Medical Professionalism

crossing a generational divide

Colin Walsh and Herbert T. Abelson

Abstract A comprehensive discussion of professionalism in medicine must include its impact on successive generations of physicians. Fifty years ago, doctors acting professionally emphasized medicine as a calling and an ability to act as the authority for patients in crisis at home and in hospitals. Therapeutic options were limited relative to the modern era, and the laying on of hands was practiced as science and art. Today, doctors balance increasing demands on time and efficiency with the sense of primacy of patient care. Technological innovation and patients’ increasing access to medical knowledge through varying media of inconsistent quality challenge physicians in novel ways. Fifty years in the future, doctors will have access to vast amounts of information through a multitude of noninvasive diagnostics. Progressively more personalized medicine should inspire doctors to become even more adept at communicating effectively with patients. Professionalism in medicine throughout these generations embodies similar fundamental behaviors, such as demonstrating compassion, respect, and humility; adhering to high ethical and moral standards; subordinating personal interest to that of others; and reflecting on actions and decisions. Despite the dynamic nature of the profession itself, the omnipresent need for such traits will define medical professionalism for decades to come.

Understanding professionalism in modern medicine occupies a great deal of attention in medical education, research, and practice. A universally recognized definition eludes the profession, but professionalism itself may be considered in the context of the era in which it applies. This exploration of pro-

Copyright 2008 by The Johns Hopkins University Press

Perspectives in Biology and Medicine, volume 51, number 4 (autumn 2008):554–64
Medical Professionalism

Professionalism across the generations of medical care utilizes three vignettes, each followed by brief discussion. The fictitious accounts chronicle doctors of three eras across 100 years of American medicine. The final vignette, which takes place five decades from today, encourages more avid speculation on what it means to be a doctor in the 21st century.

While a comprehensive definition of medical professionalism might be better served by an exposition wholly devoted to devising its meaning, a brief consideration of the term may enhance the utility of both the physician vignettes and their corresponding discussions. The issue has been approached from various perspectives, including that of sociologists analyzing the interplay of professional autonomy in exchange for social value, and that of physicians formulating the fundamental principles and professional responsibilities required to maintain “medicine’s contract with society” (ABIM 2002; Freidson 1988, 2001; Swick 2000). One attempt at a normative definition of professionalism in medicine focuses instead on core physician behaviors (see Table 1). This definition emphasizes that “the concept of medical professionalism ... must be grounded in what physicians actually do and how they act, individually and collectively” (Swick 2000, p. 614).

**Table 1**  **Physician Behaviors Defining Medical Professionalism**

| Subordinating personal interests to the interests of others |
| Adhering to high ethical and moral standards |
| Responding to societal needs and reflecting a social contract with communities served |
| Evincing core humanistic values: honesty and integrity |
| caring and compassion |
| altruism and empathy |
| respect for others |
| trustworthiness |
| Exercising accountability for oneself and for colleagues |
| Demonstrating a continuing commitment to excellence |
| Exhibiting a commitment to scholarship and advancing the field |
| Dealing with high levels of complexity and uncertainty |
| Reflecting upon actions and decisions |


Dr. John Chapman crushes the remains of his cigarette in the ashtray and adjusts his tie in the rearview mirror. He has arrived for an after-dinner call at the Reeves home. At his office and in their home over the last 20 years, the Reeves
family has received the bulk of its care from Dr. Chapman, even during occasional trips to the local hospital. Tonight the family patriarch, Henry, has taken a turn in his battle with cancer, as Mrs. Reeves informed Dr. Chapman over the phone earlier in the day.

Dr. Chapman quickly checks the contents of his bag, although he doubts he’ll use many supplies tonight. His stethoscope, morphine, digitalis, ampoules of 50% glucose, aspirin, codeine capsules, atropine, epinephrine, and others are included in his armamentarium.

After brief greetings, the doctor is ushered upstairs to Henry’s bedroom, where he performs a methodical physical exam while speaking reassuringly to his longtime patient. Mr. Reeves continues to suffer abdominal pain with intermittent fevers, and his reduced appetite has resulted in significant weight loss in the past few months. Dr. Chapman focuses on analgesia and comfort for Mr. Reeves, and he spends the next two hours at his patient’s bedside. Eventually, Mr. Reeves is comfortable enough to sleep, and Dr. Chapman leaves the room quietly to converse with the family.

He is encouraging regarding Mr. Reeves’s comfort and the level of care he is receiving at home. But he is also realistic about what can be done at this point. He reassures the family that he is always available to them. Mrs. Reeves asks few questions and is calm despite her grief; she has always appreciated and respected Dr. Chapman’s knowledge and steady demeanor. The doctor excuses himself and returns home to his family by 11 pm.

In the 1950s, the concept of professionalism suffused the work of doctors but was rarely discussed or addressed. While even today many “attempts to define professionalism as a set of virtues, obligations, and behaviors fall short of capturing its essence,” no one was trying to define it 50 years ago; issues of lifestyle, litigation, and quality assurance were simply not part of public discourse (Smith 2005). The lifestyle was doctoring itself. A solo practitioner hanging out his shingle was a commonplace sequela of completing training, and those practitioners were overwhelmingly white and male—even by 1970, only 9.2% of United States medical graduates were female (Silberger et al. 1987).

Although some timeworn therapies have persisted for decades, medical knowledge has advanced markedly in the last 50 years. In the past, both available therapies and understanding of disease processes were more limited. Patients like Mr. Reeves were told “No more can be done” more frequently and earlier in the course of disease. The doctor’s role in crisis was support at the bedside, either in hospitals or in patients’ homes.

Doctors in the community were often general practitioners; three-fourths of the population received care from family doctors who worked, in some cases, seven days per week, not including essentially nightly at-home call (Luce 1950). The pace was slower with long hours. Doctors’ lives lacked balance, and focus remained squarely on the patient; this was the norm. As Francis D. Moore (1959) said in the preface to his textbook: “The fundamental act of medical care is assumption of responsibility . . . complete responsibility for the welfare of the pa-
tient” (p. 1). Physicians of every specialization practiced the laying on of hands as a true craft, an art form. With limited imaging and laboratory-based options, these doctors had no choice but to hone their abilities in the history and physical exam. Sequential physical exams, rather than technology, allowed physicians to follow their patients’ progress. These years solidified the concept of medicine as a calling, and a noble one at that.

The patriarchal nature of health-care delivery produced the sense of “doctor knows best”—doctor as the authority. The doctor-patient relationship demonstrated notable formality; one would not hear patients calling physicians by first names other than “Doctor.” Physicians and surgeons extended “professional courtesy” often at no charge when treating doctors or their family members—whether or not any prior personal connection existed. While these aspects of doctor-patient interactions may no longer be as common, physicians like Dr. Chapman exemplified those core behaviors that rest at the heart of medical professionalism, such as subordinating one’s own interests to those of others, demonstrating high moral standards, and contending with matters of great uncertainty.

Health care 50 years ago was even more focused on “sick care” than it is now. Topics of diversity, the underserved and underrepresented, and direct-to-consumer marketing (though it occurred through radio and print) were less germane for physicians of the day. Polio held the attention of the Western world. The dawning of “the DNA era” occurred with the discoveries of James Watson, Francis Crick, and Rosalind Franklin. Prescription and nonprescription drugs were differentiated for the first time, and by 1956 the NIH had achieved a budget of $100 million (Shannon 1987). The Medicare and Medicaid Program would not be signed into law until 15 years later, on July 30, 1965; the landscape did not yet reflect the immense financial, procedural, and cultural influence of this initiative (CMMS 2008). The emphasis on research and discovery would evolve, but over the next 50 years, doctors would begin to question what it meant to practice medicine in the midst of technological advances and what it meant to be professional at the turn of the 21st century.

2000s

Dr. Luisa Martinez takes advantage of the 30 minutes before her first scheduled appointment to check her voicemail and e-mail. A general pediatrician, she still focuses more heavily on diabetes mellitus than others in her group.

She hears a mix of Spanish and English coming from Exam Room 1 and recognizes the voice of Lucy Pasley, a seven-year-old girl with type 1 diabetes. Lucy giggles in response to her mother, Maria, tickling her as she has her temperature checked with an aural thermometer. Lucy has appeared for a well-child check, and Dr. Martinez spends the next few minutes asking how Lucy has been doing at home and in school while conducting a physical exam that Lucy thinks is just play.

Maria always arrives with a list of questions jotted for the doctor on note-
paper and a separate list of “Lucy’s sugars” for the past few months. Today, she is most interested in asking Dr. Martinez about an article she read online from the Chicago Tribune describing the use of a sulfonylurea to treat another young girl with formerly insulin-dependent diabetes. The doctor discusses the implications of that article with Maria and emphasizes the rarity of the genetic mutation involved that permits the therapy to work. She’s happy spending a few extra minutes with the Pasleys, but she is not surprised to notice that the remaining exam rooms are already full.

Dr. Martinez closes the visit after 12 minutes and moves to the next exam room. She smiles at a pharmaceutical representative holding a tablet computer and toting a rolling briefcase filled with samples. Dr. Martinez distributes those samples liberally to help her patients who struggle to afford their prescriptions. Normally, the doctor would pause to greet the representative, but she needs to see the next patient before doing so. Dr. Martinez would like to complete her scheduled visits by the early afternoon in order to meet a small group of medical students at the nearby school to discuss techniques in physical diagnosis.

Today’s doctors enjoy unprecedented access to medical knowledge and therapeutic options. The wealth of clinical understanding is more readily available in formal medical education and through online resources, digital pharmacopeias, and pocket references. But doctors are also more openly introspective on the nature of the profession through reflective writing, subjective surveys of attitudes and perspectives at all levels of training, and debriefings for medical teams following emotionally challenging care experiences (Donohoe 2002; Furnham and McGill 2003; Redinbaugh et al. 2003; Shapiro et al. 2006). Indeed, the very concept of medicine as a calling is being questioned. Young would-be practitioners are choosing careers in the business of medicine over the practice of medicine—consulting, pharmaceuticals, insurance, medical administration. Veteran physicians are also moving into these pursuits and they may cite how things used to be and job dissatisfaction as motivators (Pathman et al. 2002; Zugler 2004).

Despite doctors’ seeking careers outside of clinical care, more doctors are practicing in the United States than ever before, and, while the majority of physicians are white, minorities are increasingly represented. Female matriculants to medical schools surpassed males for the first time in 2003, and women represent 25% of practicing physicians (Levinson and Lurie 2004). Foreign medical graduates compete effectively for residency positions in most specialties. Specialization is the order of the day, although general practitioners continue to emerge from family medicine and primary care training programs to work in the community. The landscape has changed.

The public enjoys dramatically improved access to medical information—admittedly of varying degrees of accuracy—through the media and the internet. Medical decision-making is being analyzed through lenses of increasing acuity. Now that doctors can do more, they fear repercussions of not doing enough, errors of omission. “Defensive medicine” comprises the practice of treating lia-
bility risk as well as the disease at hand; physicians may order additional tests to reassure patients or may avoid patients at higher risk of bad outcomes (and presumptively higher risk of filing suit) (Kachalia et al. 2005). Coping with uncertainty is one of the core behaviors of professionalism, and constant advances in understanding demand intellectual vigilance. But cogent forces have placed some doctors on the defensive, as they view staying up-to-date as a survival tool instead of a fundamental aspect of their practices (Millenson 1997). The sense of accountability and the opportunity for reflection on medical decisions fundamental to professionalism are threatened by such a harsh environment. Recent efforts to remove the specter of punishment from error-reporting processes and to assure underreporting of errors demonstrate a continuing commitment by the profession to preserve its ideals (Lehmann et al. 2007; Taylor et al. 2007).

Issues of professionalism supersede events specific to the doctor-patient interaction. Medical school career panels now include discussions of work-life balance. Lifestyle has become an issue of conflict between trainees and trainers just as it has become the topic of investigation; trainees and medical students are influenced by issues of income, locale, flexibility of hours, and familial obligations (Maiorova et al. 2008; Newton et al. 2005; Orpin and Gabriel 2005; Sanfey et al. 2006). House calls have drastically reduced in prevalence, although they have not disappeared (Meyer and Gibbons 1997). The increased use and reliance on technology in the setting of sound clinical decision-making has created the impression of doctors as technicians. Trainees balance the necessity of developing physical exam skills in the context of emerging, disruptive technologies. The doctor who spent hours at the patient’s bedside in crisis might instead be effectively and successfully treating multiple critically ill patients in intensive care.

Physicians practice in the midst of calls for health-care reform and increased activism. Doctor-patient interactions are more informal, and professional courtesy may still translate to “VIP treatment” but no longer to “no charge.” Interactions with representatives of affiliated industries—pharmaceuticals, device development, and so forth—and studies demonstrating impaired decision-making in the setting of reciprocity have limited what representatives of Big Pharma can do to educate and entertain doctors (Schneider et al. 2006; Wazana 2000).

Physicians are surrounded by changes in their profession, but what of professionalism? The medical community seeks an open discussion on the topic, a concept whose principles “are really like the bricks and mortar that ultimately build a framework of good medicine” (Stevens 2007). Metrics of professional behavior and hundreds of studies of professionalism as early as medical school admissions contribute to the literature (Reddy et al. 2007; Smith 2005).

Beyond the research, the writing, and the discussion, professionalism is most important in its impact on the quintessential act of medicine—doctors treating patients. The account of Dr. Martinez is no singular occurrence. It contains aspects of days in the lives of tens of thousands of physicians across the country. This simple outpatient visit resonates with doctors and patients alike partly be-
cause it is relatively ordinary. In fact, it is easy to take for granted the demeanor, attitude, and behavior of the doctor for precisely this reason. Dr. Martinez effortlessly combines core humanistic values and her involvement and understanding of scholarship in her conversations with her patients. Her commitment to excellence includes her desire to combine her medical knowledge with her compassion, to make this visit a normal part of life for a young girl. By so doing, Lucy will be accustomed to an aspect of her health that will always be her condition but need not be her illness.

In any system of training that relies so heavily on experience and apprenticeship, discord between trainers and trainees cannot be avoided. Unprofessional behavior (making disparaging personal remarks about patients, discussing protected health information in public places, inappropriate use of alcohol and recreational drugs) and, indeed, the very discussion of professionalism are exacerbating the generation gap in medicine (Reddy et al. 2007). Professionalism cannot be ignored, and entrants into the medical profession must be willing to engage on some level in the conversation. Medicine will never again be the lifestyle-unconscious field that veterans refer to as the “age of heroes.” But recent medical graduates also cannot assume that earning a degree means they know what they need to know about earning a patient’s trust and providing the best care, even when therapeutic options beyond palliative care have run out. In the next 50 years, this professional schism must be negotiated. If it is not, doctors in 2050 may actually be no more than technicians, as patients become increasingly more interested in “what the test shows” instead of “what the doctor has to say.”

2050s

Scanning the morning’s appointments on her digital assistant, Dr. Elisa Tang hugs her children before they hurriedly grab their bags and run outside to the driverless automated bus. Dr. Tang logs into the electronic medical records (EMR) system for her hospital from home by glancing into a scanner on her computer which recognizes her retinal imprint.

A cardiac electrophysiologist, Dr. Tang opens her “virtumail” while simultaneously selecting the remote diagnostics that have been transmitted from her patient’s devices overnight to servers for the clinical information system (CIS). She makes minor adjustments, which will be synced with her patients’ miniature “cardiosupportive” devices overnight.

Dr. Tang dictates to her digital assistant as she travels to the medical center. The audio files sync to the dictation service through the CIS at the hospital. Automated transcription follows with remarkable accuracy. Advances in communication have permitted Dr. Tang and her colleagues to accomplish more from diverse locations, but she wonders if they substitute for seeing patients in clinic. As she arrives at clinic, she recalls the words of a professor in medical school: “Touching will always be the bond and bridge between doctors and patients that is the therapeutic catalyst.”
When Dr. Tang considered going to medical school, she was told doctors just didn’t do as much anymore. Along with electronic records, genechips, and advances in tissue engineering like artificial cartilage, society has seen other major steps: robotic-assisted surgery is commonplace, and neutron imaging has become clinically feasible, improving yield in diagnosing breast cancer among other diseases more easily missed by other modalities. The entire light spectrum is now utilized to test noninvasively for and assay almost every component previously available only from blood or biopsy material.

The traditional physical exam is obsolete, but “laying on of hands” is still a central tenet of the doctor-patient interaction for Dr. Tang and her colleagues. Technology can keep track of every potential drug interaction in the pharmacy, but it still cannot triage very well. Improvements in speech recognition have made automated translation a developing reality, but software cannot read a patient’s face and choose the perfect words with which to deliver bad news. Widespread use of subjective patient surveys following clinic visits and hospital stays in Dr. Tang’s medical system continue to demonstrate that patients value the time with their doctors above any other aspects of their visits.

As the information network worldwide expanded, doctors found new avenues into homes. Dr. Tang has primary care colleagues who make “house calls” through virtual conferencing. Patients can be seen frequently using biometric systems set up by home health care. Patients still come into clinic, but doctors have found these home systems more efficient and more convenient to ensure follow-up.

Dr. Tang arrives at clinic and greets her nurse, Patricia. “You have two patients waiting, Doctor,” notes Patricia, as Dr. Tang sits at a computer in the workroom. “Both have chips—I scanned them already.” Dr. Tang nods and thanks her.

Not all patients have agreed to the recently introduced MedChip program—“your medical history in the palm of your hand”—microchips with solid-state memory inserted into their distal radii. Providers need only wave a scanning wand over the chips to download aspects of a patient’s medical record to clinical information systems. Emergency medical services throughout the country are major proponents of the new system. Early data has shown that patients tolerate MedChip implantation well and generally forget they even have the devices. Dr. Tang has seen some patients shudder when she asks them if they have MedChips as they describe their privacy concerns. She supports the system because it allows her to spend more time talking to patients about their chief concerns for that visit instead of simply rehashing lengthy medical histories and medication lists. It also allows providers to avoid repeating recent, costly, and potentially identical evaluations at unaffiliated facilities.

Dr. Tang carries no paper chart into the exam room and simply activates the touch-display on the wall as she enters. She uses these displays to illustrate some of her favorite images describing the cardiosupportive devices and procedures that she recommends for her patients. With the displays, patients participate more actively in their visits than in the days of the prescription pad, because they see images of all of the medications that “the doctor thinks they should be taking.” Dr. Tang has learned to engage her patients more effectively as a result of these
advances. As far as she is concerned, any technology that gives her more time to see and talk to patients instead of charting, billing, or simply writing about patients has a place in her clinical practice.

Five decades from today, the technological landscape will represent many advances in all aspects of care. It will be all the more crucial for doctors to be able to communicate these advances to patients in the setting of an increasingly complex array of options. The tenets of professionalism outlined for Dr. Chapman in 1950 will still hold true for Dr. Tang in 2050. Compassion, integrity, competence, and critical thinking will all be foundations for the practice of medicine. The need to balance the advancement of medicine with societal need and a sense of accountability will be even more important as the unprecedented wealth of medical, pharmaceutical, and technological options will demand the vigilance of physicians in every aspect of their practice. As their tools help them treat disease as never before, they will require more confidence in their knowledge and abilities to interpret vast amounts of data and to be able to connect with patients beyond the machines, the pills, and the lab tests. Practitioners must prevent doctor-patient interactions from becoming sterile, mechanical, technology-driven processes. Therefore, those nine core behaviors of professionalism will be even more critical for doctors 50 years from now.

The concept of personalized medicine will impact therapeutics and pharmaceuticals. Patients will come to expect medication tailored to their particular receptor maps through pharmacogenetics. Quality assurance will no longer be a topic of controversy, discussion, and heterogeneous implementation. It will be prerequisite to ensure patients are receiving care that, more than ever before, is truly "designed" for them. The widespread reporting processes described in Dr. Tang’s medical system should lessen the tension between doctors and patients. Doctors will be less reticent to admit medical error or suboptimal outcome, and patients, who in the past may have sought legal reparations, will be satisfied with honest apology. Patients will also be more apt to believe steps are being taken “so it won’t happen again.”

The nation will represent the most diverse population in history, and cultural competency will be as basic as knowledge of anatomy and physiology to medical education. Underserved areas will be the last to benefit from many improvements to the scientific standard of care. But attention to reaching these areas taken by current trainees will set a precedent for future generations. Primary care will not disappear, particularly as the advancements become more commonplace and costs associated with them decrease. Primary care doctors, free clinics, and nonprofit organizations will be able to redirect these resources to underserved areas over time.

As device and imaging technology becomes more miraculous, the perception of “doctors as technicians, mechanics, tradespeople” may only worsen among some patients. But the end result of these advances remains improved therapeu-
tic options, better outcomes, and stronger bonds between the beneficiaries of all of these changes—doctors and patients.

