COLLINGRIDGE AND THE CONTROL OF EDUCATIONAL COMPUTER TECHNOLOGY

Marvin J. Croy, University of North Carolina at Charlotte

During the past fifteen years, David Collingridge has made important contributions to the understanding of technology and the prospects for its effective control. Though philosophically sophisticated, his views have been given more attention by social and political scientists than by philosophers. In an effort to explore the rationale and applicability of his views, this article takes up three tasks. The first is to explicate Collingridge's basic argument on the topic of controlling technology. This argument is contained in his earliest works, *The Social Control of Technology* (1980) and *Critical Decision Making* (1982). The second task is to offer some critical comments on the adequacy of Collingridge's case, and the third is to apply the results of this analysis to a particular development in instructional technology (the use of expert systems and decision support systems).

THE CASE FOR CORRIGIBILITY

Collingridge builds his case by first noting that there are two conditions necessary for avoiding the undesired consequences of a technology: "It must be known that a technology has, or will have, harmful effects, and it must be possible to change the technology in some way to avoid the effects." [1] Unfortunately, one or both of the conditions are often lacking, and attempts to control technology seldom succeed. Collingridge expresses this misfortune in the form of a dilemma which seriously threatens our ability to control technology. This "dilemma of control," as Collingridge terms it, is summarized as follows.

> Attempting to control a technology is difficult, and not rarely impossible, because during its early stages, when it can be controlled, not enough can be known about its harmful social consequences to warrant controlling its development; but by the time these consequences are apparent, control has become costly and slow. [2]

Although Collingridge constantly refers to this difficulty as a dilemma, he never frames it in the explicit argument form of a dilemma. Doing so may be helpful in making clear the central difficulty to be faced, which resides in the fact that one must choose from a limited set of alternatives, each of which leads to an undesirable consequence. Reconstructing the above passage in the form of a Constructive Dilemma produces the following argument.

> Either a technology is in a relatively early stage of development when it is unknown what changes should be made, or a technology is in a relatively late stage of development when change is expensive, difficult, and time-consuming.
> If the former, then control is not possible.
> If the latter, then control is not feasible.
> Therefore, either controlling a technology is not possible,
> or controlling a technology is not feasible.

To illustrate this dilemma, Collingridge presents the example of the automobile. In 1908, changes in the design and use of the automobile were easily implemented, but accurate predictions of its eventual social consequences were impossible. It was not possible to anticipate future problems with air pollution, lead in gasoline, non-renewable resources, social dynamics, etc. Today, the undesired effects of automobile use are easily determined, but change is difficult to implement. Collingridge further reinforces the concept of a dilemma by referring to the situation in 1908 as the "prediction horn" of the dilemma (the doomed attempt to predict the consequences of emerging technologies). Likewise, the situation with the automobile today is referred to as the "control horn" of the dilemma (the doomed attempt to alter well-developed technologies). The alternatives and eventual consequences which comprise these horns of the dilemma can be made explicit by means of a second reconstruction of Collingridge's argument.

> Either one attempts to control technology by early prediction of consequences, or one attempts to control technology by reacting to consequences as they unfold.
> If the former, then the attempt fails because of predictive unreliability.
> If the latter, then the attempt fails because of developed rigidity.
> Therefore, either the attempt to control technology fails because of predictive unreliability, or the attempt to control technology fails because of developed rigidity.

Collingridge believes that most attempts to resolve this dilemma have focused on the prediction horn of the dilemma. Research in Bayesian decision theory and similar approaches involve efforts in this direction and constitute attempts to reject the second premise in the argument above. Collingridge believes that these attempts are doomed. Following Popper, Collingridge adopts a fallibilist view of decision making and denies that decisions can ever be justified. Since prediction of future states of the world is impossible early in the life of a new technology, maximizing one's expected gain is beyond reach. Rather than making decisions under risk as is countenanced on a Bayesian view, decision making actually occurs under ignorance. Having no means of achieving accurate prediction, attention naturally turns toward the control horn of the dilemma and the attempt to alter developed technologies. What makes this so difficult is the fact that technologies become entrenched by virtue of their growing interdependence. Changing one developed technology requires changing many others. Technologies can avoid this problem, Collingridge advises, by aiming to achieve the property of corrigibility. To do this, technologies should be advanced by means of decisions which are easily corrected. "A decision is easy to correct, or highly corrigible, when, if it is mistaken, the mistake can be discovered quickly and cheaply and when the mistake imposes only small costs which can be eliminated quickly and at little expense."[3] This view is elaborated in a number of directions by Collingridge. These elaborations will be considered in assessing the acceptability of Collingridge's proposals for controlling technology.

