Problems in Causal Loop Diagrams Revisited

George P. Richardson
Rockefeller College of Public Affairs and Policy
University at Albany - State University of New York

A reemerging problem

Word-and-arrow diagrams (causal-loop diagrams, influence diagrams, cognitive maps, and the like) are enjoying widespread use in the system dynamics and systems thinking communities. It is increasingly common to see these diagrams with links labeled “S” and “O” to identify causal effects in the “Same” or “Opposite” direction to changes in the causing variable at the tail of the arrow. But therein lies an old problem in a new disguise.

An illustrative example

Figure 1 shows a pair of feedback loops representing the essential structure of the spread of a disease, diagramed in the style popularized by The Systems Thinker, Kim (1992) and others, and now much in vogue in the systems thinking and system dynamics literatures. The diagram has a serious flaw, the same flaw pointed out in Richardson (1986/1976) and corrected in the definitions in Richardson and Pugh (1981), but hidden here in the S’s and O’s: two of the links do not behave as their labels claim.

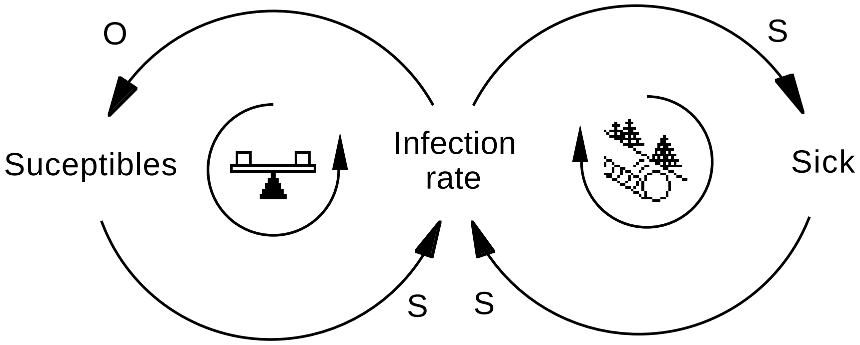

Figure 1: Causal loop diagram in “S” and “O” notation, capturing the balancing and self-reinforcing loops inherent in the spread of a disease

The story intended in Figure 1 is that a few people infected with the disease (Sick) make contact with people who can catch the disease (Susceptibles), resulting in more people becoming sick, so still more Susceptibles become infected. This self-reinforcing process continues (in this simplified picture) until the stock of Susceptibles falls low enough to slow and eventually halt the spread of the infection. (A stock-and-flow diagram for such a system is shown in the notes, along with a graph of typical behavior.)[1]

The link from the susceptibles to the infection rate is labeled "S," meaning, according to the current characterizations (e.g., Kim 1992), that as the susceptible population changes the infection rate changes "in the Same direction" (ceteris paribus). Similarly, the link from the infection rate back to the susceptibles is labeled "O" meaning that as the infection rate changes the susceptible population changes "in the Opposite direction." The labels on the other arrows in the diagram have analogous interpretations.

Yet it is clear that two of the characterizations of these arrows are false. Consider the link from the Infection rate to Susceptibles. If the infection rate were to decrease, as it does in the later stages of the spread of the disease, the susceptible population would not increase (move in the opposite direction) as its "O" label suggests — susceptibles in this system would continue to decrease as more become sick. The link from the Infection rate to the Sick population has a similar problem: when the infection rate decreases, the sick population does not decrease (move in the same direction) as the "S" label suggests — it would continue to increase. We know very well the reason for these behaviors: the infection rate always *subtracts* from the susceptible population and *adds to* the sick population. The populations are stocks, and the infection rate drains one stock and pours into the other. In this diagram, the susceptible population always decreases and the sick population always increases, whether the infection rate is increasing or decreasing. Thus, the "S" and "O" labels rather fundamentally mischaracterize the meaning of two of the links in this simple structure.