Regardless of the time period involved, professionalism imbues medicine. It simply takes different forms over the years. Professionalism was not directly discussed in the 1950s, but it remained a combination of mastery of current medical knowledge enabled by contemporary technology and delivered with compassion, humility, respect, and sensitivity to the needs of patients and their families. The discussion of professionalism occupies research, academics, and clinical practice today, but its basic characteristics are not so different. Fifty years from now, professionalism in medicine will still be personal, tactile, humanistic, supportive, and helpful whether the problem at hand is grave or trivial. Dr. Chapman would be surprised about many elements of Dr. Tang’s practice, just as she would be amazed at his ability to practice so effectively in an era of relative technological unsophistication. But both would immediately recognize the traits they hold in common: love of the craft, commitment to their patients, competence, compassion, facility in communication—simply put, those qualities that make them both not only great doctors, but true professionals.

References


The profession of medicine is based on a shared set of tacit and explicit agreements about what patients, doctors, and society at large should be able to expect from each other, a social contract that defines the profession. Historically, the development of this set of agreements depended upon the creation of social organizations that could speak for the entire profession. Over the last several decades, however, the perceived need for these organizations, and especially the umbrella organization for the profession, the American Medical Association, has waned. The reasons for this are complex, but the consequences are significant: an eroding social contract, fragmentation, lack of cohesion and integrity, and loss of the public’s confidence. The present social contract is one-dimensional, overly simplistic, and failing to sustain the public’s trust. To address these problems, a renewed social contract is necessary. Although this renewed contract should be based on foundations similar to the original, it must directly confront such contemporary challenges as resource allocation and conflicts of interest. Equally as important, to reinvigorate our social contract more physicians will need to come to grips with a basic truth: to sustain professionalism we need a strong, unified professional association.
The exact birth date of medicine as a profession is murky and depends on one’s definition of “profession.” But if one accepts a bare-bones definition—a group that publicly “professes” to share uniform training and standards of practice, which they promise to use in service to others—it is possible, roughly, to date the birth of medical professionalism. And it is much younger, and perhaps more fragile, than many might imagine it to be.

Some would date medical professionalism to the Hippocratic era. Margaret Mead has noted that Hippocrates first separated the roles of healer and sorcerer (Bulger and Barbato 2000). They famously swore an oath professing standards of conduct, and they promoted empirical observation as the basis of medical practice. Nonetheless, as the eminent historian Ludwig Edelstein (1943) has argued, the Hippocrates were a minority sect, who did not succeed in creating uniform standards of practice and behavior for all Greek physicians. Contravening some Hippocratic dicta, Greek physicians performed abortions and assisted in suicides (Baker 1993). The rich and the powerful could even hire Greek physicians as medical hit men. According to the Roman historian Tacitus, the emperor’s wife, Agrippina, hired a Greek court physician, Gaius Stertinius Xenophon (ca. 10 BCE–54 CE), to poison her husband, the Emperor Claudius (The Annals, Book XIV, 1–16). Popular acceptance of this account suggests that the Hippocratic prohibition against harming patients was not uniformly practiced by Greek physicians. Instead, most physicians of the time were simply specialists in the uses of chemicals and botanicals, unbound by a uniform code of conduct or standards of practice. As the medical historian Albert Jonsen (2000) put it, in Hippocratic times “there does not appear to have been anything like a medical profession” (p. 9).

Others might date medical professionalism to the Middle Ages or to the Renaissance, when standard curricula in medical schools, novel public-health efforts, and the hiring of “plague doctors” by towns began to clarify some of the social obligations that medical doctors should take on. For instance, in 1666 William Boghurst, a London apothecary, asserted that physicians were obliged to treat patients during epidemics. Yet these obligations and social roles were neither clearly articulated nor widely accepted—indeed, the standard advice of physicians facing the plague in this era, both for themselves and their wealthy patients, was cito, longe, tarde: go quickly, go far, and don’t come back too soon. The fact that towns had to hire specific doctors to stay and care for patients during epidemics suggests that a commitment to continue providing care was not acknowledged as part of the physician’s role.

The term medical ethics and the modern use of profession first appeared in the early 19th century, when an English physician, Dr. Thomas Percival of Manchester, introduced them in his book, Medical Ethics (1803). Percival (1803) clearly articulated specific social roles for all physicians and hoped to see these widely adopted. While it is tempting, therefore, to date the birth of medical professionalism to 1803, Percival’s efforts to get the British medical profession to agree to a written set of ethical standards for all physicians were, unfortunately,
sharply rebuffed. The sentiment in England at the time was that proper gentlemen didn’t need written ethical standards, because they already knew how to behave. In fact, as Baker et al. (1999) put it, codes of ethics were considered “undesirable” because they were “useful only to persons who, lacking decent character, wish to pretend that they had one.”

In the end, it was the American medical profession that, in the mid-19th century, created the first national set of ethical and practice standards. Eventually, similar standards were almost universally accepted, thereby creating the modern concept of the medical profession. American physicians were primed for the task of creating a full-fledged profession for several reasons. Perhaps most important was the Americans’ attraction to the notion of a social contract—a notion conceived by French, English, and Scottish Enlightenment thinkers, but implemented most fully in the young American republic, created by rebels against inequitable classism. In the United States, people were to relate as equals. Social relations were to be built upon more-or-less explicit contracts between willing parties, not such nebulous notions as noblesse oblige or gentlemanly honor. This way of thinking led to the desire to specify the terms of social relations. In medicine, this specification would take the form of a written code of ethics.

In 1847, American medicine was in disarray. There were no uniform standards for medical education, medical practice, or medical ethics. Most medical care was ineffective and often life-threateningly dangerous. Caveat emptor ruled the field. The free market was leading to the rampant production of a wide variety of uneducated and unorthodox practitioners. The survival of scientific medicine was under threat—at risk of dying before it had been fully born, let alone produced any of the miraculous cures it would later deliver. In this environment, a group of “orthodox” practitioners met to draw up a set of educational and ethical standards, by which they might define—and defend—the nascent “profession” of scientific medicine. The document they produced, the 1847 Code of Medical Ethics of the American Medical Association (AMA), was the first national code of ethics for any profession.

This code of ethics, which was hailed at the time for being as revolutionary as the Declaration of Independence (Baker et al. 1999), was clearly derived from the work of Percival, the Hippocratics, and others. Yet it was also quintessentially American. It laid out a three-part social contract, with reciprocal obligations spelled out between physicians and patients, physicians and other physicians, and physicians and their communities. In many cases these obligations were significant and specific. The three chapters of the code were drawn along the lines of these reciprocal obligations. With regard to community-physician obligations, for example, a physician is “required to expose his health and life for the benefit of the community, [and] he has a just claim, in return, on all its members, collectively and individually, for aid to carry out his measures.” In relations with individual patients, physicians were to “be ever ready to obey the calls of the sick,” “secrecy and delicacy” should be “strictly observed,” and so on. But in return,
patients were to select only properly trained physicians and to “faithfully and unreservedly communicate to their physician the supposed cause of their disease” (yet a patient should not “weary” the physician with “tedious detail”), and, of course, “the obedience of a patient to the prescriptions of his physician should be prompt and implicit” (Baker et al. 1999, appendix B and C).

These reciprocal obligations did not depend on the personal virtue of the practitioner, though it was certainly hoped that virtuous individuals would join the profession. Instead, the obligations of medical professionals were laid out, explicitly and in writing, so that patients, the community, and physicians all would be aware of these standards. The profession aimed to make uniform claims about the quality of its practitioners, which would be the basis of public trust and improved public health (and—not coincidentally—the foundation for the establishment of self-regulation and monopoly power).

One can certainly argue about the extent to which these reciprocal sets of obligations were lived out, and the degree to which physicians, in particular, lived up to the ideals they espoused in the code. One can also raise questions about the extent to which patients were a willing party to this new contract. Nonetheless, the general notion that all physicians have specific and unique obligations, and a special, privileged role in society, became widely accepted only after this new group of professionals was willing to (1) put these matters in writing and (2) develop mechanisms for self-regulation to encourage adherence to its new code (Wynia 2006). Indeed, the social status of physicians was eventually raised to near-stratospheric heights, based in part on this explicit social contract that demanded altruism, civic-mindedness, devotion to scientific ideals, and a promise of competence and quality assurance through self-regulation.

The Role of Professional Associations

Since professions are group-based social entities, being part of a collegial community is an essential feature of professionalism. In particular, when a profession is based on a written social contract—a code of ethics—the organization that writes this code becomes very important. If a practitioner wants to affect the social contract, the way to do so is through the professional association. And participation in local, state, and national professional associations became important for many other reasons as the medical profession became socially recognized and successful—that is, as the social contract played out.

Some activities of the early AMA were guild-type activities, such as the fact that bank loans and malpractice insurance were often contingent upon AMA membership. Other activities and standards more clearly promoted the public good, or were plainly altruistic—such as the obligation specified in the AMA Code that “when pestilence prevails,” physicians must continue to care for patients despite the risk to their own health and even (after 1912) “without regard” to remuneration (Huber and Wynia 2004).
Being a member of one’s professional association was also how one kept up-to-date on the evolving science of medicine, a special challenge for far-flung solo practitioners in the United States. It was how one forged collegial relations—needed for referrals and assistance during surgery, for example. The famous physician Sir William Osler repeatedly noted the importance of professional societies as the fertile ground in which professionals grew: “You cannot afford to stand aloof from your professional colleagues in any place. Join their associations, mingle in their meetings, gathering here, scattering there; but everywhere showing that you are faithful students, as willing to teach as to be taught” (Bryan 1997, p. 51).

As this quote suggests, professional associations played an important role in developing the non-monetary reward system of early medicine. According to early sociologists of the medical profession, monetary rewards were scant and a surprisingly rare motivator for those entering the medical profession. Talcott Parsons, for example, suggested that people who became doctors tended to be driven less by money than by a desire to look good in front of their peers (Latham 2002). Insofar as this was true, presenting work to one’s peer group was important not only to science, but to the development of a cohesive, collegial professional community.

Participation in professional associations was also an ethical obligation. For medical leaders in particular, participation was seen as a core altruistic obligation to the future of the profession. Again, according to Osler: “no physician has a right to consider himself as belonging to himself; but all ought to regard themselves as belonging to the profession, inasmuch as each is a part of the profession” (Bryan 1997, p. 50). Once, when Osler was asked by a medical student whether he (the student) should attend a local medical society meeting, because he wasn’t sure what he would get out of it, Osler responded, “Do you think I go for what I can get out of it, or what I can put into it?” (Bryan 1997, p. 49).

**Advances in Science, Loss of Humility**

By the turn of the century, scientific medicine was beginning to show its promise. While previous generations of doctors had believed, often falsely, that they had something of medical benefit to offer the ill, the generation of doctors that understood public hygiene and inoculation actually did save lives, and dramatically so. Between 1900 and 1920, deaths from typhoid, diphtheria, and gastritis were cut by more than half, and tuberculosis deaths dropped by one-third. By the 1940s, with the introduction of penicillin and streptomycin, influenza deaths plummeted, and tuberculosis deaths were falling so rapidly that the disease was widely expected to be eliminated. When books like DeKruif’s *The Microbe Hunters* (1926) noted both the self-sacrifice and success of physicians in combating infectious diseases, many Americans came to see physicians as heroes.

Sadly, one effect of gaining heroic status was the loss of any remnants of
humility that doctors might have retained from their Hippocratic roots. Interestingly, the Hippocrates’ emphasis on humility had been based on an awe of the gods’ powers over human life and a belief that physicians would be guilty of hubris if they intervened contrary to the gods’ plans. Later generations of physicians saw the human body as mechanistic, amenable to manipulation and measurement, and the subject of scientific scrutiny and learning. They should have (and some had) derived humility from their belief in scientific questioning—recognizing that scientific knowledge is always tenuous and subject to further refinement (Wynia and Kurlander 2007). John Gregory (1724–1773), for instance, called such scientific humility “diffidence” and held that “candor, which makes him open to conviction, and ready to acknowledge and rectify his mistakes,” is a moral duty for physicians, urging that errors in care be used to study and improve medical practice (Gregory 1772, pp. 209–10). Samuel Bard, founder of the Columbia College of Physicians and Surgeons, told graduating medical students in 1769: “Whenever you shall be so unhappy as to fail, in your Endeavors to relieve; let it be your constant Aim to convert, particular Misfortunes into generaly Blessings, by carefully inspecting the Bodies of the Dead, inquiring into the Causes of their Diseases, and thence improving your own Knowledge, and making further useful Discoveries” (pp. 13–14). Scientific humility, insofar as it drove scientific inquiry and the development of new treatments, was tremendously successful. But, perhaps predictably, as science made advances and medicine had greater success, it became harder for physicians to remain humble. Those physicians who sought out errors to learn from them, brave pioneers of quality improvement like Richard Cabot (1868–1939) and Ernest Codman (1869–1940), were often vilified by other practitioners.

Some of this vilification reflected basic human nature—the reluctance to admit error or have one’s errors exposed. But it might also have reflected an ongoing divide early in the development of the profession, between the science and art of medicine: researchers were more interested in science, while clinicians were more devoted to art. To be sure, many believe that this divide was, and remains, largely artificial, since practicing medicine without attention to science would be foolish, and caring for human beings without attention to art would be cruel: both are necessary to good medical practice. In effect, however, in some of these debates the term art was code for the notion that individual practitioners should be allowed to practice according to their own best judgment, often uninformed by the latest science and without meaningful oversight from colleagues or anyone else.
ing of the AMA, there had been an undercurrent of concern amongst practitioners over the following question: would professional autonomy mean that the profession, as a group, was to establish standards (rather than having them established by the state or through the marketplace) and ensure that all members lived up to them? Or would it mean that each individual professional, once found to be qualified, would be allowed to establish their own patterns of practice?

We’ll return to this question momentarily, but early on—certainly throughout the Progressive Era (ca. 1890–1913)—it appeared that the debate was being resolved in favor of professionals, as a group, establishing standards and mechanisms of self-regulation (Burrow 1977). For example, within a year of its founding, the AMA established committees to set standards on medical education, medical sciences, practical medicine, surgery, obstetrics, and medical literature and publications. Committees on anatomy, physiology, materia medica, chemistry, forensic medicine, vital statistics, hygiene, and sanitary measures soon followed (Haller 1981). The proposed arrangement was clear: individual practitioners would benefit from professional social privileges garnered by the AMA, but in return they were expected to follow the dictates of the profession, as set by AMA committees.

As science advanced, the divide between clinicians and scientists seemed to narrow. New scientific measurement tools, such as the stethoscope, various blood tests, and microscopy, became part of the medical care armamentarium. The clinicians’ preference for artful rather than scientific practice looked to be on the wane. Dr. John H. Musser, President of the AMA in 1904, remarked, “With the incoming of scientific precision there is the outgoing of so-called art. Diagnosis by intuition, by careless ‘rule of thumb’... is as little trustworthy as the shifting sand of the Sahara” (King 1983, p. 2478).

Other Perils of Scientific Success and Authority

Linking practice to science led to great advances in patient care and public health. Sadly, however, the downsides of this success-linked-to-science were substantial: physicians not only came to lose humility and respect for “the art,” but their customer service orientation as well. Medicine became increasingly complex, and microscopic phenomena weren’t always easy to explain. Perhaps more important, a mechanistic understanding of the human body meant that medicine could provide tremendous benefits whether or not the patient understood or believed in how these benefits came about (such as with inoculations). So physicians pushed for public-health mandates at the population level and adopted a highly paternalistic attitude towards patients at the individual level.

But pride, paternalism, and the loss of art and customer service were, sadly, not the only negative consequence of this focus on scientific competence as the
source of physicians’ social authority. Another was that physicians’ civic obligations eventually came to be taken for granted, seen as unimportant, or misconstrued; and many were nearly abandoned.

First, in the wake of vaccination, antibiotics, cardiac surgery, organ transplantation, and other miracles, any professional obligations beyond scientific competence no longer seemed necessary. Saving lives was sufficient to garner high levels of public respect. Second, some civic obligations, such as the professional duty to continue caring for patients during epidemics, were eventually seen as “anachronistic,” because the achievements of scientific medicine had made them so. As the U.S. Surgeon General put it in 1970, “the era of infectious diseases is coming to an end” (Huber and Wynia 2004). It’s not hard to imagine a profession with this level of hubris feeling little need for any ethical regulations—after all, what could be more ethical than eliminating disease?

Third—and more complex—is that the profession accrued so much credibility there was no longer any question that it should be self-regulatory. At first blush, this development might seem to promote the civic obligation of self-regulation, but gaining the unquestioned capacity to self-regulate created an unfortunate backlash. From the time of its founding, a goal of the AMA had been to develop a heavy mantle of credibility around physicians that would create a professional monopoly, or “professional closure,” with the assistance of the state. That is, those who were not qualified, according to standards established by the profession, would be closed out of practice by the state. If successful, professional closure would protect the public from unscrupulous and unscientific practitioners. It would also raise the status, and presumably the pay, of qualified practitioners. (It is, in my view, impossible to fully disentangle these altruistic and self-serving motivations.) As physicians delivered on their promises to improve medical care, and risked their own lives in doing so, the profession became extremely successful in arguing for regulatory closure. In fact, medicine was so successful in this regard that many of our self-regulatory mechanisms, such as medical licensure, accreditation bodies, and various other professionally derived structures and processes, were accepted as legally binding—which blurred the lines between the state and the profession. Victims of our own success, many physicians no longer recognized these various regulatory structures as a part of professional self-regulation and necessary to maintaining our social credibility over the long term; instead, they came to be perceived as meddlesome outside bodies, sent in by the state to scrutinize us and disrupt our practice.

Finally, though it pains me to admit it, the burgeoning field of medical ethics also contributed to the loss of physicians’ sense that professionalism entails civic responsibilities. Early bioethics, responding to legitimate concerns—ranging from paternalism, as noted above, to physician participation in Nazi crimes against humanity under the guise of obligations to society—strongly stressed the importance of autonomy as a principle of biomedical ethics and deemphasized
or even denigrated physicians’ civic duties. Some urged physicians to ignore civic considerations altogether and think only of the welfare of the individual patient before them. For instance, in 1984 Norman Levinsky wrote in the *New England Journal of Medicine* that “physicians are required to do everything that they believe may benefit each patient, without regard to costs or other societal considerations” (p. 1573). Such a statement reflects the domination of medical ethics by respect for individual autonomy, but it also illustrates the loss of a cardinal facet of the social contract that had grounded physician professionalism, and which the sociologist Talcott Parsons had described: the obligation of physicians to serve as mediators between private and community interests (Latham 2002; Wynia et al. 1999).

In sum, in the late 20th century there developed a very different sense of professionalism, epitomized by the notion that one should care only about the patient sitting in the exam room. As a simple, one-dimensional ethics, this notion of strict individual advocacy appealed to patients’ immediate interests, and it seemed easy for doctors. But it could hardly be more different from the initial understanding of professionalism as comprising a complex set of reciprocal obligations between physicians, patients, and the community.

**The Physician as Trustee**

Under the original social contract for the medical profession, doctors had obligations to patients but also obligations to the community—and it was recognized that these could come into conflict. While stewardship of shared financial resources was not an obvious issue early on (before health insurance came into existence), conflicts arose around patient wants and desires, and the hope of the community for those patients to be productive members of society. When these responsibilities conflicted, a good professional would serve as a mediator, seeking to do the best possible for all concerned.

Even more than for other professions, this mediator role was an important part of the social contract for physicians. In simple, practical terms today, the agreement is the following: physicians are given certain social privileges to protect the ill (such as by allowing time off work) in exchange for a collective promise to help society by working to return the ill to productive life. So, ethically, physicians cannot sell notes to excuse otherwise healthy people from work, despite the fact that there might be a ready market for them.

This was recognized in the 1847 Code of Medical Ethics, which noted that a physician’s skills “are qualities which he holds in trust for the general good.” And our commitment to serving the larger public good played a crucial part in the professional standing that medicine first achieved during the 19th century. As Cruess and Cruess (1997) put it: “[19th-century] legal measures for the first time granted medicine a broad monopoly over health care—along with both indi-
individual and collective autonomy—with the clear understanding that in return medicine would concern itself with the health problems of the society it served and would place the welfare of society above its own” (p. 943).

**Problems with a One-Dimensional Social Contract**

Under a simplified, autonomy-centric view, however, physician ethics came to look something like lawyerly ethics. Namely, zealous advocacy for one’s client became the primary, if not only, duty of the physician. But the practical and conceptual problems with such a simplistic stance are substantial (Sage 1999), and they are playing out today.

The main problem is that a zealous advocate cannot also serve as the opposing counsel and the judge. But in medicine, unlike in the legal system, there is no opposing counsel. And even if there were, there is no impartial judge to weigh the physician’s arguments against those of this hypothetical advocate for the larger community. To make zealous advocacy work as the physician’s sole ethical responsibility, and to produce just outcomes when the needs of individuals and communities came into conflict, there would need to be a system in place to which the physician would have to plea—and in which the physician would not have the final word.