ASSESSING COLLINGRIDGE'S CASE

Collingridge believes that technology can be controlled by means of decisions which are easy to correct. This concept is explicated in three equivalent formulations: corrigibility, control of unpredictable systems, and flexibility (maintaining a wide array of alternatives). A crucial point about achieving these conditions is that doing so comes at a high price.

> The art of deciding under ignorance may be viewed in three equivalent ways: make decisions which are easy to correct; choose systems which are easily controlled; keep your future options open. It follows from this equivalence that just as high degrees of controllability and corrigibility usually have to be paid for by high control costs, high flexibility is generally similarly expensive. [4]

Each of these three formulations deserves special attention, but as demonstrated below, the same difficulty affects each proposal.

Criticisms of Corrigibility: Some of the difficulties inherent in Collingridge's case for controlling technology can be made clear by appreciating the fact that corrigibility is costly. Ease of correction results in high control costs, and this fact produces a tension which affects many decisions under ignorance. The factors which determine these costs include the cost of monitoring, the cost of error when it occurs, and the time required for correcting the error. Although these variables are clearly identified, it is evident that their cost cannot be accurately measured, or even estimated, in the early days of a new technology. If it was indeed impossible in 1908 to anticipate the impact of the automobile on air quality, it is just as clearly impossible in 1908 to measure the cost of monitoring for air pollution. Neither the foresight nor the means required for this monitoring were present at that time. Likewise, the error cost of air pollution was beyond definition and measurement. The difficulty which this presents can be appreciated by imagining a technologist faced with the decision as to which of several directions to direct a new technology. Collingridge advises us to follow the most corrigible path, but without reasonably accurate measures of cost this recommendation is useless in particular cases. This difficulty affects technological development in another way. Once it is realized that corrigibility costs money, the question raised is how much to purchase. A developer of a technology cannot pour 100% of resources into error detection and correction, and deciding among different alternative allocations requires a reasonably accurate measure of the costs of corrigibility. Without this, talk about the virtues of flexibility and corrigibility remains pointless.

It is ironic that, after so much emphasis on the early days of new technologies as being our best hope for controlling technology by virtue of being our best hope for flexibility and corrigibility, the early days of a new technology are precisely those when measuring the cost of corrigibility is most difficult. In retrospect, perhaps some differences can be assessed. Yet the inability to make accurate predictions, which Collingridge stoutly proclaims as being characteristic of that early period, precludes any reliable measurement of the cost of corrigibility, and, hence, any practical utility for his principles.

Problems with the Concept of Flexibility: Collingridge believes that one should develop a technology in ways that keep one's options open. Developing in a non-corrigible manner supposedly reduces the number of alternatives available. Time itself, Collingridge admonishes, is a "powerful closer of options." The intent of these claims may seem familiar and straightforward. However, closer inspection reveals a set of difficulties related to those discussed above. First, it is just false that making one decision as opposed to another reduces the number of available options. Making a particular decision always closes off some options and opens up others. The difficulty, of course, is that the options closed off may be the desired ones and those opened up may be unwanted. Yet the number of options available is always practically infinite. So, acting to maximize flexibility makes no sense. Acting to maximize the number of desirable options available does make sense, but the issue again concerns how and when the desired and undesired options can be distinguished. The claim here is that decisions which maximize desirable options may be easily identified in retrospect, but, once reliable prediction is forsaken, cannot be identified early in the life of an emerging technology.

