued for treble damages of this award, but it was denied because WARF failed to meet its burden of proving willful infringement. Id. at *1, *9–11.
413 See supra note 402.
414 AUSTIN, supra note 37, at 43.
415 Birks, supra note 19, at 14.
416 Id. at 33.
interest that was lost or damaged\textsuperscript{417} so that, for example, “a valid award for breach of contract does not require exact correspondence with the particular benefits the injured party would have received had the contract been fully performed.”\textsuperscript{418} In these senses, therefore, substitutionary remedies obviously provide significant advantages over replicative remedies.

But there are several concomitant disadvantages. One is that the plaintiff must prove the interest lost or damaged had some prior value.\textsuperscript{419} (If not, there is no need to provide something else as a substitute for it.) A claim for a replicative remedy, by contrast, requires no proof either of prior value or of loss sustained.\textsuperscript{420}

Another disadvantage is that the plaintiff must prove the defendant’s act or omission was wrong.\textsuperscript{421} This requirement not only imposes an additional burden on the plaintiff, but it also means that a claim for a substitutionary remedy must be targeted at the appropriate defendant. Secondary rights are not valid against the whole world, and so substitutionary remedies can be pursued only against the wrongdoer or someone vicariously liable for the wrong.\textsuperscript{422}

Finally, while a replicative remedy involves repossessing the very debt or item itself, and so can always be exercised against the person currently in possession, it is not always true that the defendant subject to a substitutionary remedy will have the means to satisfy it. Liability insurance makes this less of a problem, of course, but there is no guarantee that the defendant will carry insurance, or that the policy limits will be sufficient to meet the remedy awarded. The plaintiff may seek garnishment of wages or a lien on the defendant’s property, but these remedies work only where the defendant has an income or property and, even then, they often mean that the plaintiff will have to wait a long time for full satisfaction.

\textsuperscript{417}See Kaye supra, note 172, at 946.
\textsuperscript{418}Advanced Micro Devices, Inc. v. Intel Corp., 885 P.2d 994, 1009 (Cal. 1994).
\textsuperscript{419}See DOBBS, supra note 345, at 308 n.3 (providing “[f]or example, the plaintiff may recover only diminished market value when his pet is negligently damaged, and if the pet has not market value or its market value was not reduced, the plaintiff may be denied recover [sic] altogether.”).
\textsuperscript{420}See LAYCOCK, supra note 36, at 5; see also supra Part IV(F).
\textsuperscript{421}See Birks, supra note 19, at 34 (“To reach a claim for consequential loss, or to justify inflicting other kinds of unpleasanctness on the other, the plaintiff must be able to reconceptualize the matter as a complaint of a wrong and show that the law makes that wrong actionable as such.”).
\textsuperscript{422}Id. at 12; see Vázquez, supra note 132, at 1089–90.
There is, of course, nothing to prevent a plaintiff from claiming both a replicative and a substitutionary remedy, provided that this does not effectively mean double recovery. Thus someone deprived of a sum of money due may claim both debt—to recover that very sum—and damages (i.e., interest) to compensate for the unjust detention of the debt.

VI. TRANSFORMATIVE REMEDIES

Transformative remedies are quite different from both replicative and substitutionary remedies. Replicative remedies vindicate primary rights; substitutionary remedies vindicate secondary rights. But transformative remedies do not vindicate tertiary rights, because such rights cannot exist without a court order to create them. (This means, of course, that there can be no such thing as a tertiary self-help remedy.) Instead, transformative remedies create a special kind of tertiary right that does not simply reflect a prior primary or secondary right, but puts the plaintiff in an entirely new position. The plaintiff’s prior (primary and secondary) rights offer no analog for this new state because transformative remedies “do not proclaim the existence of a relationship . . . but create a new one.” For this reason, they have sometimes previously been referred to as “constitutive” remedies.