This scenario is not very appealing to most physicians. An adversarial medical care system would be profoundly inefficient and frustrating for patient and doctors alike. Yet it is what *must* evolve if physicians insist on adopting a one-dimensional advocacy role. And indeed, we are developing just such a system today, with control over medical decisions devolving to health plans and purchasers, to which physicians and their patients must plea.

**Simple Contract, Complex Problems**

This new social contract, based only on advocacy for individual patients, has other ramifications as well. For instance, professional closure weakens. New groups of practitioners arise, unqualified according to the old professional standards but free to practice according to the dictates of the market that an autonomy-centric social contract promotes. We are not there yet, but we are experiencing a slow reversion towards the days before 1847, when anyone could hang a shingle and call themselves a “doctor.”

Also, in the long-running dispute over what professional autonomy means, a simplified social contract decisively tilts the playing field towards those who would redefine professional autonomy to mean the right of individual doctors to treat patients according to individual preference, rather than the right of the group to self-regulate by setting and enforcing practice standards.
As the contract devolves away from groups and towards individuals, there has been a reversion away from codes of ethics and back towards an ethics of individual virtue. Incidentally, this is not to be confused with “the virtues” à la Aristotle, who believed virtue to be habitual and based upon carefully following rules over a long period of time, until they become ingrained. Rather than emphasizing that physicians are bound by a shared set of behavioral standards, which students should embrace until they become second nature, ethics courses in medical schools today tend to focus on training students to think things through for themselves. This, of course, is laudable and a necessary brake against professional group-think, but it’s hard to believe we should depend completely on each individual’s analysis. Such reliance will predictably lead some physicians to take wrong actions that they believe they can justify, and others will start out with a very different understanding of acceptable actions. To put this in colloquial terms: the problem with teaching ethical analysis and then relying on the “red-face test” to maintain professionalism is that some people don’t embarrass easily. Sometimes, we’d be better off with clear rules and a meaningful obligation to follow them.

Finally, with a one-dimensional, individually focused contract, there is less perceived need for organizations like the AMA that wrote and enforced the old, more nuanced and group-oriented, social contract. This is hardly the only cause of the AMA’s membership woes, but it is a key part of a negative membership spiral. Ironically, AMA members—comprising practicing physicians—largely bought into the simplified social contract, in which the association itself became less important. With its loss of stature among physicians came losses in membership and social prestige, and a reduced ability to influence the environment of medical practice. Then, more doctors chose to abandon the organization, because it came to be seen as ineffectual even in its more limited role. Organizational leaders facing such a situation can easily become desperate, casting about for ways to please the remaining members. In their efforts to serve them, it is easy to further alienate those on the margins, by moving even further from the core mission around which the AMA was created: writing the social contract for medicine and ensuring that all physicians are living up to it.

Specialty associations have tried to inherit some of the AMA’s power to establish their own, independent social contracts with some success, since they can better focus on negotiating for a relatively homogeneous membership. Sadly, however, these efforts often result in the increasing fragmentation of the profession and frequent episodes of internecine conflict. As cohesion in the professional community declines, so does professional social capital, resilience, and effectiveness.
WHERE TO GO FROM HERE?

Given recent history and current trends, it seems that relatively few physicians might weep over the passing of the AMA, but since no alternative organization is being proposed to take its place, the alternative is to have no national association for all physicians. Most of us probably know, intuitively, that “every one for oneself” is not a solid basis on which to maintain a profession. “Every specialty for itself” isn’t much better. In short, without a unified professional association we cannot have a profession.

Can we rebuild medicine’s social contract to meet the challenges of the new century? Can we create a new progressive era for medicine, retaining our commitment to science while building back in and reinforcing our obligations of service to society, artful practice, humility, and professional autonomy (in its original sense)? Is it possible to rehabilitate old institutions, such as the AMA, to help accomplish this task?

We don’t want or need the same social contract today that we developed in 1847. A contemporary social contract should focus far more attention on matters of resource distribution, quality measurement, and the interactions of the various players in the health-care system. (It’s not just patients and doctors anymore: purchasers, regulators, and other practitioners must be brought into the contract.) And, in fact, these ideas are gaining traction within the AMA (Ethical Force Program 2008).

Many progressive physicians, however, have lost hope for the AMA and its capacity for evolution, even though most know little of how the AMA actually works. In my view, rumors of the AMA’s demise are premature. The fundamental role of professional associations is to write the social contract for the profession. Our options are to have multiple organizations perform this task—with different social contracts for each specialty—or to have a uniform social contract for all physicians. There are good reasons to favor the latter.

Second, the AMA remains engaged in this task, and the process through which it works (though imperfect), is, on the whole, fairly solid. The AMA is a representative democracy, with representatives from all major specialties and every state. Naturally, democratic structures reflect the majority thinking of those who are involved. So the profession of medicine, and the AMA in particular, faces something of a Pogo problem: we have met the enemy . . . and he is us.

Finally, American medicine exists within a democratic society. Physicians are not alone in establishing our social contract, we do so in constant negotiation with various communities. Often, these negotiations take place through democratic processes, and our professional associations are the means we have of projecting the voice of medicine into public policy debates. If certain physicians don’t like the tenor or content of the voice of American medicine, it is not enough to leave. There is, as Osler understood, a professional obligation to be engaged and help change what the voice is saying or how it is being said.
Nevertheless, some of us have become inured to political polarization over the last 40 years. Some might see all of organized medicine as beyond redemption—too much in hock to corporate interest, too attached to a political party, too reactive. As a result, the AMA might have lost large segments of two or more generations of physicians, who are so cynical about organized medicine that they cannot imagine an evolved AMA, one that might (at least sometimes) reflect their values and help orient the profession towards public service. Sadly, in my experience many leaders of academic medicine—though progressive at heart and generally not lacking a sense of empowerment—are in this position. They hold a deep-seated cynicism about the AMA and its ability to change—or their ability to help change it.

We should not give up on these leaders: their skills and knowledge can be invaluable. At the same time, though, we need to directly engage young professionals who haven’t yet adopted this cynical attitude. Activism among young physicians is rising, as is AMA membership, even while it continues to fall among more senior members of the profession. In the last year, membership in the AMA among physicians under 40 rose 2.2%, while membership among those older than 40 fell 2.8% (Julie Gill, AMA Membership and Marketing, personal communication, May 15, 2008). Perhaps the best we can hope for from some medical leaders will be a bemused silence, as the young progressives under them learn how to use our professional association to reinvigorate the social contract of the medical profession.

References

Bard, S. 1769. A discourse upon the duties of a physician, with some sentiments, on the usefulness and necessity of a public hospital. New York: A. and J. Robertson.


E XPECTATIONS A ND
OBLIGATIONS

professionalism and medicine’s
social contract with society

RICHARD L. CRUESS AND SYLVIA R. CRUESS

ABSTRACT As health care has become of great importance to both individual citizens and to society, it has become more important to understand medicine’s relationship to the society it serves in order to have a basis for meaningful dialogue. During the past decade, individuals in the medical, legal, social sciences, and health policy fields have suggested that professionalism serves as the basis of medicine’s relationship with society, and many have termed this relationship a social contract. However, the concept of medicine’s social contract remains vague, and the implications of its existence have not been fully explored. This paper endorses the use of the term social contract, examines the origin of the concept and its relationship to professionalism, traces its evolution and application to medicine, describes the expectations of the various parties to the contract, and explores some of the implications of its use.

social contract: A basis for legitimating legal and political power in the idea of a contract. Contracts are things that create obligations, hence if we can view society as organized “as if” a contract has been formed between the citizen and the sovereign power, this will ground the nature of the obligations, each to the other.

The subject of medicine’s professionalism has assumed increasing importance during the past decades because of the widespread belief that medicine’s traditional values, which are closely linked to professionalism, are under threat (Cruess and Cruess 1997; Freidson 2001; Hafferty 2006a; Krause 1995; Starr 1984; Stevens 2001; Sullivan 2005; Wynia 1999). There is a rich literature that defines professionalism and outlines medicine’s obligations as professionals (ABIM 2002; Cruess, Johnston, and Cruess 2004; General Medical Council 2006; Hafferty 2006b; Royal College of Physicians of London 2005; Swick 2000; Wynia et al. 1999). Virtually all observers contributing to this literature are in agreement that professionalism serves as the basis of medicine’s relationship with society, and most believe that the relationship is best described by the term social contract.

Based on the concept’s foundation in philosophy and political science, we also believe that social contract is the most appropriate descriptor of the relationship. In this article, we describe how the social contract relates to professionalism, define the concept in contemporary terms, and provide an outline of the nature of the current contract. Finally, we discuss some of the implications of a social contract approach to medical professionalism.

**Medicine and Society**

There has been a surprising degree of agreement on the fundamental nature of the relationship between medicine and society. Virtually all who have described it state that society has granted medicine autonomy in practice, a monopoly over the use of its knowledge base, the privilege of self-regulation, and both financial and nonfinancial rewards. In return, physicians are expected to put the patient’s interest above their own, assure competence through self-regulation, demonstrate morality and integrity, address issues of societal concern, and be devoted to the public good (Abbott 1988; Carr-Sunders and Wilson 1933; Cruess and Cruess 1997; Elliot 1972; Freidson 1970; Kultgen 1998; Parsons 1951; Stevens 2001; Sullivan 2005; Wynia et al. 1999). While there have been disagreements about the motivation and performance of the members of the profession and the state of health of the “bargain,” its existence and the presence of a mutual state of “dependency and obligation” between medicine and society seems accepted (Freidson 2001; Haug 1973; Johnson 1972; Klein 2006; Krause 1996; Larson 1977; McKinley and Arches 1985).

As long as both society and the medical profession were content, there was little effort to formally categorize their relationship. However, as changes in both medicine and society have sparked widespread dissatisfaction with the current state of health care, a variety of models have been proposed to describe medicine’s relationship with society. Starr (1982) first suggested that the relationship is contractual, stating that the contract was being redrawn in response to dramatic changes in health care, such that the contract was “subjecting medical care
to the discipline of politics or markets or reorganizing its basic institutional structure” (p. 380). Subsequently many observers, including social scientists, lawyers, policy analysts, bioethicists, and physicians turned to the historical concept of the “social contract.” Sullivan (2005) emphasized the link between professionalism and the social contract, stating that the “the social contract became the moral basis of professionalism” (p. 54).

Klein (1990) used the term “implicit bargain” when describing the relationship between the government, the National Health Service, and the medical profession in the United Kingdom. Although some observers in the United Kingdom refer to the presence of a social contract, it is probable that Klein’s choice of words influenced others who noted that the “bargain” had broken down (Davies and Glasspool 2003; Ham and Alberti 2002). Three recent studies of medical professionalism in the U.K. have used the term “implicit compact” which specifically includes doctors, patients, and society, and all state that the relationship involves reciprocity (Edwards, Kornacki, and Silversin 2002; Rosen and Dewar 2004; Smith 2004a). The Royal College of Physicians of London (2005) has proposed that morality is so fundamental to the practice of medicine that the social contract should be renamed a “moral contract” (though without elaborating on the details).

The term social contract also has been applied to other relationships in contemporary society, including some that touch medicine directly: those between society and its medical schools, between society and science, and between society and universities (Gibbons 1999; Inui 1992; Kennedy 1997; H.R. Lewis 2006; Lubchenko 1998; Ludmerer 1999; McCurdy et al. 1997; Schroeder, Zones, and Showstack 1989).

The Social Contract: Origins and Evolution

The concept of the social contract was developed by 17th- and 18th-century philosophers, primarily Hobbes, Locke, and Rousseau, at a time when most countries were ruled by hereditary monarchs (Crocker 1968; Masters and Masters 1978). It had two purposes: to provide an historical account of the origin of the state and society as citizens united their individual “wills,” and to explain the nature of the relationship between the state and its citizens. It outlined a series of reciprocal rights and duties as being fundamental to this relationship. While the concept has not been universally accepted as the philosophic basis of the state, it has had a continuous presence in philosophic discourse (Bertram 2004; Rawls 1999, 2003). Rawls’s (1999) theory of justice is a contemporary expression of contractualist thinking: “those who engage in social cooperation choose together, in one joint act, the principles which are to assign basic rights and duties and to determine the division of social benefits” (p. 10). The philosophers endorsing the concept of a social contract were clear that there is no formal legal contract. However, they justified the use of the term on the grounds that “the
Rights and duties of the state and its citizens... are reciprocal and the recognition of this reciprocity constitutes a relationship which by analogy can be called a social contract” (Gough 1957, p. 245).

Contemporary interpretation of contract theory leans heavily on the idea of “legitimate expectations” as being fundamental to mutual understanding (Bertram 2004; Rawls 2003). In addition, the failure of one party to meet the legitimate expectations of the other has consequences in the attitudes and actions of the other.

The social contract can be considered a “macro” contract including all essential services required by a population, but it has also been proposed that there are “micro” contracts, applying to individual essential services required by society, which must conform to the “moral boundaries” laid down by a macro contract (Donaldson and Dunfee 1999, 2002). Health care could be included in the overall relationship or, given its importance to the well-being of both individuals and society, it could be governed by its own micro contract. It appears that this latter approach best describes the structure of society, recognizing that health care is one of a number of conflicting priorities within the macro contract.

The details of the social contract between medicine and society differ between countries, being influenced by cultural, economic, and political factors. While there are many documented commonalities, there are also significant differences in the funding and organization of health care, and hence in how professionalism is expressed. What seems not to differ is the role of the healer, which answers a basic human need (Dixon, Sweeney, and Gray Pereira 1998; Kearney 2000). Those elements of the social contract that refer to the healer are relatively constant across national and cultural boundaries, while those that refer to how the services of the healer are organized, funded, and delivered may vary (Cruess and Cruess 1997; Krause 1996; Laugeson and Rice 2003).

As society and healthcare evolve, the social contract also evolves, expressing the relationship between society’s dominant constituencies. The literature on professionalism recognizes that professionalism changes in response to societal needs (Freidson 2001; Krause 1996; Starr 1982; Stevens 2001). Indeed, Castellani and Hafferty (2006) have warned against continued reliance on the “nostalgic professionalism” of the past. As the social contract changes, the professionalism that serves as its basis must also evolve. Therefore, in assessing the current state of professionalism, legitimate contemporary societal expectations must be emphasized. However, equal importance must be given to those aspects of professionalism that are valued by both society and the medical profession but may not be given the same level of importance by the commercial sector or governments (Freidson 2001; Hafferty 2006a; Melhado 2006; Stevens 2001; Sullivan 2005; Tuohy 2003).

Selecting the most appropriate descriptor of the relationship between medicine and society is important, as it has the potential to give a mutually agreed-upon framework for discussion. Originally conceived to protect both individual citizens...
and the public from the abuses of authoritarian rule, social contract theory now emphasizes the mutual rights and obligations of citizens and those governing them. Most of those analyzing the interface between medicine and society believe that the relationship involves reciprocal rights, privileges, and obligations. Of the many terms suggested, only *social contract* has an historical basis in philosophy and political science, having been in wide use for three centuries. For this reason, it does not require redefinition and should more easily serve as a basis for dialogue between the parties to the contract.

### The Social Contract in Health Care

The social contract in health care is a mixture of the implicit and the explicit, the unwritten and the written. In all countries, the explicit parts include legislation outlining the structure of the health-care system, laws establishing the regulatory framework, including licensing, certification, and discipline, and jurisprudence relating to health care (Hafferty and McKinley 1993; Krause 1996; Starr 1982). The Hippocratic Oath and codes of ethics also constitute an explicit part of the contract, outlining medicine’s commitment, as do the International Charter on Medical Professionalism (ABIM 2002), Good Medical Practice (General Medical Council 2006), and Good Medical Practice USA (2007). Many of these documents impose legal obligations on physicians and the profession (Rosenbaum 2003; Rosenblatt, Shaw, and Rosenbaum 1997). However, there are also both written and unwritten portions entailing moral commitments that are fundamental to both the social contract and professionalism (Pellegrino 1990; Stevens 2001). One cannot legislate altruism, commitment, or independent professional judgment; they must come from within individual physicians (Couléhan 2005; Kultgen 1998; May 1997).

Until recently, most observers have been content to outline medicine’s relationship to society as bilateral, while recognizing the presence of multiple stakeholders in health care. The usual statement is that “there is a social contract between medicine and society.” This seems to assume that two major players exist: a relatively monolithic medical profession made up of individual physicians and their institutions and patients and wider society.

Reality is different. Rosen and Dewar (2004) analyzed the relationships between the multiple stakeholders involved in health care in the United Kingdom and integrated them around the concept of reciprocity. In redefining medical professionalism, they proposed a new “compact” involving three interlocking societal components. The first group consists of patients and patient groups as well as the “public”; the second of health-care managers, the state, government departments, and the European Parliament; and the third of the medical profession and “professional bodies.” There are interactions within each group, and each group has reciprocal relationships with the other two. Each relationship is “mediated” by the media, the legal system, and the regulatory framework. The
commercial sector was not included as a major stakeholder or “mediator.” While we agree with Rosen and Dewar’s approach, we believe that it does not correctly outline the nature of the interrelationships.

**Parties to the Contract**

A schematic representation of our concept of the contemporary social contract in health care, including its complex interrelationships, is presented in Figure 1.

*The Medical Profession*

Medicine is not monolithic. It includes individual physicians and those institutions traditionally mandated to carry out medicine’s collective responsibilities (licensing and certifying bodies, and educational and training institutions) and their national and specialty associations. The interests of primary care physicians do not always coincide with those of specialists and sub-specialization often results in significant differences between specialists (Abbott 1988; Starr 1982; Stevens 2001, 2002). Professional associations have been described as representing an elite whose priorities may differ from those of practicing physicians. There is a constant interplay between and among individual physicians and medicine’s institutions that must take place if the profession is to develop a consensus on the issues pertaining to its social contract with society (Laugeson and Rice 2003; J. M. Lewis 2006; Peterson 2003; Salter 2001, 2003).
Society

Society is also complex, consisting of patients and the general public on the one hand, and government on the other. Physicians and medicine relate to each societal component. The primary relationship of the individual physician in both moral and fiduciary terms is with the individual patient (May 1975; Pellegrino 1990; Rosenbaum 2003; Rosenblatt, Shaw, and Rosenbaum 1997). This relationship cannot be isolated from the system within which it operates, nor from the wishes of society as a whole. Other health professionals and their organizations, disease-oriented and consumer groups, industry, individual citizens, and the unorganized general public are all partners to the contract (Brown et al. 2004; Blumenthal 2006; Callaghan and Wistow 2006; Ham and Alberti 2002; Morone and Kilbreth 2002; Rosen and Dewar 2004; Salter 2001, 2003). As within the medical profession, there is a dynamic interplay between the various nongovernmental stakeholders as they interact with each other, which results in the elaboration of what patients and the public wish from physicians and their organizations (Le Grand 2003).

In line with contract theory, physicians and those representing them and patients and the general public have expectations, “each of the other.” A proposed outline of these expectations is given in Table 1. Professionalism serves as the basis of this relationship, essentially establishing the rules of the game as outlined in medicine’s declaration of applied morality, its code of ethics.

Medicine also has an important relationship with government, because the profession operates using powers delegated to it by society through government action. Governments are also complex, being composed of elected politicians, civil servants, and (particularly in publicly funded institutions) managers. Again, there is a dynamic interaction between these individuals or groups of individuals that results in public policy. There are also a series of expectations and obli-

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Expectations: The Public and the Medical Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients’/public’s expectations of medicine</td>
<td>Medicine’s expectations of patients/public</td>
</tr>
<tr>
<td>Fulfills role of healer</td>
<td>Trust sufficient to meet patient’s needs</td>
</tr>
<tr>
<td>Assured competence of physicians</td>
<td>Autonomy sufficient to exercise judgment</td>
</tr>
<tr>
<td>Timely access to competent care</td>
<td>Role in public policy in health</td>
</tr>
<tr>
<td>Altruistic service</td>
<td>Shared responsibility for health</td>
</tr>
<tr>
<td>Morality, integrity, honesty</td>
<td>Balanced lifestyle</td>
</tr>
<tr>
<td>Trustworthiness (codes of ethics)</td>
<td>Rewards: nonfinancial (respect, status), financial</td>
</tr>
<tr>
<td>Accountability/transparency</td>
<td></td>
</tr>
</tbody>
</table>
gations resulting from the relationship between medicine and government. (See Table 2.) Because of the current dominance of the state or the commercial sector, to which the state may delegate a major role, the relationship between medicine and government is now extremely important, as are the expectations and obligations of the two parties (Freidson 2001; Krause 1996; Light 2001; McKinley and Marceau 2002; Starr 1982; Stevens 2002). Professionalism governs medicine’s actions in dealing with governments.