Moreover, as pointed out in Richardson (1986/1976), such misbehaving links will occur at least once in every causal loop that a modeler would draw to capture complete system structure, because there must be a stock in every loop in a system dynamics model. Hence, at least one of the arrows in every feedback loop in an accurate loop diagram must be an additive or subtractive influence for which the S and O notation is incorrect.[2]

At least one very experienced modeler uses the S and O notation by very carefully defining each label in terms of an increase or decrease from what *would have occurred* without the change in the causing variable. For example, he would define the link from Infection rate to Susceptibles by saying "if the Infection rate increases then the Susceptibles will be *less than they would have been* had the Infection rate not increased," thus justifying the O for Opposite. This sort of definition solves the problem, but unfortunately most of us are not this careful in the definition and usage of these labels. The definitions published in every issue of *The Systems Thinker,* for example, say S and O indicate "change in the same direction," and "change in the opposite direction." Furthermore, the problem is "solved" by the more sophisticated definition by covering it up (see Richardson 1986/1976, 161-162). The real problem is not subtle wording but rather that Susceptibles is a stock and the Infection rate is a flow — that the process of infection moves people by subtracting from the pool of susceptible people and adding to the pool of sick people — and any correct reading of the word-and-arrow diagram requires recognizing that fact.

The result is very general: in any word-and-arrow diagram that contain concepts that should be interpreted as stocks (levels, accumulations) and flows (rates), the S and O notation fails to capture for the reader the structure of what is really occurring.

As diagrams like Figure 1 have grown in popularity, the form of the diagram shown in Figure 2 has gone out of favor, yet it has a very desirable property that Figure 1 lacks. When polarities are labeled with positive and negative signs, links can be defined as *either* additive or proportional influences, and the notation works well semantically in either case. One can say that a positive arrow from A to B means that *A adds to B,* or, *a change in A causes a change in B in the same direction* (resulting in a positive correlation or direct variation). For a negative link from A to B one says *A subtracts from B,* or, *a change in A causes a change in B in the opposite direction.* Such definitions and interpretations have been commonplace since Richardson and Pugh (1981). Their applications in practice are easy, for the creator or describer of a word-and-arrow diagram knows the meaning of the concepts in the diagram and thus knows when a link is adding or subtracting versus when the influence described is proportional.

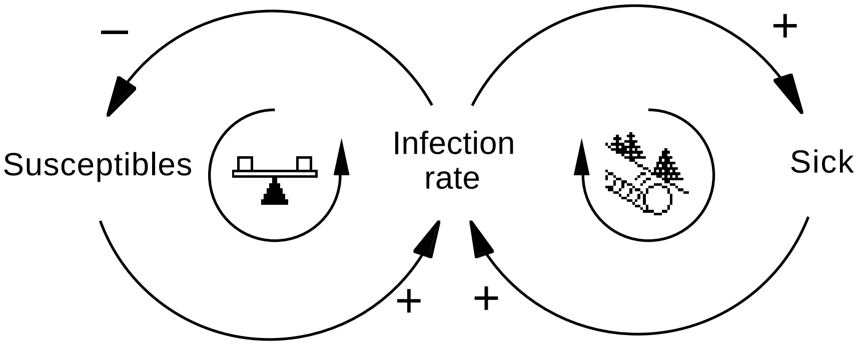

Figure 2: The loop in Figure 1 shown in arithmetic sign notation, where the signs on the two top links would be interpreted as addition and subtraction, while those on the bottom two links represent proportional change.

Unfortunately, in the current “S” and “O” notation, no analogous definitions or interpretations are possible. We’d have to say “S” can mean “adds to” and “O” can mean “subtracts from,” neither of which makes any intuitive sense.

The motivation for “S” and “O”

I conclude that the current fad of using S’s and O’s to label the polarity of links in influence diagrams is seriously flawed. It does have two desirable properties, however, which are presumably the reasons it was introduced in the first place. First, it helps to prevent the mislabeling of the polarity of links that beginners sometimes create when tracing around the loop the up-and-down implications of a change in a variable. When one comes to a link and says "when C drops then D tends to rise," some beginners have a tendency to put a positive sign on the link, signifying "rise," instead of a negative sign indicating "change in the opposite direction." The solution to this erroneous tendency is straightforward, however. Suggest that beginners assign link polarities by going to each link separately and always determining the implication of an *increase* in the variable at the tail of the arrow; then the direction of change in the variable at the head matches intuitively the correct polarity of the link: "If C increases, the D tends to *decrease,* so the link is *negative.*" Once the link polarities are correctly established individually, one can then get the polarity of the entire loop and tell its self-reinforcing or goal-seeking story.

