Supplement S1: Model documentation Confidence Trap Resubmission

This supplement is the model documentation to the paper ***When operators and regulators fall into a confidence trap: The anatomy of man-made disasters*** and is structured as follows. We start from the visualization of our dynamic hypothesis, in causal diagram notation, in Figure 1 of section 1, drawn in the software package Vensim (www.vensim.com). When formalizing this model into a simulation model, we need to translate key variables in stocks and flows, with auxiliary variables and parameters. This full model is shown in Figure 2 at the start of Section 2. Next, Section 3 describes the equations for each of the variables in this model, ordered by the five feedback loops they are part of, loop B1 to B5. Doing so requires 30 equations of variables. Section 4 contains a table with all 43 parameters used in this simulation model as well as their values in the base case version of the model. Section 5 describes the dynamic behavior of this base case, using graphs to illustrate the behavior of the main variables over time. However, these graphs are just for one set of values for the 43 parameters. In Section 6 we systematically search the entire set of parameters for changes in parameter values that lead to significant changes in the performance of the model, in particular changes in the occurrence and/or prevention of man-made disasters. The primary purpose of doing so is policy design as this analysis indicates possible high-leverage policies for improvement. An addition purpose served by doing this sensitivity analysis is testing for flaws in the model – as such, changes should not lead to unrealistic values.

1. Causal loop diagram of the dynamic hypothesis

Figure 1 below reveals the core logic of the simulation model. The model is built from five interlocking feedback loops – negative or balancing feedback loops labelled B1 to B5 – that interact causally in multiple ways to generate the dynamic phenomenon of the confidence trap.

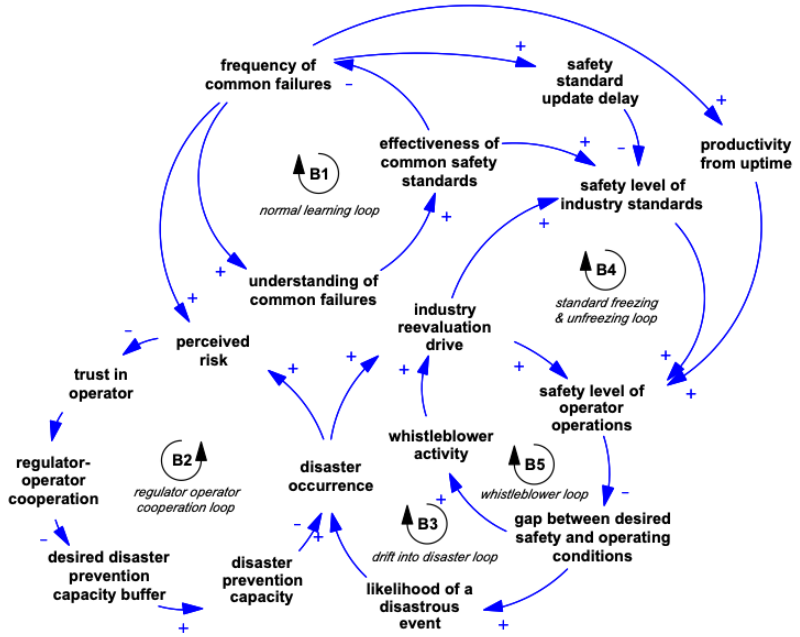


Figure 1: The dynamic hypothesis in causal loop diagram notation

This dynamic hypothesis describes what typically occurs in terms of safety over the lifetime of a complex technical asset, be it a certain type of aircraft, a chemical plant, nuclear facility, or any other high-risk complex technical installation or socio-technical system.

The hypothesis starts with the dynamics generated by loop B1. During the early stages of such a technical asset, a great deal of deep engineering expertise is brought together to develop the asset. Here, the designers consider all sort of inherent limitations to the design, how to mitigate these limitations and how to manage them. This considerations leads to a body of maintenance, operations, and safety procedures, to a set of standards. In the initial so-called "teething problems stage," previously unforeseen dynamic behavior is observed, leading to failures. Understanding of these often-occurring failures grows rapidly, which either leads to redesigns or to adequate process measures to account for these failures, be it in operations or in maintenance. Consequently, the effectiveness of safety standards for common failures grows. Over time, this process saturates as this understanding accumulates, and fewer and fewer unforeseen failures occur. This is what we call a *normal learning loop*.

At first sight, the dynamics of loop B1 may sound familiar, rather ordinary, and somewhat peripheral to our core topic of disasters. However, these obvious-looking dynamics trigger dynamics in other loops that together can lead to man-made disasters.

1. As failure frequency decreases, the perceived risk involved in operating the asset reduces. This leads to greater competence-based trust of the regulator in the operator, and consequently a greater willingness to cooperate with the operator in *reducing safety measures that can mitigate the harmful effects of disastrous incidents* right when they occur. This makes the operator more vulnerable to potentially disastrous events (loop B2).
2. Meanwhile, at the operator, where these standards are indeed implemented, continuing efforts to push productivity at the expense of safety leads to an organizational drift away from high safety (loop B3).
3. As the failure frequency decreases, fewer and fewer changes are made to the existing industry safety standards, which results in *a freezing of the safety standards* (loop B4).
4. Over time, as new uses are found for the asset, as new components and materials are used, technological changes occur, etc., *their relevance, and hence, their guarantees for safety, slowly erodes* (loop B4).