Finally, as society is made up of government and those governed, patients as citizens and the general public enjoy a relationship with government that is closer to the classical vision of a social contract. Patients, their representatives, stakeholder groups, and the general public must deal with the elected officials, civil servants, and health-care managers mandated to ensure that citizens and the public receive the preventive and therapeutic measures in health expected in a modern society. The expectations and obligations of these parties are illustrated in Table 3. While professionalism does not play a role in this relationship, the nature of the social contract between the public and government is expressed in the structure and funding of the health-care system and has a profound effect upon the professionalism of medicine, either supporting or subverting its healing role and traditional values (Freidson 2001; Light 2001; Stevens 2001; Sullivan 2005).

**The External Influences**

Three important external influences on the social contract and on the interactions between the three parties are: (1) the health-care system, including the role of the private sector; (2) the regulatory framework; and (3) the media.
The Health-Care System and the Private Sector

The relationship between the commercial sector and both physicians and patients is usually outlined by legal contracts (including insurance policies), not a social contract. However, the role of the marketplace in health care is clearly part of the overall social contract. Its magnitude is decided by government action or inaction and accounts for many of the national differences in the nature of the social contract (Hafferty and McKinley 1993; Krause 1996; Marchildon 2006; Rosenbaum et al. 1999; Tuohy 1999; Vogel 1986). Most countries have systems that combine public and private roles, with the nature of the mix ultimately determined by legislation. When medicine or the general public wishes to change the system, it must be done through the political process.

The nature of the social contract between medicine and society imposes limits on the legal contracts outlining the obligations of practitioners, the commercial sector, and government. When these limits are exceeded, the public will react. Recent examples include the gag laws that prohibited physicians from informing patients of therapeutic options not included in their insurance coverage, and the attempt to impose a 24-hour limit on hospital stays following obstetrical delivery. The public and the medical profession, supported by the media, objected and, working through the political process, established that some decisions must remain between physicians and their patients (Rosenbaum et al. 1999). There are also limits on the actions of the medical profession. A physicians’ strike in Ontario over the right to bill more than the approved fee schedule received no public support. The profession did not gain its objectives, and its reputation was severely damaged (Meslin 1987). These exemplify the often unwritten constraints on all parties to the contract to remain within the moral boundaries perceived to be part of the social contract.

The Regulatory Framework

The regulatory framework of a country impacts the social contract. Countries such as France, where the government retains the right to regulate, have differ-
ent contracts from those with systems drawn from the Anglo-Saxon tradition, where more emphasis is placed on the independence and autonomy of the profession (Hafferty and McKinley 1993; Irvine 2003; Krause 1996).

The Media

The impact of the media on the social contract can be profound, especially in contemporary society with rapid communication within and between countries. The “Bristol cases” in the United Kingdom provide a powerful example. Pediatric cardiac surgery was carried out with unacceptably high mortality rates for years. The facts were known to many with administrative responsibility both within and without the institution, but it was not until the media revealed them to the general public that action was taken (Irvine 2003). Public indignation prompted an extensive reevaluation by government of the concept of self-regulation and recommendations for significant changes in the process, including partial withdrawal of the profession’s regulatory powers (Salter 2003; Secretary of State for Health 2007). The failure of the medical profession to self-regulate constituted a breach of its obligations under the contract, and the media was instrumental in highlighting this fact, leading to a change in the contract with an alteration in the expectations of the major parties.

**Expectations Under the Contract**

As is true of all contractual relationships, there are expectations on both sides. Tables 1, 2, and 3 propose a list of the current expectations of the three parties to the social contract derived from a review of the literature (see Appendix).

Many of the expectations of the three parties have been present since the modern professions were established by licensing laws in the mid-19th century, but there have been dramatic increases in the nature and magnitude of the expectations and changes in how they are expressed. Societal expectations have increased because modern science has given the healer greater capacity to cure, increasing medicine’s importance to the average citizen (Rawls 1999; Starr 1982). Physicians have also altered their expectations from the 19th and early 20th century when physician incomes and status were relatively low (Klein 1990; Krause 1996). New expectations have been the added to the contract, such as the desire of individual physicians for a balanced lifestyle and the expectation that physicians will participate in team medicine (Blendon et al. 2006; Borges et al. 2006; Chisholm, Cairncross, and Askham 2006; Coulter 2002; Henningson 2002; Holmstrom, Sanner, and Rosenqvist 2004; Johnston 2006; Levinson and Lurie 2004; Neufeld, Maudsley, and Pickering 1998; Schoen et al. 2005; Watson et al. 2006). Because the understanding of many physicians as well as patients and the public is often based upon a nostalgic understanding of the professionalism of yesteryear, the changes that have occurred to the social contract must be understood by all parties.
In addition to the changing expectations of the various parties, the expectations of one party may conflict with those of another. An example is the realization that younger physicians of both sexes wish time for family and outside interests (Borges et al. 2006; Henningson 2002; Holmstrom, Sanner, and Rosenqvist 2004; Johnston 2006; Levinson and Lurie 2004; Watson et al. 2006). This may conflict with the altruism fundamental to the practice of medicine (Coulehan 2005; Inui 2003; McGaghie et al. 2002). If patients believe that their doctor is pursuing his or her own interests during the relationship, they will lose trust and the physician’s ability to heal may be diminished (Coulter 2002; Hall 2005; Mechanic and Schlesinger 1996; Pellegrino 1990). Faith in the morality, integrity, and honesty of physicians is fundamental to trust. For generations this trust was given blindly. Now it must be constantly earned. It is also important for physicians to trust the health-care system within which they function, and the commercial organizations with which they deal. If this trust is not present, physician motivation changes, cynicism occurs, and patient care may suffer (Gould 2001; Hall 2005).

For a century and a half, the expectation of both elements of society has been that the profession will assure the competence of its members through self-regulation. Until recently, licensure and certification obtained early in a career was felt to be sufficient. Because of the well-documented failure of the profession to self-regulate, society is now demanding proof of competence, including professionalism, throughout practice (Irvine 2003). This has added significant new obligations to contemporary professionalism.

Major changes have occurred in physician autonomy and accountability. In earlier times physicians were accountable primarily to their patients and their colleagues. They are now accountable to governments and commercial organizations for their competence, performance, productivity, and the cost-effectiveness of their activities (Broadbent and Laughlin 1997; Emanuel and Emanuel 1996; Timmermans 2005). In addition, courts have established new levels of accountability as judicial interpretation of legislation and malpractice claims have increased, particularly in the United States (Moran and Wood 1993; Rosenbaum 2003; Rosenblatt, Shaw, and Rosenbloom 1997; Starr 1982; Vogel 1986). One consequence has been a decrease in the autonomy of physicians in practice (Broadbent and Laughlin 1997; Emanuel and Emanuel 1996; Freidson 2004; Krause 1996; Rosenbaum 2003; Salter, 2001; Timmermans 2005). However, both patients and physicians continue to expect sufficient autonomy to be preserved for physicians to make independent decisions in partnership with their patients (Chisholm, Cairncross, and Askham 2006; Neufeld, Maudsley, and Pickering 1998).

Another important change in the social contract relates to the development of public policy. Health and health care are essential if citizens are to live normal and productive lives, and health-care policy has a significant impact on access, cost, and quality (Freidson 2001; Krause 1996; Moran and Wood 1993; Richmond and Fein 2005; Starr 1982). For this reason, the public wishes to influence
health-care policy. Physicians also wish to have input, as they believe that they possess expertise essential to the proper formulation of health-care policy. For their part, governments state that they welcome participation of the public and the medical profession, but they appear to wish to control their input (Le Grand 2003; Salter 2003).

As trust in the medical profession has decreased over the past few decades, it has been realized that the perception that individual physicians and the profession represent a force for good in society—a force not restricted to health care—is of great importance, something which in the past was assumed. This point has been emphasized by several eminent social scientists. Sullivan (2005) has suggested that practicing “civic professionalism” is essential for the profession, stating that in becoming a professional, one assumes a civic identity involving a duty to function “in such a way that the outcome of the work contributes to the public value for which the profession stands” (p. 23). Stevens (2001) has written that the profession must fulfill its “public roles” and be seen to be doing so in an exemplary fashion in order to gain public support for the concept of professionalism. And in his last book, Freidson (2001) has outlined what he termed the “soul” of professionalism and indicated how important its preservation is to the public good.

While there are differences in expectations between the parties, there are also areas of agreement. For example, government expects patients and the general public to have what they would term “reasonable expectations” of the system (Klein 1990; Le Grand 2003; Marchildon 2006; Salter 2001), although there are differences in how individual patients and health planners would define “reasonable.” The same holds true for the desire of both the medical profession and the public to have a health-care system that is adequately funded and staffed, with the differences focusing on the methods and levels of funding.

**Contract Theory and Who Rules**

If the term social contract is to be a valid descriptor of the relationship between medicine and society, it should be compatible with the often shifting patterns of power and influence on public policy in the health-care field, an issue well documented in the literature. In his classic work, Freidson (1970) described the dominance of medicine. Almost immediately, others questioned this dominance, suggesting that medicine was being depersonalized, proletarianized, and being subjected to bureaucratic control (Haug 1973; McKinley and Arches 1985; Starr 1982). Eventually all, including Freidson (2001), agreed that medicine’s dominance has been greatly diminished. The theory of countervailing forces emerged, according to which there is a dynamic interplay between the medical profession, government, and the corporate sector. The balance has shifted with either government or the corporate sector now assuming dominance depending upon the structure of the health-care system (Krause 1996; Light 2001; Mechan-
ic 1991). Patients and the general public were not assigned an independent role, assuming that government would represent their interests. While confirming the loss of influence of the medical profession, recent literature has suggested that the public also has been disenfranchised and has stressed the importance of medicine engaging patients and the public in both the development of policy and in decision-making at the local level (Allsop, Jones, and Baggott 2004; Cohen, Cruess, and Carpenter 2007; Krause 1996; Le Grand 2003; Morone and Kilbreth 2002; Salter 2003).

Government is ultimately responsible for establishing the structure of a health-care system, including the balance between public and private payment. Tuohy (2003) has traced the changes in accountability and governance in health care. She states that health care has long had “indirect” governance, beginning with a principle-agent relationship between government and medicine and proceeding to one based on a contract model. Neither appears appropriate to contemporary conditions, where governments are “simply one set of actors among others in complex networks linking different social and economic sectors as well as different orders of relationships from the local to the total” (p. 201). She and others have called these “loosely coupled networks,” where the role of the government is to guide, negotiate, broker, and facilitate the emergence of consensus. She believes that the nature of the issues facing contemporary health care is driving governance in this direction.

The schematic representation of the social contract appears to be compatible with the changes in power and influence among the various parties that have occurred in recent times. The role of the corporate sector and the regulatory framework, both of which have a profound influence on the social contract, result from choices made by society and expressed through legislation in the structure of the health-care system. We would suggest that the “loosely coupled networks” function under the umbrella of a social contract and that the schematic representation provided includes the major participants in the loosely coupled network and outlines their interrelationships. The reciprocity described will continue to necessitate an interaction between the parties, no matter which party is dominant.

**Implications of This Approach**

Applying the concept of the social contract to professionalism, and hence to the relationship between medicine and society, has been said to reframe the discussion of this relationship in three ways (Kurlander, Morin, and Wynia 2004). First, it identifies the parties involved in shaping the relationship. This is essential if the complexities of contemporary health care are to be fully understood. Second, it helps to focus the discussion on the issues that pose the greatest challenge to contemporary health care, emphasizing areas of disagreement as well as consensus. Finally, it assists in establishing the moral boundaries of professional concern.
We would suggest an important fourth advantage: interpreting professionalism within this framework emphasizes professionalism’s relevance to the practice of medicine and makes the profession’s obligations and the reasons for their existence more understandable. In addition, the failure of individual physicians or of the profession as a whole to meet legitimate societal expectations should logically result in consequences, including the possibility of a significant change in the contract.

There are several implications to this approach. In the first place, the reciprocity inherent in the idea of a social contract underlines the importance of correctly interpreting the expectations of both medicine and society. In addition, the idea of reciprocity legitimizes the idea that the profession has expectations of society and should encourage the profession to negotiate those aspects of the social contract that can increase the ability of individual physicians to fulfill the role of the healer (George, Gonsenhausser, and Whitehouse 2006; Wynia et al. 1999). It also highlights the current state of affairs in the United States, where it is unclear who would actually negotiate on behalf of the medical profession in a country without a national health plan, and hence a central negotiating table (Cruess and Cruess 1997; Stevens 2001). Many societal expectations of the profession must be met by medicine’s institutions. These include most aspects of self-regulation—specifically, ensuring physicians’ competence through setting and maintaining educational standards and assuring quality of care. The concept of a social contract makes these expectations and the profession’s obligations under the contract explicit and indicates why they should be the concern of every practicing physician.

The final point relates to the teaching of professionalism and the transmission of professional values, the primary responsibility of medical schools and their associated teaching institutions. Teaching professionalism as the basis of medicine’s social contract provides a rational basis for the existence of both the expectations and obligations of the various parties. Under the social contract, the collective expectations of patients, the public, and government of the medical profession constitute a functional definition of medical professionalism and a summary of medicine’s professional obligations. As Kultgen (1998) has stated: “Entry into the profession is a voluntary act, and most people who perform it are disposed to learn its ways and take its ideology seriously. They need only to be told how” (p. 366). We believe that one should add “why”—and that the social contract provides a cogent answer.

References


Perspectives in Biology and Medicine


**APPENDIX: REVIEW OF LITERATURE**

The expectations of the various parties to the social contract cannot be derived from a single source. Tables 1 through 3 are based on a review of the literature in the following fields.

A ltruism: May 1975; Pellegrino 1990; Piliavin and Charng 1990; McGaghie et al. 2002.


Patients' desires and satisfaction with their care: Neufeld, Maudsley, and Pickering 1998; Coulter 2002; Schoen et al. 2005; Blendon et al. 2006; Chisholm, Cairncross, and Askham 2006.


Regulation of the professions: Moran and Wood 1993; Stacey 1997; Salter 2001; Peterson 2003; Rosenbaum 2003; Tuohy 2003.

Moving Beyond Nostalgia and Motives
towards a complexity science view of medical professionalism

Frederic W. Hafferty* and Dana Levinson†

Abstract Modern-day discourse on medical professionalism has largely been dominated by a “nostalgic” view, emphasizing individual motives and behaviors. Shaped by a defining conflict between commercialism and professionalism, this discourse has unfolded through a series of waves, the first four of which are discovery, definition, assessment, and institutionalization. They have unfolded in a series of highly interactive and overlapping sequences that extend into the present. The fifth wave—linking structure and agency—which is nascent, proposes to shift our focus on professionalism from changing individuals to modifying the underlying structural and environmental forces that shape social actors and actions. The sixth wave—complexity science—is more incubatory in nature and seeks to recast social actors, social structures, and environmental factors as interactive, adaptive, and interdependent. Moving towards such a framing is necessary if medicine is to effectively reestablish professionalism as a core principle.

This article reviews the evolution of the modern-day professionalism movement in organized medicine. What started in the early 1980s with fears related to loss of professional stature and concerns about the corrosive forces of commercialism on core professional values has evolved into a broad-
based and formal social movement. This movement encompasses efforts that range from defining and measuring professionalism to developing curricular interventions to promote professionalism during medical training and beyond. More recently, work on professionalism has begun to consider how organizational structures might affect the ability of individuals to manifest core professional values and behaviors.

Clearly, no movement—particularly one that is taking place in an area as complex and rapidly changing as medicine—functions in isolation. Organized medicine’s formal “professionalism project” is actually one of three social movements currently underway within medicine (the other two are evidence-based medicine and patient safety), all of which fall under a broader rubric of quality of care. Related to these movements is the nascent exploration of complexity science as a conceptual framework for understanding medicine and medical practice (Ahn et al. 2006; Bell and Koithan 2006).

This article traces the evolution of medicine’s professionalism movement, focusing on the contemporary margins of the movement. We consider the potential for professionalism to move beyond its current focus as a discourse that stresses individual motives and behaviors to one that includes a more macro-perspective on how systems and structures affect individuals and how organizations themselves might embody professional principles. We then consider how a complexity science perspective might apply to medicine’s professionalism project and use the hidden curriculum literature to frame an example of how this might take place.

The modern-day (1980s to present) discourse on medical professionalism has been dominated by a “nostalgic” view of professionalism (Castellani and Hafferty 2006). This discourse has unfolded through a series of waves. The first four—discovery, definition, assessment, and institutionalization—have been highly interactive and overlapping sequences that continue to unfold to the present. The fifth wave—linking structure and agency—is nascent, while a sixth—viewing the medical professionalism movement and medical education as taking place within a complex adaptive system—is more incubatory in nature. The sixth wave would consider professionalism through the prism of complexity science, by which we mean the study of dynamic adaptive systems consisting of interacting and inter-dependent variables. We view this final wave as a necessary evolution, if the stated goal of organized medicine’s professionalism movement, the reestablishment of professionalism as a core principle of medical practice, is to reach fruition.

**The First Four Waves in Medicine’s Modern Professionalism Movement**

Medicine has only recently become preoccupied with how best to define and promote “professionalism.” For centuries, the Hippocratic Oath was considered a sufficient ethos to guide physicians and therefore had a commanding influence, even as modern codes of ethics were being implemented. The first American
Medical Association Code for Medical Ethics (1847) focused on the moral authority and independence of physicians in service to others, affirmed the profession’s responsibility to care for the sick, and emphasized individual honor (Riddick 2003). With this code serving as a normative anchor and as a point of social legitimation, organized medicine began its evolution into the type of organizational structure that social scientists would come to label “professional” (Starr 1982). Nonetheless, and in spite of organized medicine’s tenacious pursuit of professional powers and privileges, which reached its zenith in the 1950s, there was little formal emphasis within the medical training process on the tenets of professionalism. Furthermore, outside the rather limited purview of state medical boards, there was little oversight of physician work. This lacuna was fed by medicine’s desire for occupational autonomy (for example, restricting the evaluation of medical work to insiders) along with a persistent vagueness about what exactly might constitute professional or “unprofessional” behaviors (Hafferty 2006b). Instead of critical scrutiny and ongoing refinement, medicine came to treat the idea of its own professionalism as something so routine and obvious as to be taken for granted. Being a professional meant having completed one’s training—nothing less, but certainly nothing more. Even the “automatic attribution” linking professionalism to a degree was considered too extreme by some guild members, particularly those who believed that the core attributes of professionalism were an inherent part of one’s character and therefore beyond the influence of medical training. For these people, the solution to any problems of professionalism lay not in training, but in the recruitment of applicants with exceptional character, particularly those who would prove resilient to the attenuating aspects of medical training. From this vantage point, and taken to its logical extreme, professionalism is a quality that precedes, rather than emerges from, medical training.

How then, did medical professionalism—what it means and how to teach and evaluate it—become such a hotly debated topic at the end of the 20th and the beginning of the 21st centuries? The answer, albeit simplified, is that medicine underwent a number of significant challenges to its powers and privileges during the latter half of the 20th century (Starr 1982). As summarized in Eliot Freidson’s Professionalism: The Third Logic (2001), organized medicine had acquired both professional dominance and professional autonomy based on claims that it had developed an esoteric body of knowledge, an occupationally controlled division of labor and related labor market, the control of new member entry and their training, and an “ideology serving some transcendent value.” By the 1960s and 1970s, however, these powers and privileges began to unravel, fueled by a post–World War II economic boom, the emergence of information technologies (which allowed for, among other things, the monitoring of physician practice patterns), advances in the scientific basis of medical practice, government interference in health care delivery and financing (particularly with the creation of Medicare and Medicaid), and most importantly the rise of commercialism and a substantial for-profit health-care industry (Hafferty 2006a).
One early and internal warning shot about these changes came in 1980, when Arnold Relman, then editor-in-chief of the *New England Journal of Medicine*, wrote a lead editorial expressing profound concerns about the rise of the “medical industrial complex” and its impact on the autonomy and integrity of physicians. Citing the rise in business influences on physician practice, Relman cautioned about new constraints on physician autonomy, as well as the potential for conflict of interest should physicians’ financial interests affect their clinical decision-making. In many respects, Relman was eerily prophetic—for the real boom in medical commercialism would not begin to unfold until 1982, when the American stock market eased into what would evolve into this country’s second longest bull market (1982–2000). Across the 1980s and 1990s, billions of dollars flowed into new and established medical companies—each promising investors a solution to the nation’s health-care woes.

Relman’s admonitions were echoed (although not immediately) by a bevy of other medical leaders including subsequent *New England Journal of Medicine* editors-in-chief Jerome Kassirer (1995, 1997) and Marcia Angell (1993, 2000), and long-time *JAMA* editor George Lundberg (1985, 1988, 1990, 1997). By the early to mid-1990s, evidence of unease about the growing threat of “commercialism,” along with calls for physicians to “rediscover” or “return to” their “core professionalism ideals” were in full bloom (Barondess 2003; Burnham 1982; Davis 1988; McLeod 1982). This unease, with its identification of a common enemy (commercialism) and a generic solution (professionalism), constituted the first wave of the modern professionalism movement (Hafferty 2006a). By the late 1990s, Relman’s warning that commercial influences were making a “hollow mockery of professional oaths” had been elevated from a solitary voice to an occupation-wide consensus (Relman 1998).