Transformative remedies are not typically combined with replicative or substitutionary remedies. It will generally make little sense to try to create a new status through a transformative remedy at the same time as replicating a pre-existing right or substituting for a pre-existing interest. Moreover, since transformative remedies are not issued to vindicate prior rights, it follows that the source of their availability either

423 See Dobbs, supra note 345, at 38–39 (“[B]oth general and special damages may be recovered so long as the two measurements do not duplicate elements of the recovery and so long as the special damages do not run aground on the limitations imposed on them.”).
424 Pollock & Maitland, supra note 253, at 225.
425 Zakrzewski, supra note 34, at 45–47.
426 Id. at 203 (“A transformative remedy is a legal relation that significantly differs from any legal relation that existed before the court order was made.”); see also F. H. Lawson, Remedies of English Law 274–75 (O. Kahn-Freund & K. W. Wedderburn eds., 1972).
427 Lawson, supra note 426, at 274 (quoting Itzhak Zamir, The Declaratory Judgment 1 (1962)).
428 Lawson, id., refers to “constitutive remedies,” which according to Zakrzewski, is a subdivision of a broader category of “transformative remedies,” whereby “constitutive remedies are non-enforceable without further court proceedings.” Zakrzewski, supra note 34, at 204.
involves the exercise of discretion or is simply mandated by statute. 429

Divorce is an example of a transformative remedy where the discretion lies in the hands of a judge. 430 Bankruptcy protection, on the other hand, involves both judicial discretion (in approving an exit plan) 431 and a mandatory provision (in the form of the automatic stay imposed when a petition is filed for bankruptcy protection). 432 Little more need be said here about transformative remedies, because they tend to be highly particularized according to the field of substantive law in which they apply.

VII. BOWEN V. MASSACHUSETTS

It is now time to return to the case with which we began, namely Bowen v. Massachusetts. 433 With our new taxonomy of remedies in place, it soon becomes clear that this should have been a very simple case to decide.

The facts were, indeed, simple. In 1965, Congress authorized the Medicaid program 434 as “a cooperative endeavor in which the Federal Government provides financial assistance to participating States to aid them in furnishing health care to needy persons.” 435 The financing of the program works in the following fashion:

Although the federal contribution to a State’s Medicaid program is referred to as a “reimbursement,” the stream of revenue is actually a series of huge quarterly advance payments that are based on the State’s estimate of its anticipated future expenditures. The estimates are periodically adjusted to reflect actual experience. Overpayments may be withheld from future advances or, in the event of a dispute over a disallowance, may be retained by the State at its option pending resolution of the dispute.

. . . . Should the Secretary subsequently conclude that his initial determination was incorrect, the statute provides that “he shall certify restitution forthwith in a lump sum of any funds incorrectly withheld or

---

429 See, e.g., ZAKRZEWSKI, supra note 34, at 205–07.
430 See Marsha Garrison, How Do Judges Decide Divorce Cases? An Empirical Analysis of Discretionary Decision Making, 74 N.C.L. Rev. 401, 403 (1996); see also ZAKRZEWSKI, supra note 34, at 207–08, 210 (stating that constitutive remedies are common in family and divorce law and the “remedies are granted in the exercise of a high degree of discretion”).
otherwise denied."436

Bowen itself concerned an amount of Massachusetts’ expenditures, which had been disallowed as outside the scope of the program.437 Massachusetts filed suit to recover the expenditure and the federal government resisted on the grounds that the claim was barred by the Administrative Procedure Act’s (“APA’s”) prohibition on claims against the federal government for “money damages.” 438

The terminology of “money damages” was, admittedly, a slight complicating factor.439 As Justice Scalia pointed out, this involves “something of a redundancy.”440 The correct term is simply “damages,” which are, of course, always calculated in money. But this redundancy is pernicious: while all damages take the form of money, not all court ordered monetary awards are damages.441

As we have seen, damages are a substitutionary remedy awarded as a sanction for a wrong. But there was no question of wrongdoing in Bowen. No act or omission was fraudulent, negligent, or otherwise culpable. The only question concerned who should bear the burden of certain expenditures. No substitutionary remedy was, therefore, appropriate, so Massachusetts’ claim could not possibly be for damages. That was the mistake made by the dissent.