2. *Stocks-and-flows diagram of the formalized dynamic hypothesis*

In translating our conceptual dynamic hypothesis into a quantified model, we need to make choices for the formalization of each relationship into a specific mathematical form. The variables in this model are in **bold**, and their causal links are in blue. Quantifying also implies choosing appropriate auxiliary variables and (constant) parameters. Figure 2 below shows the full Vensim (www.vensim.com) model:

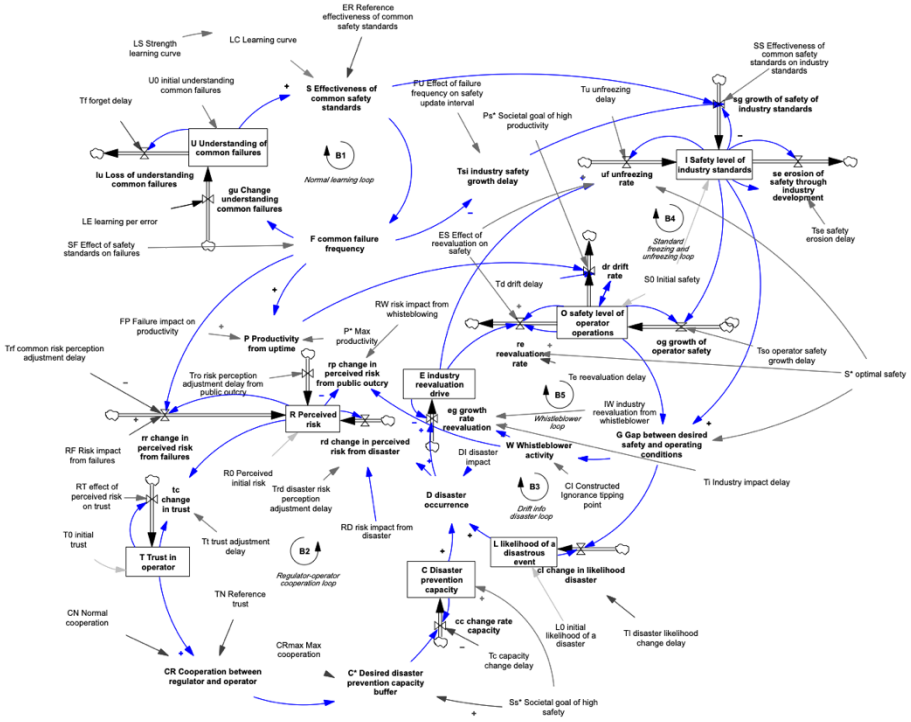


Figure 2: the formalized stocks & flows diagram of the dynamic hypothesis

Figure 2 provides all parameters in non-bold font, with the arrows showing their causal links in grey. Table 1 further down contains a full listing of all these parameters and their numerical values. In addition, Figure 2 provides the variable names for which a mathematical function needs to be specified in bold and black font. We will now describe these equations one by one, by order of the loops they appear in.

3. *Description of equations*

3.1. The normal learning loop (B1)

The first loop we describe is the learning loop for commonly occurring failures. In the early life stages of a new product or process, there is limited understanding of these common failures, and consequently quite some errors. Hence, frequent failures occur. This phase is often referred to as "the teething problems" phase (Bond, 1999). It is also consistent with the first phase of the so-called bathtub-type failure profile over the lifetime of many technical assets. Over time, as the number of past failures accumulates, so does the understanding of these failures. In other words learning occurs over time and accumulated experience. This learning initially attenuates common failures, and begins to plateau as the frequency of these types of failures decreases. In system dynamics terms, this is operationalized by a negative feedback loop, as shown in Figure 3 and as described in detail below.

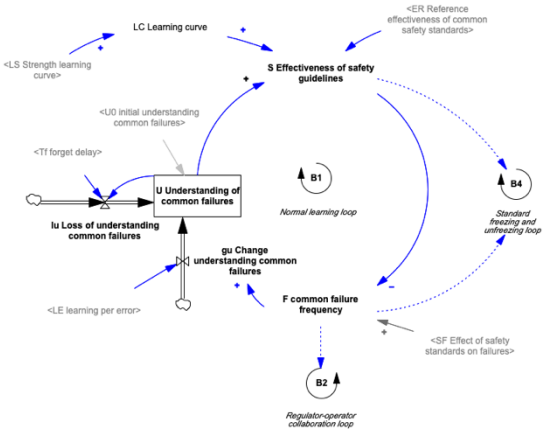


Figure 3: Loop B1: The normal learning loop

We now describe the individual variables and their equations, moving clockwise and starting with *U*. (Please refer to the figure to see what variable names refer to if not stated in the text).

Understanding of high likelihood errors (U)

The stock of understanding of high likelihood errors is *U*, initialized at U_0, with unit of measure *Knowledge*. It has two rates of change, *lu* and *gu*.

Variable *gu* refers to the learning that takes place from encountering common failures and the amount of learning that takes place as a result of overcoming these failures. This is operationalized in Equation 2 as *F*, the common failure occurrence frequency, and the constant *LF*, the amount of learning per failure:

Eq. 1 $$\frac{dU}{dt} = gu + lu$$ (Knowledge /Time)

Eq. 2 $$gu = F * LF$$ (Knowledge /Time)

Over time, organizations will lose knowledge regarding these common failures as people leave, people forget, and conditions change, etc. (Holan & Philips, 2004). This is represented by the process formalized in Equation 3, which is in System Dynamics terms a simple first order material delay:

Eq. 3 $$lu = \frac{U}{T_f}$$ (Knowledge /Time)

Here *lu* is the loss of understanding regarding common failures, and T_f is the delay over which this forgetting takes place.