This maelstrom of concerns about the corrosive effects of industry soon gave way to a new perspective—that many of these warnings and rallying cries were vague and sometimes internally contradictory. Thus was born a second wave in the professionalism movement, as medical insiders called for more formal and succinct definitions of professionalism and related concepts (CrueSS and CrueSS 1997b; Swick 2000; Wynia, Latham, and Kao 1999). One notable product of this second wave was Herbert Swick’s “Toward a Normative Definition of Medical Professionalism” (2000), with its set of nine requisite behaviors (such as that “Physicians respond to societal needs and their behaviors reflect a social contract with the communities served”), including a framing of altruism (“Physicians subordinate their own interests to the interest of others”) as core to what it means to be a professional. Swick’s definitions and conceptual framework became the basis for work on professionalism by a number of medical organizations, including the American Board of Internal Medicine (ABIM), the American Medical Association (AMA), the Association of American Medical Colleges (AAMC), and the National Board of Medical Examiners (NBME).

The ink was hardly dry on these definitional credos and charters before yet
another cry erupted from within the movement—this time to measure and assess professionalism (Arnold 2002; Arnold et al. 1998; Stern 2005; Veloski et al. 2005). Many of the concerns in this third wave were pragmatic and tied to the emergence of professionalism curricula within medical schools. Advocates advanced two related arguments. First, they claimed that any initiatives to teach professionalism in medical schools would be undermined unless students were formally assessed as a part of this effort. Advocates noted that assessment drives learning, and furthermore (in a null curriculum message), that avoiding assessment would send a message to students not to take any such professionalism initiatives seriously. The second argument (albeit related) focused more on the broader theory of professionalism, including medicine’s social contract with society and the role of peer review and organizational self-assessment in that overall framing. This argument transcended pedagogical pragmatics and went to the very heart of professionalism as a social practice. As a consequence of these and related concerns, efforts surged to assess professionalism along a number of fronts. Literature reviews of assessment efforts were compiled with a focus on admissions and on linking medical school experiences to later clinical behaviors (Etienne and Jullian 2001; Ginsburg et al. 2000; Lynch, Surdyk, and Eiser 2004; Papadakis et al. 1999, 2005; Stern, Frohna, and Gruppen 2005). All the while, discussions as to whether professionalism could be taught, let alone measured, continued unabated (AAMC 1999; Cruess and Cruess 1997a, 2006; Rowley et al. 2000; Whitcomb 2005b).

The fourth wave in medicine’s professionalism movement, and one concomitant with the definition and measurement crests, has been the rise of institutionalization initiatives across a broad constellation of medical organizations. Led by the ABIM’s Medical Professionalism Project, a number of medical organizations, specialty groups, and private organizations such as the AAMC (2004), the NBME (2005), and the ABIM Foundation (Veloski et al. 2004) began sponsoring conferences and allocating resources in what amounts to a collective “professionalism project” (Cohen 2006). Efforts to define and assess professionalism have been core to this overall effort. Examples of products included the Physician Charter created by the ABIM Foundation, ACP-ASIM Foundation, and European Federation of Internal Medicine, and the ACGME’s identification of professionalism as one of its six “core competencies” (ABIM 2002; ACGME 1999). While much of the professionalism reflected in these documents is decidedly nostalgic” in nature, some novel elements are beginning to percolate within medicine’s overall professionalism discourse. For example, in addition to the more traditional calls to place the welfare of patients ahead of provider welfare (altruism) and to promote patient autonomy, the Physician Charter included “social justice” (“The medical profession must promote social justice in the health care system, including the fair distribution of health care resources”) as one of its three “fundamental principles.” In this way, the Charter identified (normatively) medicine’s responsibility to look beyond the physician-patient dyad in framing its professional responsibilities.
A second (and already mentioned) aspect of this institutionalization wave has been the overall effort to create formal coursework on professionalism, particularly for undergraduate medical students (Curry and Makoul 1998). In short order, a wide variety of course materials were developed and implemented at medical schools throughout the United States and Canada—often in addition to curricula on medical ethics (Makoul, Curry, and Novack 1998). Researchers, in turn, began to examine the influence of such formal training experiences on student conceptions of professionalism and on subsequent practice behaviors (Ginsburg, Kachan, and Lingard 2005; Ginsburg, Regehr, and Lingard 2004; Haidet and Stein 2006).

This gaggle of discovery, definition, measurement, and institutionalization has not been without its critics. Concerns have included the lack of a curricular theory of professional development; the lack of attention to the overall learning environment for professional development (which would extend beyond the usual and customary focus on the formal curriculum); the lack of linkages between formal curricular efforts and current professionalism practices, including peer review and state medical board actions; and the lack of consistent and focused calls from within the movement to include a duty to advocate for the well-being of society and the betterment of the public health (Coulehan 2005; Wear and Kuczewski 2004). Finally, critics have cautioned that efforts to measure professionalism were creating a de facto set of implicit definitions—sometimes complementing, but sometimes clashing with, already established definitions.

**Wave Five: Reconciling Professionalism at the Micro and Macro Levels**

While both sociology (theoretically) and medicine (in principle) recognize that there are essential differences between conceptualizing professionalism at the level of the individual versus the organization, most discussion of professionalism generated within academic medicine during the 1980s and 1990s focused on defining, assessing, and institutionalizing professionalism at the individual level—thus promoting an agency-based framing of professionalism (Stark 1989; Todd and Horan 1989). As a consequence, relatively little attention was directed toward understanding how organizations (medical schools, clinics, hospitals, or medical centers) might enable or constrain the motives and behaviors of trainees and practitioners. Still further removed from consideration was the related question of how organizations themselves might behave in a professional or unprofessional manner.

There is, however, evidence that this conceptual cul-de-sac is beginning to change as medical education begins to explore the interactive and interdependent nature of the individual-setting relationship. One example is the aforementioned Physician Charter. The Charter opens its statement on social justice by calling upon the profession to manifest this principle; physician behavior is treated as a secondary concern. Nonetheless, the overall content of the Charter,
including materials in its Preamble and Conclusions, and the wording of all three “core principles” and 10 “commitments,” is firmly embedded in a tradition that establishes professionalism as a matter of individual (physician) responsibilities.

A second and more substantial reframing of professionalism as a collective/organizational responsibility is the high profile currently being accorded issues of conflict-of-interest (COI) within organized medicine (Ross et al. 2007; Stelfox et al. 1998). COI issues are not new to medicine. The Prayer of Maimonides (1135/38–1204), for example, exhorts physicians not to “allow thirst for profit, ambition, for renown and admiration, to interfere with my profession.” Nor is COI the defining professionalism issue. There are a number of other themes, including confidentiality and honesty with patients, that command attention under the rubric of professionalism. Nonetheless, COI does showcase the influence of business/industry on medical work, including research, education, publishing, and clinical decision-making. COI also highlights a “primacy of patient” message, which includes the call to place altruism and the welfare of patients ahead of provider welfare (“selfless-service”), something many medical leaders continue to identify as the sine qua non of medical professionalism (Cohen 2006).

While it is true that earlier calls by key medical organizations to address COI issues did focus on physicians and their responsibility to differentiate between acceptable and unacceptable gifts and to “manage” their relations with industry (AMA 1991), this traditional framing appears to be shifting. In February 2006, JAMA published a “policy proposal” authored by a constellation of medical luminaries that called for academic medical centers to take the lead in adopting policies to eliminate COI within medical learning environments (Brennan et al. 2006). The proposal unequivocally challenges prior COI “myths,” including the myth of small gifts (that any gift can be small enough not to evoke social norms of reciprocity) and the myth of full disclosure (that disclosing a conflict of interest neutralizes that conflict), and it urges medical schools and AHCs to adopt a series of recommended steps to eliminate a hidden curriculum of COI practices. The report’s focus is clearly on organizations and organizational responsibilities with respect to COI, not on the individual, and some critics have cited the report’s “sterile environment” approach to ensuring professional behavior. Several medical schools already have adopted key aspects of the report, highlighting a shift in professionalism orientations from the individual to a more macro-level focus on context.

Three months later, the American Medical Students Association issued a “report card” grading all U.S. and Canadian allopathic and osteopathic medical schools on their COI policies (AMSA 2007). Much to the chagrin of many deans and faculty schools, failures far outnumbered stars, with 42 schools receiving an “F” and 19 a “D”; only five received a grade of “A.” This report has nudged many schools (including those who refused to provide AMSA with initial data) to begin developing formal COI statements governing what pharmaceutical and like companies can do within the walls of medical education.
Another framework for viewing professionalism as more than the motivations and behaviors of individual physicians is laid out in a recent article lead-authored by Jordan Cohen, former president of the AAMC and a coauthor of the JAMA policy article on COI (Cohen, Cruess, and Davidson 2007). The article focuses on the nature of setting and structure as barriers to the manifestation of professionalism principles by individuals, and as such points to the interactive nature (if only uni-directional) between settings and individuals. Cohen and colleagues point out that many of the principles detailed in the Physician Charter are not under the control of individual physicians, and that practices and policies of organizations often function as insurmountable barriers to individuals who might otherwise wish to manifest appropriate professional behaviors. The article also notes that some of the principles articulated in the Charter (universal access, meaningful patient safety efforts, and safeguarding patients from COI) may even fall beyond the province of medicine as a whole—with still broader social forces (such as funding streams) casting a definitive pall over the ability of organized medicine to advance professionalism as a core orienting value. Instead, the article calls for “system wide change” and a functional partnership (a “medical-societal alliance”) between the medical profession and society.

Cohen’s article is notable in two respects. First, his call for a partnership between society and medicine evokes a somewhat overlooked literature on professionalism—one that stands just outside the two major literatures (sociology and medicine) and is sometimes referred to as the “new professionalism” (Epstein 1999; Frankford and Konrad 1998; Irvine 2004, 2006; Mechanic 2000; Sullivan 2005; Whitcomb 2005a). Appearing under several different labels—“civic professionalism,” “democratic professionalism,” “responsive medical professionalism,” “patient centered professionalism”—this body of literature is fairly small, at least relative to the more voluminous professionalism literature, and it is most often published in journals that are (strictly speaking) neither sociological nor medical. Within this subgenre, a common theme is the need to engage the public, proactively and systemically, in any move toward reestablishing a necessary trust between medicine and the public. Thus, when medical insider Troyen Brennan (2002) calls for a professional responsibility grounded in “civic professionalism” and “activist professionalism” and grounds his call within quality of care, we are beginning to see a shift from a professionalism conceived as “just” a matter of individual provider motives or organizational policies to one that resides within the relationships among system participants, including physicians and the public, medical and nonmedical organizations, industry and government. This more encompassing professionalism takes in other medical movements, including patient safety, evidence-based medicine, and quality of care, and extends across such broader social forces as health disparities, an aging population, and, in the United States, tens of millions of uninsured.
Despite an emergent recognition within the medical professionalism movement that settings, organizations, and broader social forces play a critical role in the advancement of professionalism, the operationalization of this perspective into future policies and organizational change is far from certain. As reflected in medical coursework, in documents such as the Physician Charter, and in parallel symbols of professionalism such as ethics codes, the professionalism movement’s primary focus continues to be the individual—with a basic call for physicians to “just say no” to the corrosive forces that besiege it.

One problem with these calls is that they have not worked—at least to date. In spite of a decade of professionalism coursework, legions of articles, and the development of definitions, competencies, and assessment tools, evidence of problems and disjunctures continue to riddle what is assumed to be a comprehensive and coordinated professionalism initiative. For one, traditional definitions of professionalism, which often seat altruism and selfless behavior at their core, appear to be at odds with emerging conceptions of an appropriate (“professional”) physician-patient relationship and issues of lifestyle and “balance” amongst the newest generation of physicians (Croasdale 2003; Dorsey, Jarjoura, and Rutecki 2003; Tholhurst and Stewart 2004). Further, there is some evidence that saturating students with curricula around this topic has had the unintended consequence of creating hostility toward professionalism education in general and a sense on the part of students that they are being “harassed” (Humphrey et al. 2007). Other tensions include a physician population that appears to endorse core ethics of professionalism in principle, including the importance of peer review, but that fails to act when it encounters impaired or incompetent colleagues (Campbell 2007). Meanwhile, medical school faculty persist in modeling unprofessional behavior—leaving students feeling “genuinely and tragically confused” (Brainard and Brislen 2007).

COI data reflect similar inconsistencies and dissonances. While clinicians and researchers appear willing to acknowledge that outside interests might influence their decision-making or behaviors, such an influence, they still insist, happens only to “the other guy.” Despite operating within an occupational culture that touts “scientific evidence” and scientific decision-making, despite ample data documenting the direct evidence of industry gifts and inducements on clinical decision-making and research outcomes, and despite more generic social science research on how even the smallest of gifts can create feelings of obligation, many physicians continue to insist that their clinical decision-making stands above such influences (Alpert 2005; Brett, Burr, and Moloo 2003; Chimonas, Brennan, and Rothman 2007; Steinman, Shlipak, and McPhee 2001). Medical students, meanwhile, express a similar social invulnerability (Fein, Vermillion, and Uijtdewaal 2007). Meanwhile, relations with industry have become the rule rather
than the exception. A majority (77%) of second-year medical students have received gifts from industry (Fein, Vermillion, and Uijtdehaage 2007), and a larger majority (80.2%) believe they are entitled to such gifts (Sierles et al. 2005). Ninety-four percent of all physicians have some type of relationship with the pharmaceutical industry, including food in the workplace, drug samples, reimbursement for attending professional meetings, and payments for consulting, giving lectures, or enrolling patients in clinical trials (Campbell et al. 2007a). The same is true for departments as administrative units (67%), department chairs (60%), and members of institutional review boards (33%) (Campbell et al. 2006; Campbell et al. 2007b). Similar statistics exist for the relationships between medical school research and industry (Bekelman, Li, and Gross 2003).

The overall picture is anything but coincidental when we recognize that drug companies religiously track (on a weekly basis) the prescription-writing behaviors of physicians, by combining prescription data sold by pharmacies to specialized pharmacy-information companies with Drug Enforcement Agency numbers sold by the AMA (which makes millions per year on these information-leasing arrangements). Pharmaceutical companies, in turn, send these data to their salespeople, who, so armed, adjust their inducements accordingly. This more nuanced (and real) picture is not well countered by having organized medicine urge physicians to “embrace the principles of professionalism” or by academic medical centers adopting a set of “sterile environment” COI policies (Carlat 2007).

It is with recognition of this complexity that we suggest reframing the issue of professionalism (which, in all likelihood is not a singular issue at all) from a matter of individual motives, or even as an object of remedial actions at the organizational level, to that of a complex, adaptive system where social actors, organizational settings, and environmental factors interact. As noted in one of the few articles on professionalism and complexity science, there is a considerable benefit to viewing health organizations as “complex adaptive systems [that operate] in a professional milieu,” rather than as bureaucracies in need of rational administration (Anderson and McDaniel 2000).

**The Medical School as a Complex System**

Building on Anderson and McDaniel’s point, but refocusing on a particular setting, we wish to highlight the medical school as a complex system. In doing so, we wish to situate professionalism within the multitude of learning environments that make up this system. Specifically, we wish to focus on the impact of three such forces—the formal, informal, and hidden curricula—on medical student learning.

There are three major benefits in adopting this framework. First, this framework recasts the formal curriculum from a singular focus to one of three environmental systems, all of which impact on how students learn about and practice professionalism. Second, this shift from a singular to a multiple learning
environment perspective helps to focus on the dynamic interplay that exists among those learning environments. Third, this focus on system dynamics and interactions underscores a point basic to complexity science, namely that medical student learning is more than the sum of the respective system’s parts.

Speaking to our first point, if our goal is to understand (and possibly shape) student learning with respect to professionalism, then it is both counterproductive, and ultimately distorting, to treat the formal curriculum as the sole—or even principal—seat of student learning. There are other learning domains at work that are far more influential, both overall and at certain times and in certain settings, than that formally provided to students in the classroom or at the bedside (Hafferty and Franks 1994; Haidet and Stein 2006). This point stresses the importance of conceptually grounding medical student learning within the full range of experiences that comprises the educational experience.

Our second, and more fundamental, point is that these multiple learning environments function within a web of interdependent relationships, each with its own distinctive identity, yet each dependent upon and shaped by the others. Thus, the learning that takes place in the classroom or at the bedside is shaped by what takes place within the informal social interactions among and between faculty and students as they come together in hallways, cafeteria, and on call rooms—and vice versa. This second point is about the power of interactions.

Our third benefit to adopting a complexity science perspective is that the totality of learning that takes place within the space created by these interactions and intersections is greater than the sum of its constituent parts. Just as the formal curriculum is so much more than the sum of individual courses, medical student learning involves more than stacking what takes place within the formal, informal, and hidden curricula, one on top of the other. This is a point about synergy—a key concept within complexity science.

The fact that medical student learning is both dynamic and interdependent is reflected in a frequently raised “hidden curriculum” question—how medical schools might “do away with” the hidden curriculum, with the question usually phrased so that the hidden curriculum is cast as a singular alternative to the formal curriculum. The very phrasing of this question, while admirable in its recognition of how inconsistent and contradictory messages may negatively impact student professionalism, is incorrect in depicting the hidden curriculum as a thing that can be changed in isolation—and thus changed without altering the content and structure of the other domains of learning. One can no more get rid of the hidden curriculum than one can get rid of protons, both in an absolute sense and in the sense of disrupting the fundamental nature of the overall system. One certainly can target the hidden curriculum, but one must also be willing to track the impact of these changes in the hidden curriculum as they play out within the formal and informal curricula. Furthermore, if one resists changing the formal curriculum after purposefully altering the hidden, then the overall system is placed under an additional stress, thus further distorting overall stu-
dent learning. In sum, medical student learning is multi-dimensional, multi-situational, multi-contextual, and interdependent, and to treat it otherwise is to create a false and misleading picture of the overall educational environment.

**Closing Comments: Limits and Limitations**

Our framing of professionalism as a complex system grounded in social interactions and in the dynamics of multiple learning environments is distressingly incomplete. We have not, for example, even mentioned other occupations (health or otherwise) that are enveloped in their own professionalism movements. Furthermore, we have given short shift to the social dynamics of professionalism itself, including the successful entry of women into medicine and failed efforts to increase the number of underrepresented minorities into physician ranks. We have also neglected to discuss how professionalism, as a social movement, is heavily dependent on the broader socioeconomic and political context in which it evolves. One should thus expect important differences between the professionalism movements in the United Kingdom, Canada, and the United States (Hafferty and Castellani 2006). While one can define a system using different units of analysis (a given medical school class, a given school, all U.S. medical schools), it is also true that professionalism functions as an overall attractor point within medicine. Thus, one can take the professionalism thread, begin to pull, and eventually reach any topic within the medical sciences. Pull a little more and one can move beyond medicine into the broader sociopolitical arena. Everything is interconnected.

This is not, however, a paper “about everything.” Our goal is more targeted. We seek to outline medicine’s professionalism project as a social movement, and in doing so, capture the evolution of this movement as it shuffles and stutters towards recognizing its inherently complex nature. Even the call to recognize professionalism as a complex system is not, in and of itself, a solution to the “problem of professionalism” (at least as defined by medicine). There is no one problem of professionalism, any more than there is one professionalism (Castellani and Hafferty 2006). While professionalism can be depicted as an ideal type for analytical purposes as did Freidson (2001), there is no ideal solution, nor can professionalism adequately be conceptualized from the viewpoint of any one participant in the system. Finally, professionalism does not reside in the motives of individuals (although such motives, particularly the internalization of core professionalism values, are a core element in any understanding of professionalism), or within organizational structures and policies (although structure and process are inherent elements in any professionalism movement). Instead, professionalism exists within the dynamic interplay of system actors, system structures, and broader environmental influences.

Also important to note is that medicine’s modern-day professionalism movement is in its infancy—and continuing with this metaphor, we see medicine just
beginning to recognize the existence of other players and other sandboxes outside its own. We also believe that the modern professionalism movement is shaped by its defining conflict (the dynamic interplay between commercialism versus professionalism) and that this conflict will continue to evolve. We feel quite comfortable (although not sanguine) in concluding that the discourse of professionalism 30 years from now will be a much different discourse than the one we currently face—just as today’s discourse differs from that of the 1970s. Finally, we want to reemphasize that professionalism is not a thing. Rather it is a dynamic. In this respect, professionalism is much like those illusive and evanescent particles in physics that have no mass except in movement. In short, professionalism has no meaningful existence independent of the interactions that give it form and meaning. There is great folly in thinking otherwise.