Criticism of the Controllability of Systems: Collingridge states that being ready to correct decisions in response to feedback is equivalent to being able to control a system whose future behavior is unpredictable. A system is easy to control when the time required for responding to error is short and when performance loss due to error is small. Much of Collingridge's case here and elsewhere is based on a belief in the inevitability of error. Because mistakes are bound to occur, systems which are most responsive to errors are to be preferred. This view, however, involves an over-emphasis on the inevitability of mistakes. Without considering the frequency or severity of the mistakes to be encountered, investments in corrigible systems cannot be guaranteed to be worth the price. The mere fact that unforeseen mistakes will materialize cannot in itself justify the high cost of error detection and correction. The mistakes may be trivial or infrequent. Given past experience, it is reasonable to assume that mistakes will occur. But it is not reasonable to assume that those mistakes will be severe enough to justify the amount of controllability purchased by a particular proportion of resources.

It is not the case that more controllable systems should always be preferred over less controllable ones. This is because technologies are developed in the context of problems to be solved and goals to be attained. Achieving these goals often dictates that only as much controllability as is needed be purchased. The price of controllability, if too high, may interfere with the attainment of important goals.

Given our ignorance of the errors which await us, Collingridge urges us to pay the price of controllability. Yet whenever problems are pressing, and particularly when time is of the essence (the space race to the moon, the search for a cure for AIDS, the rush to be first in the market, etc.), it can be equally or more reasonable to proceed in a less flexible manner.

The basis for Collingridge's rejection of these points is made clear in his discussion of controllability and its assessment.

> Consider a system that interacts with the environment surrounding it in ways that confer various benefits and impose various costs on the system's controller. The controller receives signals that tell him how the system is behaving and he can alter its behavior by adjusting one or more of a number of decision variables of the system, each adjustment taking time to become effective. We may think of the controller as steering the system through the environment by means of the system's decision variables. The pay-off over time is a function of the interaction of the system and environment, but in the cases which are of interest to us this cannot be predicted because the controller has to make decisions about the system's decision variables under ignorance. If the controller has to choose which of two systems to steer, he cannot, therefore, base his choice on knowledge of the payoffs he will receive from each system. His choice can only be based on his knowledge of the system, not on its interaction with the environment. . . . Flexibility is thus to be judged from the system itself, no information about how the system will actually interact with its environment being needed. [5]

For Collingridge, this separation of the system from the environment in the process of assessing controllability is crucial. Nevertheless, this view neglects the fact that the aim of technological development is neither flexibility nor controllability in itself. The aim is to develop systems that achieve goals such that payoffs exceed costs. The connection between payoffs exceeding costs and either corrigibility or flexibility or controllability is never made by Collingridge. Making that connection would require that the characteristics of the environment, not merely the system, be taken into account. Without that connection, the high cost of controllability cannot be justified. That cost can only be justified by arguments showing (i.e., reliably predicting) that the inevitable mistakes will or probably will be severe enough to justify the investment in controllability.

MAKING A BETTER CASE FOR CORRIGIBILITY

Collingridge rightly points to the importance of corrigibility, and a good case can be made, not for its being the sole means to the rational control of technology, but for its being an important ingredient in that control. The basic weakness in his case is that, once the possibility of reliable prediction is undercut, corrigibility and its cognates collapse as a basis for controlling technology. The ability to predict consequences to at least some degree of accuracy is crucial for intelligent choices concerning corrigibility. Collingridge was correct in making the prediction of harmful effects a *necessary* condition for consistently avoiding those effects. His case abandoned that claim, and yet his views are intuitively compelling, primarily because of the intuitive appeal of the concepts of flexibility and ease of correction. Flexibility is seen in a positive light because of the inferred connection between having more options and solving problems. As indicated above, this connection can best be made, not via a mere increase in options (flexibility), but rather via an increase in useful or desirable options (versatility, perhaps). Making this connection requires the existence of usefully accurate prediction. This is precisely what Collingridge denies, and making a better case for corrigibility will require a rejection of that denial. In addition, the claim in favor of corrigibility should be moderated. Rather than claiming that all technologies should be corrigible, it is more reasonable to conclude that, among the alternatives under development, some corrigible options should always be included. To see this, more needs to be said about ignorance and the nature of new technologies.