Effectiveness of safety guidelines (S)

This level of understanding from scientific study of these failures is translated into safety guidelines that address these known failures. This is represented as a learning curve process (see Sterman 2000, Chapter 12, p. 507), as Equation 4 shows:

Eq. 4 $$S = \left(\frac{U}{ER}\right)^{LC}$$ (Effectiveness)

The learning curve exponent *LC* in this formula depends on the strength of the learning curve, parameter *LS*, according to the following standard learning curve formula, according to Sterman (2000), with LN standing for natural log:

Eq. 5 $$LC = \frac{LN(1 + LS)}{LN(2)}$$ (Dimensionless)

Common failure frequency (F)

There is an obvious inverse relation between effectiveness of safety standards and frequency of failures. This relation is formulized in Equation 6:

Eq. 6 $$F = (1 - S) * SF$$ (Failures/Time)

Here, SF stands for the multiplier effect of safety standards on failures, with units (*Failures/ Effectiveness/ Month*).

F then feeds back in *U* via *gu*, as specified in Equation 2, becoming a negative feedback loop: the higher the understanding of failures, the fewer failures will occur, and hence, the lower the growth in understanding. *F* also feeds into two other key constructs from two other loops, loops *B4* and *B2*. In *B2*, *F* affects the perceived risk *R*. We discuss *R* in the context of loop *B2* next.

3.2. The regulator-operator cooperation loop (B2)

Figure 4 shows the full diagram for loop *B2*, which is labeled the regulator-operator cooperation loop. Variable *F* reappears here in the top left corner in "shadow variable" format, as a key driver of one of the three inflows of R, perceived risk.

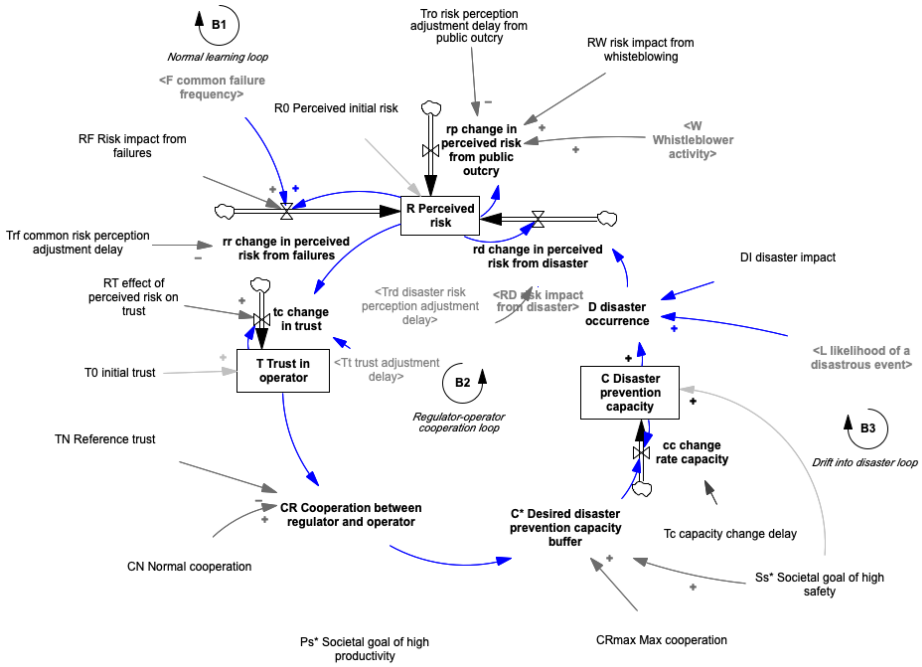


Figure 4: Loop B2, the regulator-operator cooperation loop

Perceived risk (R)

The risk that is perceived by the actors in the system has three drivers, as formalized in Equation 7 and further specified in Equations 8-10.

Equation 8 is structured as an information delay (Sterman 2000, Chapter 11.3). As Sterman formulates: "All beliefs, expectations, forecasts and projections are based on information available to the decision maker at the time, which means information about the past. It takes time to gather the information needed to form judgements, and people don't change their minds immediately on the receipt of new information." (2000, p.426). The common choice to represent such information processing is through an information delay as shown in Equation 8, where the rate of change in the perception is proportional to the gap between the current value of the input (here: *F*RF*) and the perceived value (*R*), over a time delay

(T_{rf}). We will find this structure more often in the current model, also in Equation 9, but for now a different perception process driven not by failure frequency but by the occurrence of a disaster:

Eq. 7 $$\frac{dR}{dt} = rr + rd + rp$$ (Risk /Time)

The higher the failure frequency, the higher *rr* becomes, which is the change in the perceived risk from these failures, as shown in Equation 8.