The type of remedy sought by Massachusetts was undoubtedly replicative. This much the majority got right. Unfortunately, however, they then indulged in misleading terminology of their own—and proceeded to confuse themselves. Instead of referring to the remedy involved as “replicative,” they chose instead to call it “specific.”442 Professor Laycock has adopted the same label,443 which has seen usage in other courts from time to time. But because it is so redolent of the

436 Bowen, 487 U.S. at 883–85 (footnote omitted) (quoting 42 U.S.C. § 1316(c)).
437 Id. at 885–87.
438 Id. at 887, 891; see also 5 U.S.C. § 702 (2012) (“An action in a court of the United States seeking relief other than money damages and stating a claim that an agency or an officer or employee thereof acted or failed to act in an official capacity or under color of legal authority shall not be dismissed nor relief therein be denied on the ground that it is against the United States or that the United States is an indispensable party.”).
440 Id. at 914 (Scalia, J., dissenting).
441 See generally Colleen P. Murphy, Money as a “Specific” Remedy, 58 ALA. L. REV. 119 (2006) (noting that specific relief can also take the form of a monetary award).
442 Bowen, 487 U.S. at 893, 895.
443 See LAYCOCK, supra note 36, at 5.
remedy of “specific performance,” its use tends to confuse the unsuspecting lawyer into thinking that an equitable remedy must be involved. That was the mistake made by the majority.444

In fact, almost since its inception, the common law has made available the replicative remedy Massachusetts sought: the repayment of debt.445 The Social Security Act had effectively created a system of power-conferring rules analogous to those in the common law of contracts, and Massachusetts had followed those rules when incurring certain expenditure. Since the same rules provided that such expenditure would be defrayed by the federal government,446 it followed that the latter owed a debt to reimburse the state. The fact that the debt arose outside the law of contracts was irrelevant. As we have seen, the remedy of debt actually pre-dates the creation of contract law by several centuries.447 All that is required is that the plaintiff has a primary right to a specific sum of money.448 That was undoubtedly the case in Bowen: the suit thus effectively amounted to nothing more exotic than Massachusetts attempting to recover its own money.

VIII. CONCLUSION

Replicative remedies exist to vindicate primary rights. We call them “replicative” because they are designed to replicate the very right that they vindicate. They may be maintained against the whole world and require no proof of harm or wrongdoing.

Substitutionary remedies exist to vindicate secondary rights. But secondary rights differ from primary rights in that they are not themselves worthy of legal protection. Instead, they protect legally-recognized interests. When those interests are harmed by someone else’s wrongdoing, substitutionary remedies become available to sanction the wrongdoer by providing to the plaintiff a substitute (typically money) for the interest that was violated.

Transformative remedies are quite different. They do not look back at the nature of the rights or interests that require protection, but instead focus on the future with the object of putting the plaintiff in an entirely new position for which the plaintiff’s prior rights offer no analog. Unlike

---

444 See Bowen, 487 U.S. at 893, 895, 900.
445 SIMPSON, supra note 252, at 53–54.
446 Bowen, 487 U.S. at 883 (citing Harris v. McRae, 448 U.S. 297, 308 (1980)).
447 See supra Part IV(B)(1).
448 Id.
replicative and substitutionary remedies, which reflect prior rights, transformative remedies thus generate entirely new, tertiary rights.

This is the foundation of a sound taxonomy of remedies. As we have seen, it avoids confusing labels and helps to unpack problems caused by ambiguous terminology. It can both explain the outcomes of past cases and point the way to solutions to new controversies. It is surely time that this taxonomy became more widely articulated and understood.