References


Toward Reducing the Prevalence of Chronic Disease

A life course perspective on health preservation*

Jeremiah A. Barondess

ABSTRACT  Chronic disease is now hyper-endemic in the United States and is the central problem to be addressed in efforts to enhance the health of the American population. Efforts to reduce the prevalence of chronic disease through diminished exposure to risk factors have achieved significant success in recent decades, but most have been expressions of secondary or tertiary prevention. Current knowledge suggests it would be more effective to extend efforts directed at reduction of risk to earlier phases in the biology of chronic diseases, and to maintain them over the life course. This approach lends itself to a health preservation perspective—in other words, to an orientation around protection of the future health of the individual across the lifespan, from preconception to old age. This will require linked efforts of the clinical, public health, and policy communities, together with private-sector collaborators in information management, marketing, and other areas of expertise.

President Emeritus, New York Academy of Medicine, 1216 Fifth Avenue, New York, New York 10029.
E-mail: jbarondess@nyam.org.
*Adapted from the Duncan Clark Lecture, New York Academy of Medicine, March 21, 2007.

Perspectives in Biology and Medicine, volume 51, number 4 (autumn 2008):616–28
© 2008 by The Johns Hopkins University Press
Reducing the prevalence of chronic diseases is the central problem to be addressed in efforts to enhance the health of the American population. Although successes have been achieved, such as reductions in deaths from heart disease and stroke (Thom and Rosamond 2006), significant additional progress will require more effective efforts to apply knowledge about risk factors and their determinants, to interfere with the pathogenetic pathways involved, and to take advantage of the prolonged time courses followed by most chronic diseases before they become clinically evident. Health-protective efforts should be moved to earlier stages in disease biology. We should, in other words, orient ourselves more clearly around efforts to preserve health, to interfere with the erosion of health by chronic disease beginning with the earliest germane events. Success would be marked by health improvement of two kinds: first, by reduction in the prevalence of the clinical manifestations of chronic disease or delay in the appearance of clinical manifestations until progressively later in the life course, the so-called “compression of morbidity” formulation (Fries 1980); second, at a more fundamental level, by prevention or slowing of the rate of progression of the underlying pathologic processes.

Over the past 100 years or so, as a result of the development of science-based public health and clinical practice, a real ability to preserve health has emerged, and life expectancy in the developed world has increased at an extraordinary rate. From the beginning of the 20th century to the present, life expectancy in the United States has been extended by some 30 years. It has been estimated that about 75% of the overall gain reflects various public health measures, and about 25% has been due to clinical interventions (CDC 1994). Early efforts were concerned with the dominant disease pattern of the time, acute infections, and were reflected especially in reductions in childhood mortality, while most of the gains in the past 50 years have occurred toward the end of life (Fried 2000), with clinical advances contributing to an increasing degree, for example with relation to coronary heart disease (Hunink et al. 1997).

These trends have resulted in a remarkable paradox, namely a sharp increase in life expectancy and the parallel emergence of hyper-endemic chronic disease as the major morbidity pattern, currently responsible for some 80% of deaths in the United States (Mokdad et al. 2004). Chronic disease presents a set of challenges and opportunities strikingly different from those of acute infections. As Fries has noted (Fries 1980, 2005), infectious diseases tend to have unitary causes and relatively short courses. They are frequently preventable through immunization or other public health modalities, and in many instances are susceptible of cure. Chronic diseases, on the other hand, tend to have multifactorial causes and are characterized by risk factors as well as preclinical courses that in general last for years, and are usually not amenable to cure. In addition, because of the ubiquity of chronic disease, individuals do not vary so much in terms of whether they have particular chronic diseases, but rather in terms of the rate at which their diseases progress: every individual has the likelihood of gradually progressive silent
atherosclerosis, as well as an increasing statistical possibility of malignant disease, osteoarthritis, and degeneration of articular cartilages (Fries 2005). Finally, chronic disease risk is more tightly tied to personal behaviors than are most infections. For all these reasons, in most instances chronic disease prevention is difficult to achieve. Even though lung cancer and acute myocardial infarction can be prevented in many—ultimately perhaps most—instances, the complexities of dealing effectively with the relevant behavioral and environmental forces make delay or mitigation of clinical expression of most chronic diseases through reduction of risk a more realistic approach. Most current efforts reflect secondary or tertiary prevention, often applied only after years of health-adverse personal behaviors and environmental factors have threatened health, and too often in the face of overt or longstanding clinically inapparent but progressive disease. As a result, much of our clinical effort is now directed at late-stage expressions of atheromatosis, cancer, chronic lung disease, diabetes, and musculoskeletal disorders. Greater effectiveness would likely be achieved through linked age-appropriate interventions applied across the life course, organized conceptually as efforts to move closer to primary prevention—in other words, to preserve health.

The health preservation paradigm derives from a number of basic concepts:

- Each individual possesses at birth a certain “dose,” or quantum, of health expectancy, determined by the health of the parents prior to conception, by genomic factors, and by environmental forces, especially the quality of the intrauterine environment in which the fetus develops, determined in turn largely by the health of the mother and her health-related behaviors during the pregnancy.
- The health quantum is affected over the life course by the individual’s biologic, psychologic, and behavioral characteristics as they interact with environmental, socioeconomic, and educational factors, and with the timeliness and quality of health care received. These determinants expose the individual from birth through old age to a variety of modifiers, some of which may act to preserve or even enhance health, while others have an erosive effect. The risks of encountering erosive forces and their nature, dimensions, and effects all vary with life stages. In addition, mechanisms of effect vary, and disease outcomes may be long delayed. For example, increasing evidence suggests that the health impacts of some adverse events occurring during critical periods of development may reach clinical expression only years later; low birth weight and early childhood growth rates, and their impacts on adult cardiovascular and metabolic disease are powerful examples (Forsén et al. 2000; Singhal and Lucas 2004). Other health-adverse forces, appearing later and acting through continuing environmental and behavioral exposures, are superimposed on such early determinants. In some instances, health-erosive mechanisms are synergistic, coalescing toward final common pathways,
for example the combined effects on atherogenesis of smoking, dyslipidemia, and diabetes (Biondi-Zoccai et al. 2003; Fuster and Gotto 2000). The net erosive load at any time in the life trajectory, expressed through varieties of pathologic processes, reflects the accumulated impacts of prior health-adverse exposures and events.

• Important chronic diseases may arise very early and exhibit long latency, progressing for long periods of time before reaching clinical expression, and may therefore provide multiple opportunities for initiation of effective interventions. This is strikingly true, for example, in the case of atheromatous disease, in regard to which there is evidence that the lesions may arise very early in life: human fetuses have been found to display aortic fatty streak formation, greatly enhanced by maternal hypercholes-
terolemia (Napoli et al. 1997); early atheromas have been identified in the aortas and coronary arteries of six-year-olds in the Bogalusa Heart Study in Louisiana (Berenson 2002); and established atheromatosis was found to be common in American combat casualties in Korea and Vietnam, largely men in their late teens or early twenties selected for military service on the basis of their apparent good health (Enos, Beyer, and Holmes 1955; McNamara et al. 1971). One implication of such observations has to do with our conception of health, traditionally framed as the absence of overt, clinically expressed disease. The health preservation paradigm suggests that if we are to develop health-protective programs that relate more closely to the biology of disease, those ideas should be revised to include clinical latency.

A clearer and more vigorous age-sensitive orientation of efforts to preserve health across the life course would allow a more coherent and linear approach to the way determinants and precursors of chronic disease arise throughout life, coapt in their impacts, erode the health quantum, and contribute to ill health and reduction in life expectancy. These considerations begin prior to conception. Higher maternal age and a history of smoking, for example, have been associated with higher newborn systolic pressure, and both early and late maternal age are associated with low birth weight (Oken et al. 2005; Yang, Greenland, and Flanders 2000). Maternal pre-pregnancy BMI is of potential importance as a risk factor for gestational diabetes, which is associated in turn with high birth weight and consequent adolescent overweight in the offspring, a predictor of adult obesity, and fetal risk and adverse pregnancy outcomes have been linked to preconception maternal hypertension and use of alcohol, tobacco, and illicit drugs (Chattenoud et al. 1998; Gillman et al. 2003; Rosenberg et al. 2005; Salsberry and Reagan 2005). Because of especially high risks to the fetus during the first weeks of gestation, pre-pregnancy counseling is particularly important. For example, for optimal reduction of the risk of neural tube defects, folic acid supplementation should start at least three months before conception (Werler, Shapiro, and
Mitchell 1993). The contributions of preconception paternal health are less clear, although there is a significant association between advancing paternal age and the risk of low birth weight, congenital anomalies, autism, Down syndrome, and perhaps schizophrenia, and paternal HIV and hepatitis C infection threaten both prospective mother and fetus (Fisch et al. 2003; Friedman 1981; Malaspina et al. 2001; Reichenberg et al. 2006; Reichman and Teitler 2006; Savitz, Shwingl, and Keels 1991). Overall, while pre-pregnancy and inter-pregnancy counseling are of clear importance, only 16% of obstetricians/gynecologists or family physicians provide this preconception care to most patients (Henderson, Weisman, and Grason 2002).

During pregnancy, maternal smoking, heavy use of alcohol, and exposure to illicit drugs are associated with adverse fetal outcomes (Huestis and Choo 2002; Maconochie et al. 2007). Low birth weight, a reflection of an adverse intrauterine environment and/or shortened gestation, may result from multiple factors and appears to be a significant determinant of health over the life course; low rates of fetal growth are associated with higher risk and higher mortality from cardiovascular disease in adult life, as well as higher rates of type 2 diabetes, hypertension, the metabolic syndrome, and adult obesity (Barker 1993; Forsén et al. 2000).

In the neonate and infant, low growth rates as well as small body size at one year are predictors of coronary heart disease in adult life (Forsén et al. 2004). Paradoxically, underweight neonates tend to bounce back quickly and ultimately suffer an increased risk of overweight amounting to some 25 to 30% at age seven (Chen et al. 2006). Unusually rapid postnatal growth in low birth weight babies appears to program later cardiovascular risk, insulin resistance, and obesity (Forsén et al. 2000; Hovi et al. 2007), suggesting that very early events are important in determining organ capacities and in setting a number of chronic disease–related metabolic clocks in the fetus or neonate.

Breastfeeding, especially more prolonged breastfeeding, appears to be associated with a reduced incidence of childhood obesity in a dose-response manner (Harder et al. 2005; Mayer-Davis et al. 2006). Current infant feeding patterns in the United States show a strong socioeconomic gradient. For example, only 4% of infants in the WIC program remain exclusively breastfed at six months of age, compared with 17% of nonparticipants, a potential contributing factor in socioeconomic disparities in overweight and obesity in adult life (Gidding et al. 2006). In addition, compared with parents who bottle feed, mothers who breastfeed appear to allow infants to take a more active role in controlling intake, and this in turn may promote later feeding practices that can foster better self-regulation of energy intake (Taveras et al. 2004). But distortions in early eating patterns abound; for example, some 28% of one-year-olds have consumed sweetened beverages, and at age two years French fried potatoes are the most commonly consumed vegetable (Gidding et al. 2006). Training infants and children to seek quality nutrient intake and to avoid excess calories is important in the development...
of health-protective eating patterns, since childhood feeding behaviors appear to be major determinants of eating patterns later in life (Whitaker et al. 1997).

In addition to these early determinants of adult chronic disease, increasing evidence indicates that important disease appears in childhood. The Bogalusa Heart Study identified fatty streaks and fibrous plaques in the aortas and coronary arteries of children and young adults. The lesions not only appeared early, but they were significantly more marked in older children and were correlated with obesity, high systolic blood pressure, and high serum triglyceride and low-density lipoprotein cholesterol concentrations, risk factors that tended to cluster in individuals (Berenson 2002; Li et al. 2003). As noted above, the early appearance of atheromatous disease has been underlined by the demonstration of established and often severe atheromatosis in young U.S. combat fatalities in Korea and Vietnam. In light of the known risk factors, the correlations observed to date, and the demonstrated progression of atheromatous lesions in the young, the strong inference is that heart disease prevention should be a lifelong effort, beginning in childhood (Graziano 1998). The importance of ongoing monitoring is plain, since obesity, elevated blood pressure, and dyslipidemia tend to track over time from childhood into adult life, and since lifestyle choices influence these risk factors. Parental health literacy relative to these factors is crucial.

Adolescence is characterized by behaviors with major implications for future health. The 2005 Youth Risk Behavior Survey, covering students in grades 9 to 12, found that 23% were cigarette smokers, 13% were overweight, 43% were current alcohol users, and 25% had drunk heavily on at least one of the 30 days preceding the survey (Eaton et al. 2006). Some 20% were characterized as current marijuana users, more than 3% as current cocaine users, and 2% had used a needle to inject an illegal drug into themselves on at least one occasion. A third were sexually active, and one-third of those reported that neither they nor their partner had used a condom during last intercourse. Most were not eating recommended levels of key foods, and sedentary behaviors were common; for example, 37% were watching three or more hours of television on an average school day. Overall, it is fair to say that effective individual or parental concern for optimization of health is already widely compromised in the teen years, and that clinical and public health efforts in the interest of future health require bolstering.

In the years immediately following adolescence, individuals emerge from pediatric care and begin an irregular interface with adult clinical caregivers. Health insurance coverage is uneven in these years, as young people leave the umbrella of parental coverage, often see little need to purchase health insurance, and begin to enter the labor force at levels that may not include employer coverage. The health risks of this period of life are not trivial, and a number have important implications for chronic disease later, especially health-adverse behaviors. Screening for current disease and monitoring for potential precursors of later health erosion, for example, hyperlipidemia, obesity, type 2 diabetes mellio-
tus, and hypertension, are important, but opportunities for both become fragmented during this life stage.

In adulthood, the major health-protective issues relate to immunizations, reduction of health-adverse behaviors, screening for disease precursors or silent disease, and management of clinically emergent disorders, especially cancer, ischemic heart disease, hypertension, diabetes mellitus, obesity, depression, and musculoskeletal disease. The 2006 Health Behaviors Surveillance Survey (CDC 2006a) indicated that about 20% of American adults still smoke cigarettes, only a quarter engage in vigorous physical activity three or more times weekly, and two-thirds are overweight or obese. More intense efforts to mitigate health-adverse behaviors are justified by data linking improved health and survival to cessation of smoking, control of body weight, and increased physical activity even after long periods of neglect (Anthonisen et al. 2005; Wang et al. 2002). With regard to screening, the Survey found substantial underuse: the national median figure among adults for never having had a sigmoidoscopy or colonoscopy was 43%, and while most women had had a cervical pap smear at some time, some 24% of women over the age of 40 had never been undergone breast cancer screening. Adult immunizations also lagged: among individuals over the age of 65, about a third lacked influenza vaccination in the current year, and the same proportion had never received pneumococcal vaccine.

Among the elderly, erosion of health accelerates as multiple diseases reach clinical expression and functional importance and are superimposed on physiologic senescence (Fries 2005). Subjective well-being is further compromised as exposures to adverse social circumstances, including financial pressures and the inevitable narrowing of social networks, are added, and as chronic feelings of stress, depression, loss of autonomy, and reduced social utility further compromise functionality. In addition, physical activity is often reduced due to social isolation, physical limitations, and inadequate opportunities in the environment. The importance of an active stance relative to the incapacies of old age is often lost on both informal caregivers and the clinical community. Physical activity, cognitive involvement, and social engagement are generally feasible, frequently produce sharp enhancements of mood, and contribute to vigor and mental acuity. Innovations such as the Experience Corps, in which inner-city elderly are assigned, after indoctrination and training, to fixed and significant responsibilities in elementary schools in their areas, have demonstrated sharp enhancements in subjective well-being and levels of physical activity, as well as reduced hours of television watching and increased feelings of usefulness and social engagement (Fried et al. 2004).

* * *

Review of the health risks emergent in the various life stages creates the image of a crescendo of health impacts, of an accumulating net erosive effect on the health expectations possessed by each individual at birth. The accumulated risks
are summative in their impacts and would be more effectively approached by considering them from a life course perspective.

The concept of a life course approach to health preservation has significant harmonics with the compression of morbidity construct first advanced by Fries in 1980. In essence, this idea focuses on reduction of the interval between the appearance of clinically evident disease and death, with consequent preservation of function and reduction in the necessity for clinical care. The compression of morbidity hypothesis, like the health preservation paradigm, revolves around the fact that our dominant health problem is chronic disease.

Progress is occurring with regard to compression of morbidity. Largely due to the application of effective screening procedures, clinical advances, increasing emphasis on risk factors, and rising educational levels, postponement of clinically evident chronic disease and reductions in disability have shown progressive widening over time compared with control groups in some studies (Fries 2005; Manton and Gu 2001; Willcox et al. 2006). An important aspect of this malleability relates to the fact that some 40% of deaths in the United States are currently associated with potentially modifiable health-adverse personal behaviors (Mokdad et al. 2004), suggesting that efforts to engage individuals more effectively in managing their own health is critically important. A key corollary of this approach going forward will be marketing the idea of increased personal responsibility for the management of health preservation.

In a linked set of considerations, there are now convincing studies indicating that even the limitations imposed by universal phenomena of senescence in tissues and in organ function, the factors that ultimately determine the biologic limits of life expectancy, are also modifiable to a significant degree, a phenomenon sometimes referred to as the plasticity of aging (Fries 2005). Cardiac reserve, glucose tolerance, intelligence test performance and memory, osteoporosis, strength and physical endurance, pulmonary reserve, and reaction time are modifiable by the individual, even at advanced age. Modification in most instances results from training and practice in the specific faculty; there is relatively little crossover from training in one attribute to another. Compression of morbidity and the plasticity of aging are related concepts; the ideal outcome would be compression of morbidity inside a life span closer to the biological limit, both functionally and chronologically.

In addition to appropriate management of identified disease at any age, powerful further impacts on health and its preservation would result from earlier, more vigorous, and longitudinally maintained management of risk factors, disease precursors, and health-adverse behaviors, and wider application of screening procedures, an orientation that would be facilitated through emerging electronic records and management systems. In the case of atheroma, for example, efforts in childhood toward the avoidance of overweight and excessive saturated fats and
salt in the diet, as well as the adoption of good exercise habits, should be supplemented in adolescence by intensive efforts to deter the adoption of health-adverse behaviors, and periodic screening for hypertension, diabetes, and dyslipidemia should be started or continued. This array of efforts should be maintained throughout the adult years and old age, with special attention to groups at particularly high risk, notably the poor and racial and ethnic minorities. A key conceptual shift is needed, more clearly framing the health of the individual over the long term as an active concern of the current caregiver, beyond dealing with immediate clinical or developmental issues. For example, it is increasingly clear that the health of the emergent adult is to a substantial degree in the hands of the pediatrician, and in fact, that the future health of the individual is in every instance to an important degree in the hands of the current caregiver. In each clinical encounter, the clinician is confronting the health needs not only of the individual as he or she is presenting at the time, but the health of that individual in every era of his or her subsequent life. The inculcation of such attitudes among physicians should begin during undergraduate medical education and should be emphasized and extended during residency training in every clinical specialty, interwoven with the disease-oriented material that is the primary focus of educational efforts during these years, and should be recurrently introduced, especially by clinical teachers. Analogous emphases and linkages are needed in education, training, and practice in public health, nursing, social work, and the other health professions.

A corollary of enhanced and more effective clinical and public health efforts is a more muscular individual sense of responsibility for the protection of health. Awareness of the need for personal efforts has heightened in recent years: witness, for example, the decrease in cigarette smoking among adults, the widespread adoption of running and other exercise programs, and widening concerns about fat and salt consumption (CDC 2006a). To help in promoting active individual stewardship of health, a serious effort toward enhancing health literacy is needed. The health sector will need to join with others, and should turn, for example, to the corporate community for alliances relating to the work force and retirees, a very large sector of the adult population (Okie 2007). In addition, we should link with experts in marketing, advertising, and electronic information management in efforts to go beyond addressing the needs of individuals seeking information relevant to specific health issues. A broader approach would introduce health-protective information in a wide variety of settings with no necessary relation to immediate health or disease concerns. Presenting reminders concerning screening procedures, diet, exercise, smoking, and immunizations in supermarkets, shopping malls, coffee shops, and other venues frequented by large numbers of individuals, and presenting such information in brief, attractive, tested formats on monitor screens or on health kiosks, and doing so persistently over long periods of time, would blend modern electronic communication techniques with a massive national need. Making common cause with experts in
advertising and marketing makes good sense: the health professions can identify and validate the issues, behaviors, and interventions to be “sold,” but providing the message in an effective manner in ordinary life contexts requires the special talents of people familiar with the techniques needed.