Eq. 8 $$rr = \frac{(F * RF - R)}{T_{rf}}$$ (Risk /Time)

Eq. 9 $$rd = \frac{MAX(0, (D - R))}{T_{rd}}$$ (Risk /Time)

The change in perceived risk from a disaster (*rd*), is driven by *D*, the disaster occurrence, defined in detail below. Such a disaster will increase the perceived risk level. Of course, there will only be an increase in *R* if *R* is lower than that the occurrence of the disaster warrants. When $D < R$, there should be no change in *R,* hence the MAX function in Equation 9. [This is also the case when there is no occurrence of a disaster, so long as *D* stays 0. Also, *rd* remains 0]. Again, this is an information delay, so there is a time delay of T_{rd} over which the adjustment in *rd* takes place.

Eq. 10 $$rp = \frac{MAX(0, (W - R))}{T_{ro}}$$ (Risk /Time)

The third rate of change to stock *R* is *rp*, which is the change in risk as a result of a public outcry. This is driven by *W*, the variable which denotes whistleblower activity, which will be described in detail further down in Loop *B5*. When the perceived risk associated with the amount of whistleblower activity *W* becomes greater than the current level of risk, so when $W>R$, this public outcry triggers an additional increase in perceived risk, now over the delay T_{rd}, which is the risk perception delay from a public outcry. *W* is explained further on, in the description of loop *B5*, the disaster trigger and impact loop.

Trust in Operator (T)

The lower the perceived risk, the greater the trust in the ability of the operator to work safely. This is a well-known relation in the literature on organizational trust, where this type of trust is called competence-based trust (Connelly et al., 2019; Nooteboom, 1996), denoted by *T* in this model. *T* is an accumulation, starting with T_0, and affected by of changes in its inflow *tc*, the change in trust. Again, trust accumulates from a process of perceiving behavior over time, so the structure of Equation 11 is that of an information delay:

Eq. 11 $$tc = \frac{((1 - R) * RT - T)}{T_t}$$ (Trust /Time)

We conceptualize *T* as an inverse of *R*, mitigated by *RT*, the effect of perceived risk on trust. So, the rate of change in the perception of *T* (*tc*) is proportional to the gap between the current value of the input (*1-R)*RT* and the perceived value (*T*), over a time delay (T_{rf}).

*Cooperation between regulator and operator (*C*R*)

Th higher this competence based trust *T*, the greater the level of cooperation *CR* will become. We assume a simple linear relation between *T* and *CR*:

Eq. 12 $$CR = \frac{T}{TN} * CN$$ (Cooperation /Time)

Here *TN*, which stands for a normal (reference) level of trust, and *CN*, the normal level of cooperation, are two constants that can modify this linear relation into a more or less steep inclination. This cooperation is not infinite, there is a certain maximum to it. That is operationalized in Equation 12 as CR_{max}. As this equation suggests, always $CR < CR_{max}$. The unit of measure of *CR* is *risk*, because this is really a measure of how risky the size of the buffer for unforeseen events such as potentially disastrous failures becomes.

Desired disaster prevention capacity buffer (C)*

The greater the trust, the greater the willingness to make safety standards more lax including delaying maintenance (neglecting preventive maintenance), tolerance limits., all of which exist to accommodate for unknown failure modes and or to mitigate the magnitude and impact of failure. How fast disaster prevention capacity erodes depends on how high the societal goals for safety are, labelled S_s^*.

Eq. 13 $$C^* = \frac{CR_{max} - CR}{CR_{max}} * S_s^*$$ (Risk)

3Disaster prevention capacity (C)

Over time, the desired capacity *C** will turn into the actual capacity to prevent disasters, *C*, with initial value C_0. That time is here T_c, the capacity change delay, as shown in Equation 14, which formalizes the relation of the inflow rate for *C*, the change rate for capacity *cc*, in the form of another information delay:

Eq. 14 $$cc = \frac{C^* - C}{T_c}$$ (Risk/Time)

Disaster occurrence (D)

Similar to *C**, *C* is conceptualized as a certain risk level that follows out of a certain buffer capacity. Initially, we assume this level to be at the desired societal level, so S_s^*. This becomes clearer also in the next equation, for the variable associated with the actual disaster occurrence, *D*. Normally, because of the MAX function it entails, this variable remains 0, because *C*, the disaster prevention capacity, is greater than the likelihood of a disaster event (*L*, unit of measure *risk*). However, when *L* grows beyond *C*, this variable suddenly becomes positive, and a major disaster occurs. How big the impact of this disaster will be, is determined by constant *DI*.

Eq. 15 $$D = \text{MAX}(0, L - C) * DI$$ (Impact)

D's unit of measure is *impact*, and we already saw *D* appear in Equation 9, where it closes the loop *B2*, the *regulator-operator cooperation loop*. The greater *D* becomes, the greater *rd* will become, and hence the greater *R*, where we started our description of this loop.

3.3. The drift into disaster loop (3)

Figure 5 shows the full diagram for loop *B3*, which is labeled the drift into disaster loop. Variable D reappears here in the bottom left corner. However, we will start out discussion at the variable that directly links loop *B1* with *B3*, which is *P*, the productivity from uptime.

Productivity from uptime (P)

The more frequent failures (variable *F*, from loop *B1*) occur, the lower the uptime of the assets becomes and hence the more the productivity from uptime (*P*) lowers.

Eq. 16 $$P = P^{max} - FP * F$$ (Productivity)

Equation 15 specifies a linear relation between *P* and *F*, where *F* reduces *P* from its optimum P^{max} multiplied by a factor of *FP*, the failure impact on productivity (with unit of measure *Productivity/ Failures/Month*).