Collingridge uses the term “ignorance” in two different contexts. One use occurs in the context of formal decision theory. There, the emphasis is upon the inability to know all possible future states of the world. A second use occurs in the context of developing particular technologies or systems in the real world. Here, the term refers to our inability to predict the future consequences of those technologies well enough to warrant their control. In neither of these cases, however, would Collingridge's conclusions concerning corrigible paths of development follow. In the event of gross ignorance, there is no reason to prefer one development path over another. One may as well lay out all the known alternatives and randomly choose one. Or, available resources might be divided equally or randomly among all known options. If, on the other hand, gross ignorance does not prevail, some degree of reliable prediction has evolved. When that degree is high, usually as a result of a long history of trial and error, future states of the world can increasingly be anticipated with confidence. The most interesting and perhaps crucial cases are those that lie between these extremes.

The closer a technology is to the initial period of predictive unreliability, the less confidence should be placed in its portraits of the future, and the less convincing should arguments be as to the appropriate development path. But when is a technology genuinely new? According to Collingridge, the automobile in 1908 was a new technology. But what about the electric car in 1990? While it is impossible to predict all that follows from implementing this technology, surely its consequences can be forecast much more accurately than was the case with the automobile in 1908. Nevertheless, Collingridge provides no guidance in respect to identifying new technologies, and the issue at stake here is an important one; namely, when is it that corrigibility should be preferred in the development of a technology? Resolution of that issue depends on an assessment of realistic cases in which ignorance is not practically complete.

While genuine ignorance logically supports random choice, periods of very little knowledge logically suggest variability. Since knowledge is too scarce to justify the choice of one development path over others, a diverse set of alternatives should be pursued. Corrigible and flexible paths are as likely to succeed as any other at this point and should be included in the set of alternatives. As knowledge and the number of successful predictions increases, arguments in favor of particular paths of development will become more cogent, and the importance of variability will decrease.

One part of this formulation, namely risk, is conspicuously ignored by Collingridge (perhaps because of his aversion to Bayesianism). Nevertheless, experience often progresses to the point where dangers resulting from technological failure can be reliably predicted. In these cases, corrigibility and slow rates of development can be justified, not by reference to the systems alone, but by reference to the environment and the difficulties produced by certain interactions with it. Given this knowledge, flexibility as a means to additional, *relevant* alternatives is also to be valued. In any event, corrigibility and flexibility may play extremely important roles in the development of a technology, but arguments in their favor rely upon the existence of some degree of predictive accuracy. Once a transition is made from the claim that every technological development should be corrigible to the claim that, of the alternatives being developed, at least some should be corrigible, the degree of predictive accuracy required is lessened.

In sum, the dilemma of control can best be resolved not by accepting one horn or the other but by slipping between the horns. It is not the case that we are, technologically speaking, either in a state of insufficient predictive power or in a state of knowledge obtained too late. Many, if not most, technologies are in a state where reliable predictions can be made by degrees, and discovering the limits of those degrees and the best corresponding developmental strategies is of immense importance. This is an empirical endeavor, and this point leads to an examination of the development of educational computer technology.

APPLYING COLLINGRIDGE'S FRAMEWORK TO SOME DEVELOPMENTS IN EDUCATIONAL COMPUTER TECHNOLOGY

In recent decades the technology of instruction has been advanced by two major forces. One of these has been the rapid development of computer technology and the other has been the emergence of artificial intelligence. Computer science and cognitive science have made reciprocal contributions, and one important impact on educational technology has been the rise of intelligent computer-assisted instruction (ICAI). Patrick Suppes summed up the potential of this endeavor in 1979, by drawing an analogy with the sort of education provided by Aristotle's tutoring of Alexander the Great.

> We should have by the year 2020, or shortly thereafter, CAI courses that have the features that Socrates thought desirable so long ago. What is said in Plato's dialogue *Phaedrus* about teaching should be true in the twenty-first century, but now the intimate dialogue between student and tutor will be conducted with a sophisticated computer tutor. [6]

The practical significance of this concept was demonstrated by Richard Bloom and his research associates. Their work documented the relative effectiveness of tutoring. Bloom found that students receiving one-to-one tutoring scored two standard deviations higher than students receiving traditional instruction with a standard teacher-student ratio of 1-30. [7] In 1982, Sleeman and Brown's *Intelligent Tutoring Systems* documented the advances and agenda of a diverse research program focused on tutoring. [8] With the invention of expert systems, tutoring came to be viewed as a rule-governed process, and the next step was clearly to explicate the relevant rules, whereupon they could be captured in a computer program. For example, Reiser (1989) initiated an empirical study of one-to-one tutoring in the area of learning to program in LISP. [9] The impact of the success of these efforts on the educational process, students, and teachers has not been given primary consideration by the developers of ICAI. Nevertheless, some possibilities have been recognized.