Broadening support in the policy community for health preservation activities and related research will depend to a significant degree on quantifying the likely enhancement of population health and the reduction in health-care costs that would result, as well as developing more clearly the opportunity costs of lives that are shorter and sicker than they have to be. Such costs are measured in blunted job and career prospects, reduced productivity, and lost wages and tax revenues. The pattern of current health-related expenditures is markedly unbalanced and should be revised to encourage more vigorous efforts to protect health. Currently some 83% of the national investment in health goes for personal health care, with prevention in the remaining 17%, along with administrative costs, research, and physical infrastructure (CDC 2006). The health professions should proselytize for tax and conditional spending incentives by government to encourage private-sector health-protective activities (Gostin, Boufford, and Martinez 2004). Changes that would encourage individuals to pursue health-positive behaviors, perhaps in the form of insurance premium reductions or other financial rewards, should be promoted and extended, a kind of pay for performance at the individual level. In addition, policy efforts should be developed with clearer and stronger links to upstream health determinants, including socioeconomic and racial and ethnic forces such as disparities in health-care access and quality, housing quality, segregation, and job training, as well as the minimum wage.

Important gradients are to be considered in promoting a stronger orientation around health, including difficulties in developing consensus on the relevant measures of population health, the need to make clinical and other silos more permeable, resistance to reallocation of resources, and difficulty in allocating significant resources to upstream determinants (Kindig 2006). Federal and state efforts to blunt the obesity epidemic, especially in relation to the eating patterns of children, have demonstrated the complexity of such issues (Mello, Studdert, and Brennan 2007). Gerberding (2005) has noted as additional problems the lack of near-term direct financial benefits in the protection of health, the coordination of multiple funding streams that integrated health promotion programs require, diffusion of accountability for health, the lack of a single unifying message, counter-marketing by other powerful agents, and social preferences. A more proximate issue is that the structure and organization of the personal clinical care system make it difficult for the public health enterprise to rely on the private sector to deliver preventive and therapeutic services essential to protecting community health, a health services issue of the first magnitude.

One of the most powerful factors in shaping funding priorities is largely lacking in the case of health, namely a vocal and affected advocacy group. Unlike systemic lupus, end-stage kidney disease, or Alzheimer’s dementia, for example, no
analogous group in the general population has emerged to champion health and its protection. That is a striking paradox—no group advocating for the thing everyone wants. Health in our society is an assumed good. Its absence—disease—is heavily and appropriately pursued, but its protection is seen as a lesser priority.

We perhaps need a new definition of preventive medicine; better than that, we should abandon the term in favor of a more comprehensive stance organized around the preservation of health. For a long time we have accepted health as the norm and disease as the aberrancy, and our funding patterns and priorities reflect those perceptions. It may be time to modify this framework: disease, in a real sense, and because of its ubiquity, might be thought of as the “norm,” and health, or at least good health extending somewhere close to the biologic limits of human life, as the aberrancy. Such a changed orientation might facilitate rearranging our priorities and our advocacy.

**References**


Toward Reducing the Prevalence of Chronic Disease


Essay Review

**Babies by (Intelligent) Design?**

**MARY B. MAHOWALD**

**ABSTRACT** Advances in reproductive technology and genetic interventions raise questions about the possibility of using these procedures to promote the birth of children with socially advantageous conditions. In *Babies by Design*, Ronald M. Green supports this goal and accuses its opponents of a “status quo bias.” Unfortunately, some of Green’s own arguments also show a status quo bias. Moreover, although he attempts to avoid the thorny issue of the moral status of human embryos, he implicitly takes a stand on it by endorsing prenatal interventions that inevitably entail the creation and loss of some human embryos. This essay identifies these and other flaws in Green’s account.

Ronald M. Green probably rejects the theory of intelligent design as an alternative to evolution, but he clearly believes that using our intelligence to produce babies with superior traits is acceptable, possibly even commendable—at least in some cases. The title of his recent book, *Babies by Design*, could thus be readily expanded to *Babies by Intelligent Design*—without altering its content. Green is professor of ethics at Dartmouth, founding director of the Office of Genome Ethics at the National Human Genome Research Institute, and former president of the Society of Christian Ethics. He also serves as chair of the Ethics Advisory Board of Advanced Cell Technology, Inc.


At some point in their lives, most adults would like to have children. Typically, they plan to fulfill their desire through sexual intercourse, fertilization, and gestation within the body of the woman. We might call this the traditional recipe for making babies. For a substantial number of people, however, crucial ingredients for the traditional recipe are unavailable or undesired, and an alternative recipe may be followed. Some people choose instead to remain childless and others pursue adoption, but most potential parents still want to have a child who is biologically related to at least one of them, whether through genetics, gestation, or both. Reproductive endocrinologists refer to the various procedures by which to accomplish this through a veritable cookbook of acronyms, such as IVF (in vitro fertilization), ET (embryo transfer), ED (embryo donation), OD (ovum donation), AID (artificial insemination by donor), AIH (artificial insemination by husband), ICSI (intracytoplasmic sperm injection), GIFT (gamete interfallopian tube transfer), ZIFT (zygote interfallopian tube transfer), and PGD (preimplantation genetic diagnosis). Usually, the selected recipe calls for a number of the procedures identified by these acronyms.

Because the desire to have biologically related children is a natural and fulfilling human tendency, engaging in the ordinary means for doing so is not only defensible but commendable. Some people consider parenthood a privilege—even when the recipe followed is the traditional one. Others, however, regard it as a basic human right that should be guaranteed through provision of whatever technological means are necessary to its fulfillment. For most people, the pursuit of biological parenthood is questioned only when a nontraditional recipe is followed. The more closely the alternative resembles the traditional recipe, the less the controversy. For example, IVF with gametes of married partners, with one of them gestating and giving birth to the desired infant, is broadly accepted and practiced. Using the genetic or gestational contributions of third parties is more controversial.

Another source of controversy regarding nontraditional recipes is that they usually require the creation and loss of some human embryos. Deliberate destruction of these embryos is, for some, morally equivalent to killingborn human beings. Others find the destruction of in vitro embryos unproblematic because they do not consider embryos persons who, as such, have a right to life.

Among those for whom loss of embryos is morally acceptable, concerns still surface when a nontraditional recipe is intended to produce not just a baby, but one who has traits unrelated to health. As with sex selection, the selection of other traits evokes graver moral concern than efforts to avoid disease or disability. For example, prenatal interventions to promote social advantages such as superior intelligence or appearance are troubling to many people. Most troubling are efforts to ensure that a child has the same socially disadvantageous conditions that parents themselves possess (for example, dwarfism or deafness). In Babies by Design, Green analyzes both types of intervention.
While developing his views about “The Ethics of Genetic Choice” (his subtitle), Green explicitly distances himself from the “more natural bioethics” of the Bush administration, its “conservative religious base,” and the President’s Council of Bioethics. He particularly targets the latter. “Under the direction of the conservative bio ethicist Leon Kass,” he writes, “the President’s Council has issued a series of reports largely seconding the president’s views and questioning whether biomedicine had not already gone too far” (p. 3). The Council’s positions, for Green, proceed from a “status quo bias” that has triggered “biomedical and bioethical retreat” from scientific and medical progress.

In his introduction, Green summarizes the framework from which he resists the alleged retreat and “status quo bias”: “I disagree profoundly with this conservative direction. . . . I have dedicated myself to supporting new technologies aimed at assisting people faced with infertility, helping us understand and prevent the causes of birth defects, and using human embryo research to develop new approaches for tissue regeneration and organ replacement” (p. 4). Green believes “we should begin considering deliberate interventions in our own and our children’s genetic makeup—to both prevent disease and enhance human life” (p. 4).

Because the question of the moral status of human embryos is interminably controversial and apparently irresolvable, it is understandable that Green avoids this issue. Nonetheless, his general endorsement of prenatal technologies for enhancement necessarily implies approval of the destruction or loss of embryos that occurs through these interventions. Despite his attempt to avoid the issue, therefore, Green implicitly takes a stand on the moral status of embryos: they do not have the right to life that is attributed, both legally and morally, to persons as such.

In chapter 1, Green introduces the topic of genetic enhancement through an analogy with steroid use by athletes. The two reasons he cites for opposing use of steroids to improve athletic performance involve safety and fairness. Green worries about safety mainly because the rewards of athletic success are highly seductive to potential steroid users. On the unfairness their use may bring to sports, however, he is ambivalent. Fairness, he observes, is already affronted in competitive sports by the genetic lottery through which some people are athletically advantaged even before they are born. Although he says he himself was “athletically challenged” in grade school (he couldn’t catch or hit baseballs), he acknowledges that this is hardly a problem for the professor of bioethics that he has become. “In athletics,” he says, “life’s not fair,” but those who are challenged in that regard can generally be successful in other ways (p. 28). Ironically, Green’s rationale that “life’s not fair” suggests the same status quo bias that he challenges elsewhere.

Green’s second chapter—“How Do We Do It?”—is an impressive example of the interdisciplinarity that is indispensable to credibility in bioethics. His account relies heavily on extensive interviews and correspondence with scientists whose research has contributed to pivotal advances in genetics. Of particular interest is
the work of Nobel laureate Mario Capecchi, who developed the technology of homologous recombination. This technology, Green says, “could change our world,” because it makes gene alteration and site-specific gene targeting possible through utilization of the body’s own gene repair mechanisms (p. 33). While explaining this and other relevant research accurately and accessibly, Green inter­sperses some concerns or reservations articulated by scientists themselves about ethical issues raised by their research. Capecchi, for example, told Green of his uneasiness about actively modifying the human genome. As yet, Capecchi said, “we’re not close enough to understanding the issues to make wise decisions or predict the outcomes” (p. 33).

In chapter 3, Green considers two distinctions, one between therapy and enhancement, and the other between somatic cell and germline interventions. Therapy, or treatment that restores an individual’s health, is routinely viewed as appropriate, even commendable, whereas moral justification for enhancement beyond that level is dubious. Similarly, somatic cell interventions that affect a single individual are more easily justified than germline interventions, because the latter affect the individual’s posterity, who could not have consented to the intervention in the first place. By combining the distinctions between therapy and enhancement, and between somatic cell interventions and germline interventions, we can create a spectrum of options that range from somatic cell treatment, which is easily justified, to germline interventions for enhancement, which is least justifiable, and perhaps condemnable. Between these is somatic cell enhancement, which is closer to the defensible end of the line, and therapeutic gene modification, which is closer to the other end. Between therapy and enhancement, however, there are different degrees of impact regarding both somatic cell and germline interventions. Whether this impact is based on biology or societal structure, some interventions provide effective therapy for very disadvantageous conditions or less disadvantageous conditions, and the same interventions may promote relatively advantageous or very advantageous conditions for particular individuals.

Several additional factors make line drawing problematic with regard to genetic interventions. First, some somatic cell conditions can only be effectively treated through therapies that involve gene modification. Second, combinations of different conditions, coupled with how “normal range” is defined and the life choices of individuals, influence where they fall along the spectrum. Third, the range defined as “normal” is a legitimate matter of debate; the same condition, such as intelligence or athletic ability, can occur anywhere along the spectrum, whether within a “normal range” or not.

In his critique of sharp line-drawing in genetics, Green proposes a role for “prevention” as an intermediary between the two distinctions. Through widespread endorsement of vaccine use, he says, society already regards some types of prevention as therapeutic. These preventions, he says, are a kind of enhancement because their goal “is to surpass normal levels of functioning now to prevent them
from ever occurring” (p. 61). That prevention through genetic manipulation may also be seen as therapeutic is clear from recent work on a potential cure for sickle-cell anemia. Through research with a humanized mouse model, Rudolf Jaenisch and his colleagues have demonstrated “proof of principle” that genetic reprogramming can prevent the onset of this condition (Hanna et al. 2007).

In chapter 4, Green begins his examination of the medical as well as social risks and challenges of genetic interventions by discussing “Sisters,” a science-fiction story by Greg Bear. The story takes place 60 years from now in a world where “genetic engineering” of children has become commonplace. The main character, Letitia, lacks the superior traits that the majority of her peers possess and blames her parents for having “refused to use the new genetic technology before she was born” (p. 82). As events unfold, however, Letitia’s supposedly advantaged classmates develop fatal illnesses, while she enjoys continuing health. While using this story to show how methods of gene therapy now available may cause serious but unpredictable harms, Green considers the possible reversibility of genetic changes as a means by which to achieve benefits without such risks. Capecchi’s research involving gene modification in mice is cited as a strong indicator of this possibility in humans.

The title of chapter 5, “Parents: Guardians or Gardeners,” aptly suggests Green’s ambivalence about whether potential parents should follow whatever recipe they choose to produce the offspring they desire to raise. While repeating his criticisms of the conservative positions of the President’s Council on Bioethics and its alleged status quo bias, Green acknowledges that he cannot now offer unarguable counter-positions on prenatal genetic modification. He fully expects, however, that if we move in that direction, we will know enough to answer the relevant ethical questions in a matter of years. These answers, he believes, need to be pursued now in order to show us “what kind of preparations are likely to maximize their values to children and families” (p. 121).

Like many bioethicists, Green’s main worry about prenatal genetic interventions involves the principle of distributive justice or fairness. As he puts it in chapter 6, such techniques could create a “genobility”—a class of individuals who are genetically programmed to enjoy advantages that are missing from the lives of those who are not similarly programmed. While attempting to calm fears about this potential genobility, he discusses the difference principle of John Rawls. This is the principle by which Rawls proposed that the advantages enjoyed by a limited number of people should be permitted only if the disadvantages of others are simultaneously reduced. Green regards the difference principle as outdated, because it was based on the notion that genetic differences are unchangeable, and now we know otherwise. If Rawls had known that genetic differences were changeable, Green believes, he would have been open to interventions through which genes may be redistributed in the interest of fairness. He even sees Rawls’s theory as disposed “to rehabilitate the concept of eugenics” that Green himself attempts to rehabilitate in chapter 7 (p. 153).
Since Green is a religious ethicist, it is not surprising that his discussion of eugenics involves a critique of the notion that efforts to produce babies by design is equivalent to “playing God.” In this regard he cites a question posed by William Hurlbut, a physician member of the President’s Council on Bioethics: “Is the world good either by the benevolent purposes of a creator or by the harmonious balance of a subtle evolutionary force or both” (p. 181). To Green, Hurlbut’s question encapsulates the view that “both evolution and God have conspired to perfect the human genome,” and this belief makes it seem “morally and spiritually dangerous to tamper with it” (p. 185). Because the human genome is not perfect, Green embraces an alternative religious view—that human beings are co-creators with God, each charged with the role of making the whole of creation better. He thus views human beings as fulfilling their God-given role by pursuing the perfection of the human genome through genetic modification.

Green’s concluding chapter proposes four policy guidelines that would be acceptable to many of those who might disagree with his relatively strong endorsement of prenatal genetic modification. By using the term “guidelines,” he avoids the force of regulatory language and allows for exceptions that may or should be permitted. Whether this latitude is desirable is of course debatable. In addition, the generality of key terms in his guidelines allows for different and conflicting interpretations. The first guideline, for example, says that “genetic interventions should always be aimed at what is reasonably in the child’s best interests” (p. 216). But what counts as “reasonable” to some may be unreasonable to others, and what counts as the “best interests” of someone is often impossible to ascertain with any definitiveness. His second guideline asserts that “genetic interventions should be almost as safe as natural reproduction” (p. 218). But “almost” for some is far from “almost” for others, and whose safety is the focus of concern remains unspecified. Even if safety is defined only in medical terms, the risks of intervention are different for the parties involved—not only the potential mother, potential father, and potential child or children, but also, in some cases, the gamete providers or gestators.

The remaining guidelines are apparently intended to limit the injustice or inequity that prenatal genetic interventions may introduce. Green’s third guideline says “we should avoid and discourage interventions that confer only positional advantage” (p. 223). By “positional advantage” he means an advantage that places those with whom one is playing the game of life at disadvantage. It is difficult to see, however, how mere “discouragement” could effectively prevent the positional advantage that those who have access to prenatal genetic modification might obtain for their children. In the fourth guideline, Green avers that “genetic interventions should not reinforce or increase unjust inequality and discrimination, economic inequality, or racism” (p. 225). He thus articulates a sentiment with which few would disagree, even while disagreeing about whether specific inequalities are unjust.
Not surprisingly, Green’s interpretation of his own guidelines supports his overall endorsement of prenatal genetic modifications. Although my own interpretation of his guidelines would not support that endorsement, it would take more than this brief essay to develop an adequate explanation and defense of this interpretation. Suffice it to say that Green’s neglect of three major issues leaves me dissatisfied. First is the unavoidable and morally relevant link between prenatal or preimplantation interventions and the destruction of human embryos; as indicated earlier, Green does not address this issue but implicitly takes a stand on it without acknowledging or defending his stand. Second, he doesn’t adequately identify or address the different and potentially inequitable medical and social impacts of the procedures he supports on those who are immediately affected. And third, Green’s proposed guidelines include no practical suggestions on how to avoid the exacerbations of discrimination against people with disabilities that are, I think, bound to occur if his recommendations regarding the creation of babies by design are followed. In a pluralistic society such as ours, this last neglect is regrettable even if it is unavoidable.

Finally, Green’s repeated criticisms of a so-called status quo bias is at odds with some of the positions he supports on grounds that society is already engaging in comparable activity without challenging its moral legitimacy. Green doesn’t seem to recognize a possible bias on his own part when he invokes the status quo in support of some of his positions. Rigid adherence to the status quo can, and sometimes does, thwart progress, but whether it does so depends on what is meant by “progress.” Presumably, Green would agree that what has already proved effective has an a priori validity when compared with possibilities whose effectiveness has not been tested.

Despite my serious reservations about some of Green’s positions and arguments, I believe his defense of interventions to produce babies by (intelligent) design deserves a careful reading. Not only is it well-informed and well-written; it is also generously interspersed with discussions of relevant popular sources, mainly science fiction, that make it highly readable for those who are otherwise unfamiliar with the topics addressed. Most readers of *Perspectives* are likely to join both Green and me in approval of his guidelines, even while they may disagree with both of us in their interpretation and application to genetic interventions to ensure the birth of advantaged offspring.

**Reference**

Essay Review

INTEGRATING EVOLUTION AND DEVELOPMENT

from theory to practice*

EHUD LAMM AND EVA JABLONKA

ABSTRACT This volume joins a growing list of books, monographs, and proceedings from scientific meetings that attempt to consolidate the wide spectrum of approaches emphasizing the role of development in evolution into a coherent and productive synthesis, often called evo-devo. Evo-devo is seen as a replacement or amendment of the modern synthesis that has dominated the field of evolution since the 1940s and which, as even its architects confessed, was fundamentally incomplete because development remained outside its theoretical framework (Mayr and Provine 1980). As the volume attests, there is now a strong feeling that the time is ripe for the consolidation of evo-devo, and that the field is mature enough so that mapping the theoretical terrain and experimental approaches is both feasible and scientifically productive. Now is an appropriate time to try to weave the strands of reasoning leading to the developmental perspective and offer a synthesis.

INTEGRATING EVOLUTION AND DEVELOPMENT: From Theory to Practice is a collection of papers by central researchers in the field of evo-devo, ranging from conceptual high-level surveys to applications of the evo-devo perspective to cur-
rent debates. The first two papers (Laubichler and Maienschein, and Callebaut, Müller, and Newman) frame the subject and position it relative to other contending attempts to relate development and evolution. The last two papers on evolutionary psychology (Griffiths) and the evolution of culture (Wimsatt and Griesemer) use the evo-devo perspective to shed light on contentious current debates (specifically, evolutionary psychology and the meme theory of cultural evolution). These papers employ a set of concepts central to the evo-devo discussion that have been the focus of the work of the authors for many years, among them homology (Griffiths), generative entrenchment (Wimsatt), and the notion of reproducer as an alternative to Dawkins’s replicator/vehicle model (Griesemer). The middle three papers (by Nijhout, Schlosser, and Sansom) try to extend and elaborate the evo-devo framework in various ways and will mostly be of interest to those already familiar with the field.

Although this wide range of topics makes it difficult to discuss the collection as a whole, we agree with the editors that it is necessary to explore the scope of the revision that a developmentally focused perspective presents to evolutionary theory. We therefore concentrate on the major themes running throughout the book.

### What Is Included in Evo-Devo?

While no paper in the collection is devoted to formulating a fully fleshed-out, self-contained, conceptual framework for evo-devo (the description of the organismic systems approach [OSA] by Callebaut et al. comes closest to this goal), between them the papers in this collection offer a wide range of conceptual resources. Before considering this conceptual apparatus it is useful to consider the fundamental question: what is evolutionary developmental biology? Griffith’s paper provides a useful characterization:

Evo-devo is associated with the idea that paying attention to development problematizes both the idea that form is shaped in a one-sided manner by the demands of the environment and the idea that the unit of selection is the individual gene. Evo-devo problematizes the lock-and-key model of adaptation because the developmental biology of organisms is an input to the evolutionary process as well as an output. . . . Evo-devo also problematizes the idea that the unit of selection is the individual gene because it describes emergent levels of organization in the developing phenotype . . . [that] retain their identity when they are constructed using different developmental resources. (pp. 195–96)

What remains in the wake of such conceptual upheaval? Callebaut et al. give a list of elements of a conceptual framework for evo-devo: bauplan, canalization, developmental constraints (Schlosser’s paper in the collection is devoted to explicating a notion of constraints), inerency, evolvability, developmental modularity, evolutionary origination, innovation, and novelty (the OSA “innovation
triad”), homology (one of the main topics of Griffith’s paper), robustness, and developmental mechanism. As Callebaut et al. note, evo-devo is still in search of a comprehensive conceptual framework, and they do not attempt to provide an exhaustive list. For example, they do not mention generative entrenchment and scaffolding, the two coordinating concepts used by Wimsatt and Griesemer in their paper in the collection, nor do they mention the crucial notion of genetic accommodation and related phenomena.