Safety level of operator operations (O)

This productivity then influences *O, the* safety level of operator operations, via one of its three outflows, the drift rate *dr*.

Eq. 17 $$\frac{dO}{dt} = dr + og + re$$ (Risk/Time)

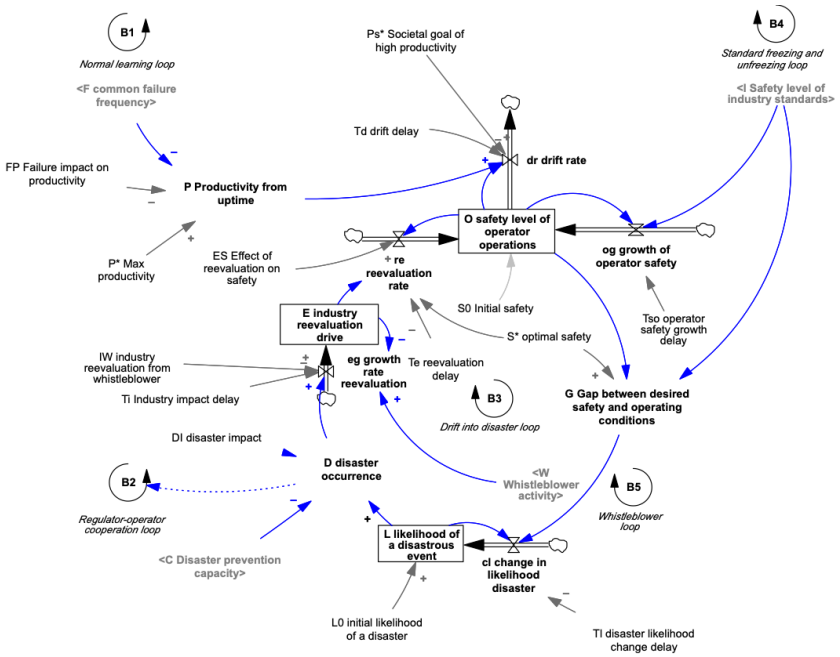


Figure 5: Loop B3, the drift into disaster loop

The outflow (drift) rate *dr* captures the phenomenon of organizational drift (Turner & Pidgeon, 1997; see also Reason, 1997), the natural tendency to move gradually towards higher efficiency, at the (unintended) expense of safety.

P affects *dr* as can be seen in Equation 18.

The first fraction formalizes the remaining productivity improvement potential, which only becomes 0 when $P = P_s^*$. Here, organizations can still improve productivity and tend to do so at the expense of *O*, the safety level of operator operations. The greater the *dr*, the greater the drop in *O*. However, how quickly this gap is closed depends on time delay *Td*, the drift delay constant.

As shown in Equation 16, *O* has three rates affecting it, not just *dr*. From loop *B4*, which we will discuss later, comes the main driver for the second flow, the growth of operator safety, *og*.

Eq. 18 $$dr = \frac{P_s^* - P}{P_s^*} * \frac{O}{Td}$$ (Risk/Time)

Eq. 19 $$og = \frac{I - O}{T_{so}}$$ (Risk/Time)

Equation 19 proposes that the level of safety of operations that the operator achieves (*O*) follows, in a lagged manner, the level of the safety standards in the industry (*I*). (We will describe *I* further on, as part of loop *B4*). As inherent to an information delay, which Equation 19 is, this happens over a certain delay, which is called T_{so}, the operator safety growth delay.
The third rate of change for *O* is the realign rate *re*, with its formalization shown in Equation 20. Rate *re* becomes greater than 0 when *E* is activated, which stands for the industry reevaluation of safety practices that tends to take place after a major disaster has occurred. Here T_e stands re-evaluation delay. for How big the impact of such a reevaluation will be on safety is determined by the factor *ES*, the effect of reevaluation on safety, with dimensions *risk/impact*.

Eq. 20 $$re = \frac{\text{MIN}(E * ES * O, S^* - O)}{T_e}$$ (Risk/Time)

This formula suggests that *O* increases by a certain fraction of itself, over a certain time period. The MIN function is only there to assure that, even in extreme situations, *O* will not become unrealistically large through large growth rates from *re*. *O* cannot become greater than the optimal safety level S^*, so $E*O <= S^*-O$.

Gap between desired safety and operating conditions (G)
S^* reappears in the next equation. The bigger the gap between the actual safety of operator conditions and the desired safety, which has S^* as its maximum value, the greater the likelihood that some disaster may occur. This causal relation is formalized in Equation 21.

Eq. 21 $$G = S^* - (O * \frac{I}{S^*})$$ (Risk)

G is the gap between the actual state of affairs and a theoretical optimal safety level S*. This consist of two elements: (1) perfect industry standards are (2) perfect translation of those in operator operations. A gap G occurs when these two are not perfect. We assume that even when industry standards are at 1.0, they would still not be perfect (Simon, 1957: 198), so we divide I by S* to correct. To capture the inherent interdependencies of these variables, the equation employs a multiplicative function of the effect of *O* times the effect of *I,* as recommended by Sterman (2000: 525-528).

Likelihood of a disastrous event (L)
As explained above, the greater the gap between desired and actual safety, the greater the likelihood that some disastrous failure event will occur. This is formalized by the stock variable *L*, with initial level L_0 and rate of change *cl,* the change in likelihood of a disastrous event.