One obvious possibility is the impact on human interaction in education, and in fact, Sleeman and Brown hold out the possibility that human interaction might serve as a "congenial and effective backup." Grabinger and Pollock (1989) developed an expert system which eventually replaced graduate teaching assistants in the role of evaluating student work. [10] This development and others described above highlight a distinction between different kinds of intelligent systems. Some systems are designed to make important decisions previously made by humans. Expert systems are often good examples of this type, being designed to function in the role of experts. Other systems are designed not to execute decisions but to support and increase the effectiveness of decisions made by humans. Programs referred to as decision support systems are often good examples of this type. In education, this distinction is crucial. It sets up a conflict between two different attempts to improve education. One attempt would proceed by making computer programs intelligent enough to take over significant components of decision making in the teaching process. The other attempt would design computers to supply information and sophisticated analyses that would improve the pedagogical decisions made by teachers. [11]

In the context of Collingridge's views on controlling technology, the question naturally raised is whether expert systems or decision support systems provide the more corrigible and flexible path for development. There are a number of reasons for answering that question in favor of decision support systems. First, decision support systems leave most of the control in the hands of versatile, human teachers. Perhaps the principal trait of human intelligence is adaptability, and the characteristics of students and learning environments often vary unpredictably. Teachers have the ability, via human judgment, to adapt previous experience to cope with novel situations. Moreover, human intelligence is supported by the power to learn quickly, sometimes from single cases. In contrast, the strength of automated instructional systems is the ability to sustain repetition and the ability to reach an almost unlimited number of students. The issue of whether computers can be programmed to exercise judgment and to learn from experience is still the subject of research. That research may well turn out positively, but it should be clear that unless computers rival or exceed the human capacity for judgment and learning, systems which depend more on computer processing will be less versatile and less easy to correct than systems which depend more on human processing. [12]

Given the current state of development in educational computer technology, corrigibility, versatility, and slow rates of development have an important role to play. The importance emerges precisely because of what is known about the risks of computerization, some of which has been learned by decades of experience with automation. Experience with promising educational media such as television, radio, and computers, and with automated learning systems, such as programmed texts, complete-course CAI, and computer managed instruction, have made two points clear. First, these innovations have often failed to live up to expectations, and second, their implementation has threatened certain educational values. [13]

The conclusion to be drawn from this is not that new educational technologies should be avoided. Rather than judging technologies prior to implementation because of the mere possibility of costs outweighing benefits, innovations should be pursued in conjunction with adequate monitoring. Monitoring is perhaps the most important concept elaborated by Collingridge. Monitoring means more than evaluating the effectiveness of a system, and it is fundamentally different from traditional formative and summative evaluations. Monitoring involves a continuous "scrutiny of a decision's real consequences with the aim of finding error." [14] Here, the decision being monitored is that of implementing a new educational technology, and the consequences being scrutinized involve wide-ranging social, attitudinal, and behavioral effects. These effects cannot be predicted ahead of time, yet enough is known to identify at least some of the risks. If instructional systems must be implemented, at least partially, in order to discover the consequences of their use, the question looms: in what respects should educational computer technology be monitored? What exactly should be scrutinized? The answer to these questions derives from educational values, namely such values as concern for individuals, individuality, cooperation, privacy, equal opportunity, etc. The role of these values in defining and shaping the educational process has long been recognized.

One avenue to corrigibility consists of a slow rate of development.

Collingridge articulates this idea in the guise of "incrementalism" in his later work. [15] Although Collingridge neglects the relevance of risk in decision making, slower rates of development produce a better understanding of the risks involved. Developing in a slow, piecemeal fashion will produce fewer mistakes in a given time frame and can provide more lead time for reacting to mistakes. In this way, the consequences of becoming dependent on particular systems of hardware, software, and human-machine interaction can more readily be predicted. This is extremely important in the context of educational technology given recent discussions of distance learning and the future of American education. Discussions of implementing intelligent computer-assisted instruction and distance education rarely consider the risks of implementation. The emphasis of those discussions is upon the defects of the educational system. Nevertheless, the strengths of the American higher educational system, in particular, far outweigh its weaknesses. This, of course, does not rule out the possibility of improvement through change. But it does indicate that any attempted improvement may have effects which are undesirable. Any change involves the risk of losing something of value, and modern technologies are notorious for the ability to produce fundamental change. The concept of risk is one that should play a more prominent role in discussions of educational technologies.