The attention given to the developmental perspective in evolutionary theory calls for a historical perspective. Although being in the thick of things is arguably not the best perspective for historical analysis, relating current views to groundbreaking work and hypotheses of previous generations is an integral part of defining and legitimating a new field. Laubichler and Maienschein open the collection by offering a historical overview focusing on the relation between development and evolutionary theory. They outline the sources of the developmental approach to evolution during the 19th and early 20th centuries, and the separation of heredity and development following the establishment of classical genetics. A return to the developmental perspective, they argue, requires an elucidation of the relationship between evolution and development. Is evolution necessary in order to better understand development (evo-devo), or are developmental considerations needed in order to understand phenotypic evolution (devo-evo)? This is an important question but before it is addressed, one must know what exactly should be included in an evo-devo or devo-evo synthesis, and identify research lines that are included (or excluded) from present synthesis attempts.

We identify five major lines of development-oriented research that were relevant to evolutionary issues and have been studied before and during the establishment of the modern synthesis, but that were seen as peripheral or relatively unimportant until recent years. Some of these are central to the evo-devo framework presented in this volume, while others are neglected. The first line, which had little impact on the synthesis, was that pursued by people like Gavin de Beer, who used the comparative anatomy of adults, embryos, and fossils to study the relationship between ontogeny and phylogeny, and which was revived by Gould (1977). The second was developmental genetics, which was studied by Richard Goldschmidt and other more conventional geneticists, like Hans Gruneberg. Goldschmidt investigated the effects of chromosomal changes and gene mutations that had very dramatic effects on development (e.g., homeotic mutations), and from these studies tried to extrapolate to evolutionary transformations. Gruneberg studied the “pedigree of causes” during development, attempting to map genetic changes onto ontogenetic changes. This line of research is central to the evo-devo synthesis, as books such as those by Carroll (2005) and Wilkins (2002) show.

The third line of research that remained marginal was that of epigeneticists like Conrad Waddington and Ivan Schmalhausen, who focused on canalization
Integrating Evolution and Development

and plasticity and who were interested in the evolution of adaptations and norms of reaction. Today this direction of research is represented by Schlichting and Pigliucci (1998), who stress that it is reaction norms rather than genes that are units of evolution; by West-Eberhard (2003), who focuses on the mechanisms and evolutionary effects of plasticity; and by Gilbert (2001), who highlights the intersection of developmental biology and ecology. West-Eberhard’s seminal book provides a general framework for integrating developmental plasticity and evolution, because it introduces a set of helpful theoretical concepts and argues convincingly that evolutionary change usually originates as a developmental response to changed environmental conditions, with genetic fine-tuning following later.

The fourth line of research was the structuralist line, which focused on the generic physical and chemical properties of biological matter and on fundamental biological processes, a tradition whose roots can be found in Goethe and D’Arcy Thompson, and which is well represented in the evo-devo literature in general and in this volume (in the second paper by Callebaut, Müller, and Newman, and in the fourth paper by Schlosser). The fifth was that of cell biologists such as Ruth Sager, Boris Ephrusi, and Tracy Sonneborn and his school, who were interested in non-Mendelian or nongenetic (epigenetic) transmission in microorganisms, and Harris (1982), who studied heritable variations in cells in culture. Some of these scientists were also interested in the evolutionary implications of such variations, but in general their work was ignored by evolutionists. Jablonka and Lamb (1995, 2005) have focused on this research perspective and discussed the way in which it challenges the modern synthesis. Both this last line of research, which is concerned with nongenetic inheritance, as well as the third line of research, especially the direction taken by West-Eberhard and Gilbert, are peripheral to the discussions in the volume and are not included in the historical discussion of Laubichler and Maienschein.

We believe that it would have been helpful if Laubichler and Maienschein had explicitly discussed these five lines and explained how they became incorporated (or failed to be incorporated) in the attempts to construct a new developmental synthesis. This would have enabled them to discuss the tenets of specific theoretical frameworks that were proposed in recent years and that try to straddle the divide between development and evolution. For example, the proponents of developmental systems theory (DST) argue that it is logically impossible to separate the various inputs leading to the development of the phenotype, such as genes and environmental inputs (the so-called parity principle); West-Eberhard presents a plasticity-focused conception of evolution; and Jablonka and Lamb suggest a view of evolution that is centered on a developmental notion of heredity. All these frameworks suggest concrete answers to the important questions raised by Laubichler and Maienschein.

Callebaut et al. devote more attention to DST (referred to as the developmental systems perspective, or DSP) and make several important observations. First,
while they reject the parity principle of DSP according to which genes do not have a privileged role in ontogenetic development, they argue that DSP fails to provide a full causal picture of development and evolution by neglecting to explore the generative potential resulting from the physical constitution and organizational structure of the phenotype. Second, they highlight the opposing notions of plasticity and canalization, two fundamental concepts required for the phenotypic-level view. It is important to stress, however, that plasticity and canalization are complementary and interdependent. Every case of canalized development (in the face of genetic and environmental “noise”) requires plasticity at underlying levels of organization. For example, a knockout mutation that does not lead to a phenotypic effect may be the result of a dynamic reorganization of a developmental gene network. An increase in the number of red blood cells at high altitudes, which is a plastic response if we look at the number of red blood cells (which changes), is an illustration of canalization if we look at the concentration of oxygen in the blood (which remains constant). It is the plasticity at the level of adjusting the number of red blood cells that allows the invariance at the level of oxygen concentration. Griffiths mentions a similar relation when he discusses “levels of homology”: a structure can be preserved on one level while the underlying mechanisms generating it may change, and vice versa. Third, Callebaut et al. convincingly argue that distinguishing between developmental and evolutionary biology based on the distinction between “how?” questions (proximate causes) and “why?” questions (ultimate causes) is an oversimplification, since the developmental-causal perspective of developmental biology is necessarily incomplete in the absence of epigenetic and environmental determinants, and the functional-teleological analysis remains incomplete in the absence of the causal mechanisms leading to phenotypic variation.

In spite of the focus on development, the authors do not discuss the role of genomic mechanisms that create genetic variability, an additional developmental layer that is related to but distinct from the mechanisms they discuss. Genomic plasticity, especially as a response to environmental stress (e.g. heat shock, starvation) and genomic stress (e.g., hybridization, polyploidization), may result from epigenomic coping mechanisms that create increased variation in times of need (Kidwell and Lisch 2001; Lamm and Jablonka 2008; McClintock 1984). The study of the machinery underlying genomic organization highlights the fact that the genome is itself a developmental unit that changes during ontogeny in response to environmental cues, and that the very same ontogenetic reorganization mechanisms may operate during phylogeny. The relations between the genomic environment and the outside environment, as well as the borderline between the plasticity of the genome and the plasticity of the organism, raise conceptual as well as empirical questions that should be addressed by a contemporary new synthesis of evolutionary and developmental ideas.

The “organismic systems approach” championed by Callebaut et al. introduces the concept of inherency, the propensity of biological materials to assume
preferred forms. This structuralist notion is used effectively by Newman and Müller (2006) to argue that many fundamental properties of organisms—such as compartmentalization and segmentation—emerged as a result of the inherent properties of biological materials. Although inherency is defined with reference to the physical/chemical properties of biological materials, fundamental evolved biological processes may have similar emergent effects. For example, once a mechanism such as RNAi comes into being, it begins to influence the types of regulation and variability that are generated and tolerated, and these become more and more an inherent property of the system. This is close to the notion of generative entrenchment (which is applied to the cultural context by Wimsatt and Griesemer in the last paper in the book), which is a measure of how many things depend on a constituent element (structure or process) and are thus likely to need to change if this element changes. The more generatively entrenched a constituent element, the harder it is to change. Inherent properties of materials do not need entrenchment to have a constraining effect, since they reflect essential physical and chemical properties of biological materials, rather than evolved properties. On the other hand, some types of evolved, entrenched, or functionally important properties are not only resistant to evolutionary change but also determine “preferred forms” (act as attractors) for the organization of the system. In other words, they have an effect similar to inherency. Hence, while the notions of inherency and entrenchment are distinct, both emphasize related aspects of developmental and evolutionary robustness.

As the volume attests, the evolutionary aspects of epigenetic inheritance and the control mechanisms that participate in the restructuring of the genome are still generally ignored by most evo-devo biologists, although Sansom discusses one aspect of the generation of genetic variations. Clearly, the discovery of the molecular nature of the gene has been a major breakthrough in the study of heredity, but there are many additional basic discoveries that have to be incorporated into the new evo-devo synthesis if it is to provide an adequate alternative to the modern synthesis. We suggest that in order for this to happen a new view of inheritance is necessary. We will briefly outline what such a view entails and how the relations between genetic variation and phenotypic traits are treated in this volume.

**A Developmental View of Inheritance**

Given the goals of evo-devo, it is unsurprising that both development and evolutionary change are problematized by most papers in the book. In contrast, the view of heredity throughout most of the papers is conservative. Nijhout’s account of the relation between genotype and phenotype is based on the classical view of genes—the complexity he introduces to the traditional picture is the result of the incorporation of interactions between proteins that are the products of genes. Since these interactions are complex and often nonlinear, selection of
phenotypes that result from such interactions must result in complex evolutionary dynamics. Nijhout suggests a mathematical model that attempts to capture this complexity. In this model the genotype-phenotype relationship is represented as a multidimensional graph called a phenotypic surface. By considering how variation in one gene affects the phenotype, in the context of specific values for other genes, it is possible to show the correlation between genetic and phenotypic variation, as well as how the genetic background influences the effect of gene changes and provides constraints. Although Nijhout suggests a sophisticated and dynamic picture of evolutionary change, he does not question the basic assumptions of the modern synthesis about the origin of genetic variations, nor does he incorporate into his evolutionary view genome-wide systemic change that occurs under conditions of stress.

It seems to us that while the modern synthesis black-boxed development, compartmentalizing developmental processes as immaterial to evolution, it is now inheritance that is being black-boxed. The “generic” properties of the mechanisms of inheritance are at the focus of much recent research in genetics, which should inform the evo-devo discussion but is as yet largely absent. The mechanisms we have in mind are mechanisms that lead to genome-wide changes during ontogeny. Many of these mechanisms are epigenetic control mechanisms and, as such, also underlie cellular epigenetic inheritance (for example, chromatin remodeling, persistent silencing by micro RNAs).

Genome restructuring can be developmentally induced, and some of the mechanisms involved in this restructuring are involved in epigenetic inheritance. Epigenetic inheritance allows the formation of new foci for selection (heritable epigenetic variations), and the mechanisms underlying it are central to the regulation of development. This confluence not only shows how heredity, far from being an independent process, is interwoven with development in terms of mechanisms, but that black-boxing it in order to better understand evolution misses an important factor of evolutionary change.

One result of ignoring the way heredity and development are mechanistically intertwined is that the border between a developmental change and an evolutionary change is for the most part assumed to be fixed and inherent. While questioning what happens at this boundary is fundamental for evo-devo, the boundary itself remains in its traditional modern synthesis location, with genes and development each operating on their own level. The discussion of what Callebaut et al. call the “pre-Mendelian” world (an ancient world where the relation between genotype and phenotype was not yet fixed), however, does problematize the border between heredity and development. It suggests that morphological plasticity is a primitive, physically based property, carried over to a limited extent into modern organisms. While this is an important idea, it remains to be explained how maintaining plasticity influenced the dynamics and trajectory of the evolutionary process (e.g., by influencing selection pressure). Moreover, although the relation between heredity and development has evolved and the
matching between genotype and phenotype has become, in some ways, more
fixed, the interaction between the two remained close, flexible, and reciprocal,
and our present world is, in fact, far less “Mendelian” than we have been led to
believe.

The discussion of the Goldschmidtian notion of hopeful monsters by Sansom
concentrates on Stephen Jay Gould rather than Goldschmidt and ignores central
parts of Goldschmidt’s careful discussion. Goldschmidt (1940) applied the evo-
devo–like notion of norm of reactivity—according to which “the genotype is
. . . the inherited norm of reactivity to the ensemble of conditions which may
influence phenotypic expressions” (p. 250)—to the discussion of macro-evolu-
tionary change. Goldschmidt observed that the range of modifiability of one
species under conditions of developmental stress is on a similar scale as the range
of phenotypic differences between related species under natural conditions (p.
253), and he argued that such differences were the result of systemic mutations
(a term by which he meant chromosomal repatterning). Based in part on the
observation that new species are usually chromosomally different from their
parental species, he suggested that evolution above the species level usually
involves chromosomal restructuring. McClintock (1984) suggested that genomic
repatterning occurs under conditions of stress and that transposable elements
play a major role in this process. Recent data on the response of organisms to
stress by recruiting epigenetic control mechanisms that lead to genome-wide
changes suggest that a synthesis of Goldschmidt and McClintock suggestions is
needed (Jorgensen 2004; Lamm and Jablonka 2008).

Goldschmidt’s analysis of macro-mutations was in the context of how new
developmental systems arise (reaction norms), a question that was also of central
importance to Waddington. How genetic changes lead to new integrated develop-
mental systems is central to the evo-devo research program, but the strategy
of genomic changes—which is employed, for example, by many organisms in
conditions of stress—is hardly discussed by its practitioners, including the
authors of the present volume. While Sansom argues in his paper for the adap-
tive advantage of traditional micro-mutations leading to gradual change and
acknowledges the role of environmental stress leading to increased variation, he
does not go beyond proposals that suggest that the amount of variation is sim-
ply increased under stress. However, stress-induced variational processes often in-
volve specific genome coping mechanisms and lead to targeted genome-wide
effects. For example, nutritional stress causes epigenetic and genetic changes in
r-RNA genes and repetitive sequences in flax; heat shock has some similar ef-
fects in _Brassica_, and hydrostatic pressure causes genome-wide changes in methy-
lation patterns in rice (Cullis 2005; Long et al. 2006; Waters and Schaal 1996).
Radiation seems to induce both genetic and epigenetic genome-wide instabili-
ties that last for several generations in both animals and plants (Dubrova 2003;
Molinier et al. 2006). Transposable elements—usually relatively silent—are acti-
vated as a result of various stresses, such as wounds and pathogen attacks, just as
McClintock suggested, and the activity is in many cases restricted to germ cells and hence transgenerational (reviewed in Kidwell and Lisch 2001). Such repatterning problematizes many of the received views regarding directed mutational change, genomic organization, and plasticity.

Although their main focus is cultural rather than genetic heredity and evolution, the discussion of cultural evolution by Wimsatt and Griesemer is based on a nuanced view of the relationship between heredity and development. According to their approach, information has to be developmentally integrated by carriers in order to be assimilated, used, or transferred. This requires scaffolding, a supporting framework, which has to be propagated alongside the information. This argument favors the multiple channels account of both cultural and biological transmission, with extra channels being used to propagate the scaffolding. Wimsatt and Griesemer stress this by noting that the “statistical independence” test—according to which there are no multiple channels since a “parity argument” shows different inheritance channels not to be statistically independent (one of the tenets of DST)—fails to take into account physical separation between the channels which, according to them, is crucial for the analysis of developmental systems. The separation of channels is a claim about their physical separation, not about their statistical independence. Wimsatt and Griesemer emphasize the large number of channels and the role of sequential acquisition that are involved in the transmission of cultural information. This approach may be applied to other ways of generating and transmitting information that involve multiple transmission channels and sequential acquisition, such as those involved in the sequential unfolding of information in the zygote.

**Constraints and Affordances**

The role of developmental constraints, and hence attention to constraints on evolutionary change in general, is central to evo-devo. Schlosser’s paper in this collection attempts to provide a framework for understanding developmental and functional constraints. Schlosser rightly points out that while constraints are often presented as opposed to selection, the former being understood as internal and the latter as external, constraints can arise from the need to maintain a stable/functional organization after variation has been introduced. He thus describes constraints generically as the “boundary conditions on a process whose dynamics (under certain conditions) are described by selection.”

Schlosser makes a distinction between physical impossibility (e.g., elephant-sized mice) and constraints. He argues that the notion of universal constraints is not productive. Rather, he argues, constraints should be conditions that prohibit the realization of certain states or events that are physically possible. According to Schlosser, what is physically possible is determined by a set of universal and immutable laws of transformations and a set of initial and boundary conditions permitted by our theories of the universe. Schlosser notes that according to his
definition of constraints, gaps in the morphospace cannot be used to infer the presence or absence of constraints, since such gaps may be due to physical impossibilities, which Schlosser prefers not to regard as constraints. A different point of view is suggested by McGhee (2007), who defines the notions of geometric constraint, functional constraint, phylogenetic constraint, and developmental constraint, and illustrates how analysis of the morphospace can be used to distinguish between the different types of constraints. Clearly, physical impossibility (which can manifest itself as a geometrical or functional constraint) is a useful starting point when analyzing form, and being extrinsic to the systems, it is not part of the constraints that result from system organization. Furthermore, it can be assumed to be a constant boundary that did not change during the time span of the evolution of life. These observations explain why physical possibility may provide a clue as to what can constitute a “theory of the possible,” which Schlosser argues is necessary for recognizing constraints. But while physical possibility is a boundary, so are chemical properties of organic molecules, the structure and properties of DNA (once it becomes the repository of genetic information), the composition of proteins, and—as we suggested—the mechanisms of genome organization and reorganization. Clearly, elements such as these constrain the realm of possibility, in a way similar to that in which physical reality determines the realm of possible shapes of organisms. When considering shape it is tempting to privilege physics, but when delving into the internal organization of organisms, it becomes hard to justify why some divisions in the realm of possibility should be understood as constraints while others should not. Many of the constraining attributes we just listed are as hard to change for contemporary organisms as are the laws of physics.

Modularity is another central concept in the evo-devo world, and one around which there is a lively discussion. Both Schlosser and Griffiths contribute to this discussion. Schlosser defines units of evolution as units of constituents that tend to coevolve because they constrain each other’s evolution. Units of evolution, so defined, act as modules of evolutionary transformation that operate above the level of the gene. Schlosser argues that modules can only be analyzed relative to a specified set of permitted perturbations in the face of which the module is stable, in the sense that the constituents of the module operate in an integrated and context-insensitive manner with high probability. The types of perturbations that help define developmental modules (e.g., environmental noise) are different from the types of perturbations defining evolutionary modules (e.g., mutations). Schlosser argues, however, that given certain conditions, units of evolution will be congruent with developmental modules. Developmental modularity can thus be used to study units of evolutionary modularity. This analysis is helpful in pointing out that the units of phenotypic evolution are not equivalent to the units of genetic variation: they can be higher-level complexes that are stable as groupings in the context of the relevant types “noisy” inputs. However, Schlosser does not consider genomic perturbations that operate beyond the level of individual genes.
Griffiths is concerned with a different distinction, that between developmental modules and functional modules. Developmental modularity is a tool for understanding the matrix of developmental resources responsible for the development of the organism, while functional modularity attempts to capture the architecture of the system at a give stage. Using these distinctions, Griffiths argues that “mental modules” of the type posited by evolutionary psychology need not be neural-functional modules. He maintains that evolutionary psychologists’ argument that selection will favor many domain specific modules rather than a few general cognition mechanisms relies on a “thin” notion of modules, according to which any architecture that produces dissociations between performances in different domains is “modular.” The thin notion of modules is irrelevant from the point of view of neuropsychology, Griffiths argues, and mental modules should not be assumed to be neural modules.

Griffiths’s paper also explores the notion of homology and argues convincingly for its relevance to psychology, as a science focused on mechanisms. The interest of philosophers in homology has increased in recent years, and Griffiths shows the importance of the homology concept for evolutionary discussion of psychology and argues for an analysis of evolution in which history and homology, rather than adaptation, are central. He maintains that such an analysis can uncover intrinsic developmental processes and deep similarities among organisms.

The wide-ranging issues discussed in this collection show clearly why the evo-devo research program is relevant for biologists working in a variety of domains, and how evo-devo practitioners are engaged with the rest of biology. The book is, however, too specialized to convince outsiders of the importance of the evo-devo perspective and the impact it can have on the way evolutionary science is being done, and given its eclectic nature, it cannot serve as an introduction to the field. Nor does the book as a whole engage critically with related approaches, although some papers make more of an effort on this front than others. Practitioners in the field, however, will find a lot to discuss and argue about. As expected from a rapidly evolving field such as evo-devo, establishing the boundaries of the field remains an ongoing process. The papers in this collection point to areas of active research, open questions, and new research directions. Exciting times—conceptually and empirically—lie ahead.

References