Eq. 22 $$\frac{dL}{dt} = cl$$ (Risk/Time)

As Equation 23 shows, *L* trails *G* over time, over time interval T_l to be precise, in yet another information delay:

Eq. 23 $$cl = \frac{(\text{G} - L)}{T_l}$$ (Risk/Time)

Disaster occurrence (D)
Once *L* becomes too great, an actual disaster occurs. This effect is captured by the equation for *D*, the actual disaster occurrence, and we already specified this relation in Equation 15 while describing loop *B2*:

Eq. 15 $$D = \text{MAX}(0, L - C) * DI$$ (Impact)

Industry reevaluation (E)
Loop *B2* described what happens after such a disaster with the cooperation between regulator and operator, and how that leads to higher investments in disaster prevention capacity, so the capacity to limit the impact of a failure after it has occurred. Loop *B3* describes what happens with the safety measures at the operator that can help prevent such potentially disastrous failures from occurring. After major disaster, many practices are reevaluated, leading to safer operations. We formalize the industry evaluation drive *E* as an information delay in equations 24 and 25:

Eq. 24 $$\frac{dE}{dt} = eg$$ (Risk/Time)

Eq. 25 $$eg = \frac{(DI * D + IW * W) - E}{T_i}$$ (Impact)

Multiple actual occurrences of whistleblower activity at specific points in time and/or a disaster will lead to a sustained drive over this time interval T_i to reevaluate industry practices. The formulation of Equation 25 achieves this sustained behavior by implementing a moving average over T_i, the industry impact delay. The rate of change *eg* is driven by two effects are combined:

- The effect of whistleblower activity, formalized by *IW*W*, with *W* the level of whistleblower activity (with unit of measure *Communication*) and IW the amount of industry reevaluation as a result of this activity (unit of measure *Impact/Communication*).
- The effect of an actual disaster, formalized by *DI*D*, with *D* standing for the actual disaster occurrence (unit of measure *Risk*) and *DI* for the disaster impact (unit of measure *Impact/Risk*).

Variable *E* feeds back into *O* as shown earlier on in Equation 20, which closes this third feedback loop.

3.4. The standard freezing and unfreezing loop (B4)

Figure 6 shows the relevant variables and causal links for loop *B4*, the standard freezing and unfreezing loop. This becomes a loop because it starts from level *E* and feeds into rate equation *eg*, which are both connected as variables within loop *B3*, as visualized in Figure 6. A central variable in *B4* is *I*, the safety level of industry standards. This is strongly affected by two variables from loop *B1*, the normal learning loop: *S* and *F*. We will start with a variable that connects *F* with *I*, which is T_{si}, as shown in the top left corner of Figure 6.

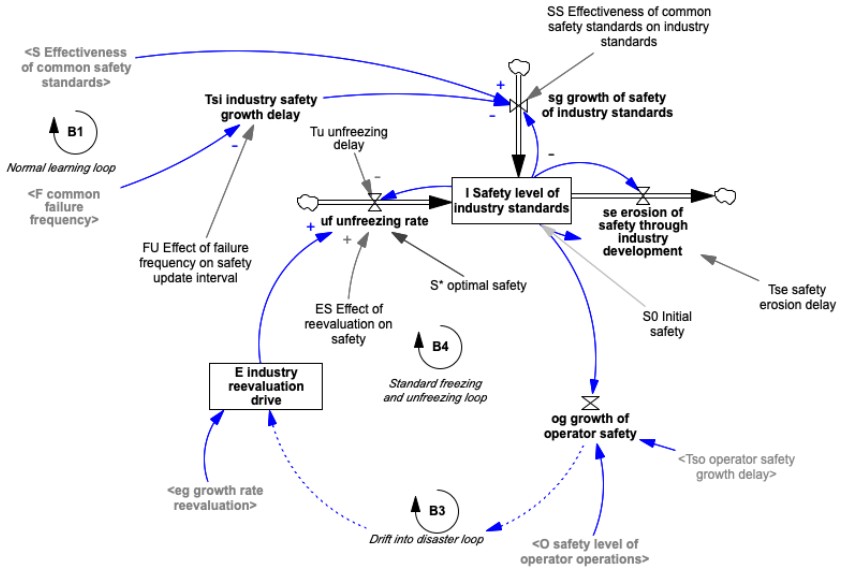


Figure 6: Loop B4, the standard freezing and unfreezing loop

Industry safety growth delay (T_{si})

We hypothesize that the less frequently common failures still occur, the slower changes in industry standards are made. This slow change corresponds with the phenomenon that after the teething problems stage of a new technical asset has passed and all the obvious failures that can occur indeed occurred and have found their way into improved industry-level safety procedures, these safety procedures start to "freeze". As shown in Equation 26, we implement this relation between *F* and T_{si} as an inverse relationship, attenuated in strength by a variable *FU*, which stands for the effect of failure frequency on the safety update interval.