CONCLUSIONS

The consequences of implementing many current educational technologies cannot be predicted with accuracy. Nevertheless, those projects should be developed as long as monitoring is initiated early in the development process so as to provide empirical guidance as soon as possible. Projects which choose slow rates of development and corrigible techniques should be supported and funded. This means that systems which seek to enhance human decision making as opposed to automating it should be developed with as much vigor as other forms of intelligent CAI. Which alternatives are best suited for specific contexts, which solve current problems without introducing more serious problems, and which produce benefits outweighing costs will be discovered over time.

Collingridge moves the discussion of technology and its control in the right direction. The concepts of corrigibility and versatility are important despite the inability to establish their universal superiority. These concepts are particularly relevant to intelligent instructional technologies. In an era when many educators are scrambling to introduce the latest computer technologies into their classrooms, alternatives which acknowledge risk, expect failures, and, using past experience as a guide, monitor the wide-ranging impact of their implementation, fill an important niche. The effort to make educational computer technology more intelligent should benefit from their pursuit. While the success of particular corrigible approaches cannot be guaranteed, the success of the overall attempt to improve teaching and learning is made more likely by their pursuit.

NOTES

1. David Collingridge, *The Social Control of Technology* (New York: St. Martin's Press, 1980), p. 16.

2. *Ibid..*, p. 19.

3. *Ibid.*, p. 32.

4. *Ibid*., p. 38.

5. David Collingridge, *Critical Decision Making* (New York: St. Martin's Press, 1982), p. 146.

6. Patrick Suppes, "Observations about the Application of Artificial Intelligence Research to Education," in D. F. Walker and R. D. Hess, eds., *Instructional Software: Principles and Perspectives for Design and Use* (Belmont, CA: Wadsworth, 1984), p. 306.

7. R. S. Bloom, "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring," *Educational Researcher,* 1984, pp. 4-16.

8. D. Sleeman and J. S. Brown, *Intelligent Tutoring Systems* (New York: Academic Press, 1982).

9. Brian Reiser, "Pedagogical Strategies for Human and Computer Tutoring," Annual Meeting of the American Educational Research Association, 1989, ERIC Document ED 316 195.

10. S. Grabinger and J. Pollock, "The Effectiveness of Internally Generated Feedback with an Instructional Expert System," *Journal of Educational Computing Research*, 5 (1989): 299-309.

11. Systems which combine varying degrees of these functions are also possible. For example, some decision support systems contain expert systems as components. Nevertheless, in the following discussion, the term “expert system” will be used to denote a system designed to make a decision on its own, while “decision support system” will be used in respect to supporting a decision made by humans.

12. Analyses of relevance to the choice between expert systems and decision support systems can be found in the following: Dianne Berry and Anna Hart, eds., *Expert Systems: Human Issues* (Cambridge, MA: MIT Press, 1990); Marvin Croy, James Cook, and Michael Green, "Human vs. Computer Supplied Feedback: An Empirical and Pragmatic Study," *Journal of Research on Computing and Education*, 26 (1993): 185-204; and Marvin Croy, Michael Green, and James Cook, "Assessing the Impact of a Proposed Expert System via Simulation," *Journal of Educational Computing Research*, 13 (1995): 1-15.

13. For historically enlightened treatments of these issues see Larry Cuban, *Teachers and Machines: The Classroom Use of Technology since 1920* (New York: Teacher’s College Press, 1986), and chapter 5 of Hubert and Stuart Dreyfus, *Mind over Machine: The Power of Human*

Intuition and Expertise in the Era of the Computer (New York: Free Press, 1986).

[14] Collingridge, *The Social Control of Technology*, p. 32.

[15] David Collingridge, *The Management of Scale: Big Organizations, Big Decisions, Big Mistakes* (London: Routledge, 1992).

[16] This work was supported in part by funds provided by the University of North Carolina at Charlotte.