The second desirable property of the "S" and "O" notation is that it strives to sidestep any nonmathematical tendencies of folk we are trying to reach with systems insights and thereby strives not to put them off. It thus enables a superficial following of lines of reasoning in a systems thinking exercise. Unfortunately, that very superficiality runs completely counter to the purposes of systems thinking consultants and clients because it reinforces tendencies to not think deeply or clearly. If one does think clearly in any instance of a diagram with additive or subtractive influences, one either becomes confused by the "S" and "O" notation, or one realizes the letters do not capture what is actually happening in the system.

There is a third possible motivation — to use "S" and "O" notation to avoid thinking of "positive" and "negative" as "good" and "bad," but I doubt this is a significant motivator since the issue is so easily dismissed by simply mentioning it. In any case, I suggest that none of these properties of the "S" and "O" notation outweigh the dramatically undesirable property that the notation is simply wrong at least once in almost every loop one would construct to capture causal structure accurately.

Rekindling motivation for "+" and "–"

Teachers learn that if there is a subtlety in a line of reasoning, and for the sake of efficiency or misplaced compassion they hide the subtlety, someone is sure to become either confused or disenchanted. Both reactions are possible with "S" and "O" notation from folk who think deeply enough to see the problem and who note that it is not being addressed. Confusion can be dealt with moderately easily: we can point out the double meaning we require in polarities in causal loop diagrams, note that S's and O's don't really do the trick, and suggest either positive and negative signs or full stock and flow representations to cure the problem. We could even combine the S and O notation with the + and - notation, reserving S's and O's for proportional change links and using +'s and -'s for additive and subtractive links. But disenchantment is more difficult to deal with, for it usually results in rejection of the enterprise — in school, dropping out of a course that persists in such glossings over; or in our case, abandoning a serious systems thinking effort.

We have three viable options for our word-and-arrow diagrams: 1) use plus and minus signs for all links to indicate link polarity;   2) use plus and minus signs for additive and subtractive links, preferrably with boxes drawn around the stocks they point to, and use S's and O's, if one must, for the other links, which would represent proportional links for which the S and O notation works correctly; 3) show explicit stocks and flows (as tubs and pipes) with the remainder of the links labeled with either plus and minus signs or, if desired, S's and O's. Any of these options can yield a logically consistent diagram. Taste can dictate the choice. Personally, I favor number 1 (with boxes around the obvious stocks) and number 3, depending on the audience, but I suspect one can find merit in any of these three options.

Only the fourth option — labeling all links in a word-and-arrow diagram with S's and O's — is logically flawed and must be ruled out. I submit that our current enthusiasm for S's and O's is significantly misguided and needs to be curbed. Positive and negative polarities stand up much better to the deep thinking we are striving to facilitate in and about complex systems.

References

Bass, F. M. 1969. A New Product Growth Model for Consumer Durables. *Management Science* 15: 215-227.

Kim, Daniel H. 1992. *Systems Archetypes.* Toolbox reprint series. Cambridge, MA: Pegasus Communications.

Richardson, G.P. 1986/1976. Problems with Causal Loop Diagrams. *System Dynamics Review* 2,2 (summer): 158-170.

Richardson, G.P. and A.L. Pugh. 1981. *Introduction to System Dynamics Modeling with DYNAMO.* Cambridge, MA: MIT Press; reprinted by Productivity Press, Portland, Oregon.

The Systems Thinker. Cambridge, MA: Pegasus Communications.

Notes

[1] To keep the picture simple in Figure 1the Sick unrealistically never get cured. A full stock-and-flow diagram of the structure, together with typical model behavior, is shown in figure 3. Note that the two links from the Infection Rate in Figure 1 become a single pipe in the stock-and-flow diagram.