Eq. 26 $$T_{si} = \frac{FU}{F}$$ (Time)

Safety level of industry standards (I)

Whereas *O* refers to the safety level specific to the operator, *I* covers the entire industry, and to the formal and legal safety standards and regulations that apply to it. Similar to *O,* level *I* (initialized at level S_0) is affected by three flow rates, as shown in Figure 6 and in Equation 27:

Eq. 27 $$\frac{dI}{dt} = sg + se + uf$$ (Risk/Time)

The time interval T_{si} that we defined in Equation 26 affects the rate equation for *sg*, as formalized in equation 28:

Eq. 28 $$sg = \frac{\text{MAX}(0, SS * S - I)}{T_{si}}$$ (Risk/Time)

Hence, the current level of industry standards *I* is updated with the gap between *I* and the effectiveness of the technical common safety standards. This latter term is *S* times *ES*, where *SS* stands for the effectiveness of these common technical standards on the actual safety level achieved by these industry standards (unit of measure *Risk/Effectiveness*). When the initial safety level S_0 is fairly high, *sg* will initially turn negative, which makes no sense, hence the max function to prevent it from

doing so. This adjustment takes place over time interval T_{si}, which, as indicated, tends to become longer, and so the adjustment times longer, leading to a process we call freezing of standards.

The effect of the progressing of time is also crucial in the second rate equation for *I*, se. Here, the progress of time leads to what we label as an *erosion* of safety. This process is formalized as a draining process, or a 1st order material delay in SD-terms (Sterman, 2000), in Equation 29:

Eq. 29 $$se = \frac{I}{T_{se}}$$ (Risk/Time)

As time passes by, technologies and product types evolve, new components are added, new usage modes become more prevalent, etc., a certain percentage of the existing safety standards becomes outdated. When T_{se}, the safety erosion delay, is large, this is obviously a small percentage, but some erosion of safety quality appears nonetheless inevitable.
The third rate equation becomes active when there is an active industry reevaluation drive, so when *E* has become >0 as a result of resent whistleblower activity or an actual disaster. Then, safety regulations from past decades are revaluated, and looked upon with fresh eyes, leading to an unfreezing of safety regulation and standards. This is *uf*, the unfreezing rate, with its rate equation specified in Equation 30:

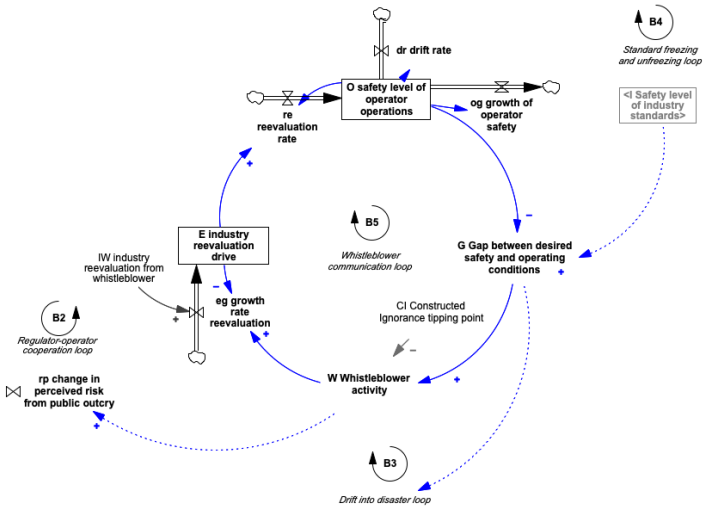
Eq. 30 $$uf = \frac{\text{MIN}(\ ES * E * I, S^* - I,\ )}{T_u}$$ (Risk/Time)

Here, *I* is increased by a certain percentage, as a result of *E* becoming >0. That percentage equals ES**E*, with *ES* the effect of revaluation on safety. The MIN function in Equation 30 is there to prevent *I* from growing greater than S^*, the theoretical maximum safety. So, *uf* cannot be greater than the room for improvement that still exists between S^* and *I*.
All the other relevant equations for loop *B4* have already been specified. Equation 18 describes how *I* affects *og*, the growth of operator safety *O*. *O* then affects causality in the rest of loop *B3*, leading to a rise in *E*, which feeds back in into *I* via *uf*, as shown in Equation 30 above. This then closes loop *B4*.

3.5. The whistleblower communication loop (B5)

The fifth and final balancing feedback loop in this model is visualized in Figure 7. This loop is called the whistleblower and communication loop. The point of this loop is that the general public does not hear about increasingly unsafe operational practices at the operator until a disaster happens. Inside the operator, obviously some agents are aware that things are not going as they should (Johnson, 2003). When their calls for increased safety are ignored, at what point will they decide to communicate their concerns publicly? In many, perhaps in most, instances or major disasters, there were advance warnings, from outside agencies but also quite often from agents within the organization affected. This is why we model this feedback loop as an integral part of the Confidence Trap phenomenon. Pre-disaster whistleblower activity is not a special case, it appears to be the rule with most disasters that occur. Unfortunately, by definition in those cases the whistleblowing has not been effective, since the disaster did happen, nonetheless.

Figure 7: Loop B5, the whistleblower communication loop



Whistleblower activity (W)
As becomes apparent from looking at Figure 7, from a modeling perspective, whistleblower activity (W) works as a "shortcut" in the causality of feedback loop *B5*. *W* is triggered by the same increase in *G* as actual disaster occurrence does (*D*), and it leads to the same effects as *D*, i.e., a growth in *E*. However, *W* does not need an actual disaster to happen to become effective;, it can be triggered sooner. As Equation 31 shows, this trigger is *CI*, the constructed ignorance tipping point. When *G* becomes greater than *CI*, the whistleblower no longer finds it possible to look the other way.
W feeds back into E, via *eg*, thereby closing loop B5. We showed this in Equation 25. *W* also feeds into *R* in loop B2, via *rd*, as specified in Equation 10, and visualized in Figure 7.