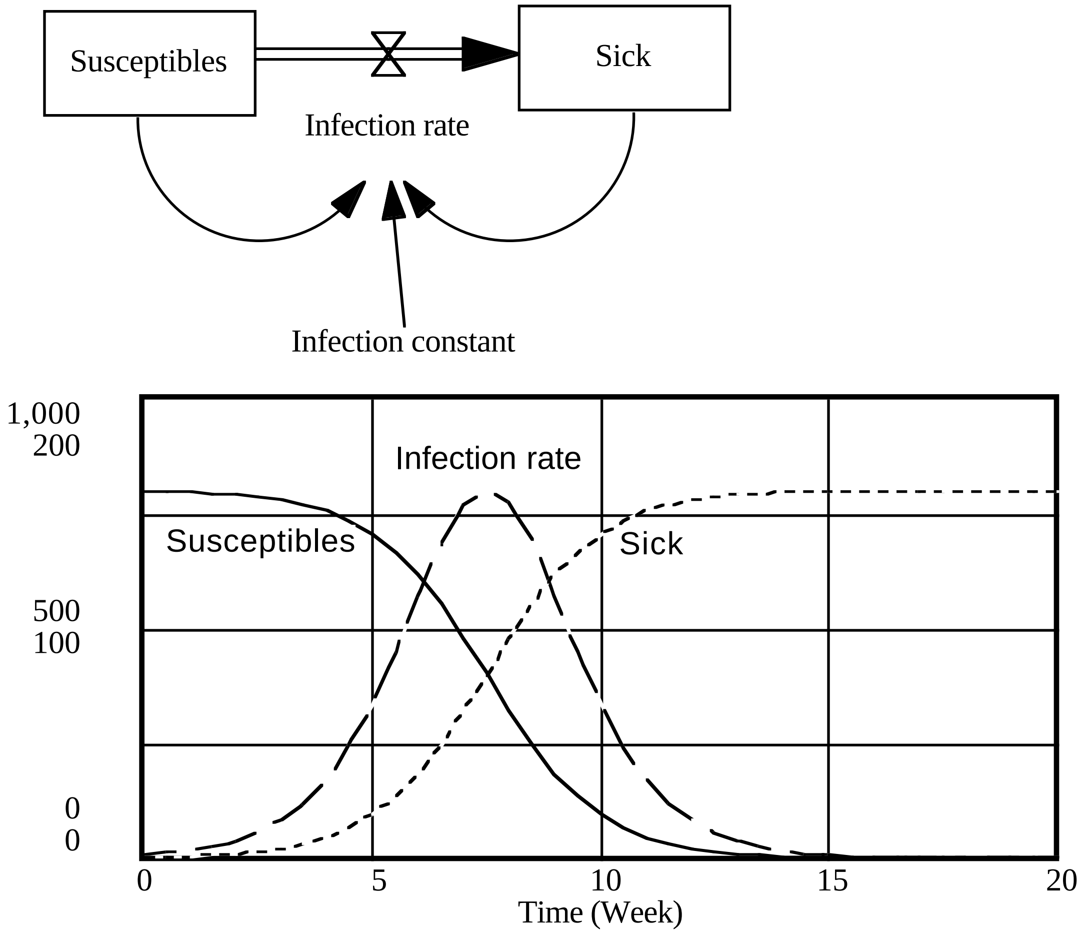

Figure 3: The structure and behavior of a trivial system dynamics model fitting the causal loop diagram in Figure 1. The Infection Rate is modeled simply as a constant times the product of the Susceptibles and the Sick populations.

The graph in Figure 3 makes it clear that the Infection Rate moves, at different times, in both the Same and the Opposite directions as the two stocks — the S and O notation fails completely here for the links from the Infection Rate.

The structure here is a variant of the “limits to growth” archetype (Senge 1990, 95-101) and is related to diffusion models such as the Bass model of market development (Bass 1969). For a more complete epidemic model see Richardson and Pugh (1981, 95-98).

[2] One can draw word-and-arrow diagrams that do not contain concepts one would model as stocks, and in such cases the S and O notation itself causes no problems. However, such diagrams would be incomplete from a modeler’s perspective and probably tell at best only a loose story about the dynamics of the system. In any case, usage of the S and O notation has not been limited to “no stock” causal loop diagrams so such special cases do not solve the problem.