This concludes the descriptions of the equations in the model. The model also contains 43 constants. The values and units of measure of these are listed in the next section.

Eq. 31 $$W = \begin{cases} G > CI : 1 \\ G \leq CI : 0 \end{cases}$$ (Communication)

4. *Parameter listing*

Table 1. Parameter Values

Name	*Symbol*	*Type*	*Value*	*Unit*
FINAL TIME	*<builtin>*	*built*	*120*	*Month*
INITIAL TIME	*<builtin>*	*built*	*0*	*Month*
TIME STEP	*<builtin>*	*built*	*0.25*	*Month*
Constructed ignorance tipping point	*CI*	*tip*	*0.72*	*Risk*
Max cooperation	C_{max}	*max*	*1.00*	*Cooperation*
Normal cooperation	*CN*	*ref*	*0.50*	*Cooperation*
disaster impact	*DI*	*trans*	*100*	*Impact/risk*
Reference effectiveness of common safety standards	*ER*	*ref*	*1.00*	*Knowledge*
Effect of reevaluation on safety	*ES*	*trans*	*1.00*	*1/Impact*
Failure impact on productivity	*FP*	*trans*	*1.00*	*Productivity/Failures/Month*
Effect of failure frequency on safety update interval	*FU*	*trans*	*12*	*Failures*
Industry revaluation from whistleblower	*IW*	*trans*	*1.00*	*Impact/Communication*
Initial likelihood of a disaster	L_0	*init*	*0.00*	*Risk*
Learning per error	*LE*	*trans*	*0.10*	*Knowledge/Failures*
Strength learning curve	*LS*	*aux*	*0.10*	*Dimensionless*
Max productivity	P^*	*max*	*1.00*	*Productivity*
Societal goal of high productivity	P_s^*	*max*	*1.00*	*Productivity*
risk impact from disaster	*RD*	*trans*	*0.50*	*Risk/Communication*
risk impact from failures	*RF*	*trans*	*1.00*	*Risk/Communication*
effect of perceived risk on trust	*RT*	*trans*	*1.00*	*Trust/Risk*
risk impact from whistleblowing	*RW*	*trans*	*1.00*	*Risk/Communication*
perceived initial risk	R_0	*init*	0.80	*Risk*
Optimal safety	S^*	*max*	*1.00*	*Risk*
Initial safety	S_0	*init*	*0.70*	*Risk*
Effect of safety standards on failures	*SF*	*trans*	*1.00*	*Failures/Effectiveness/Month*
Effect of common safety standards on industry standards	*SS*	*trans*	*1.00*	*Risk/Effectiveness*
Societal goal of high safety	S_s^*	*max*	*1.00*	*Risk*
Initial trust	T_0	*init*	*0.10*	*Trust*
capacity change delay	T_c	*delay*	*24*	*Month*
drift delay	T_d	*delay*	*12*	*Month*
reevaluation delay	T_e	*delay*	*12*	*Month*
forget delay	T_f	*delay*	*120*	*Month*
industry impact delay	T_i	*delay*	*3*	*Month*
disaster likelihood change delay	T_l	*delay*	*24*	*Month*
Reference trust	*TN*	*ref*	*0.80*	*Trust*
disaster risk perception adjustment delay	T_{rd}	*delay*	*3*	*Month*
common risk perception adjustment delay	T_{rf}	*delay*	*24*	*Month*
risk perception adjustment delay from public outcry	T_{ro}	*delay*	*3*	*Month*
safety erosion delay	T_{se}	*delay*	*120*	*Month*
operator safety growth delay	T_{so}	*delay*	*24*	*Month*
trust adjustment delay	T_t	*delay*	*12*	*Month*
unfreezing delay	T_u	*delay*	*6*	*Month*

initial understanding common failures	U_0	*init*	*0.1*	*Understanding*

These 43 parameters are of different types:

- 13 Time delays regulate at what speed levels change as a result of changes in their rate equations;
- 5 parameters set upper or lower limits for variables, denoted with an * superscript or *max* subscript;
- 3 parameters indicate a reference, or normal value for the variable it refers to;
- 5 Initial values initialize the levels in the model, where these are not either 0 or some target level;
- 8 auxiliary variables help a stock in dimension A to transition into a stock in dimension B, so usually with the unit of measure B/A;
- 1 threshold parameter indicates a tipping point;
- 1 dimensionless auxiliary variable is used to calculate an exponential function;
- 3 built-in parameters regulate the time flow in the simulation.

Regarding the values chosen for these parameters, the following:

- For different sets of values for these parameters, the model can exhibit very different behaviors. These values have been chosen so that the model's behavior mimics the empirical behavior as described in the main paper, such as the Alaska Airways disaster.
- Apart from that, as simple as possible values have been chosen. Specifically:
    - The time delays have monthly equivalents that correspond to feasible real-world delays, we conjecture often in multiples of 3 of 6 to suggest quarters or years.
    - The minimum and maximum levels are either 0 or 1, which is also the range for most variables that refer to an intangible construct such as Trust, or Safety, and of course also Risk.
    - Some specific parameters, such as the ones for the tipping point *CI*, have values chosen specifically to generate the history-friendly behavior referred to above.
- Further down in this documentation we will systematically change the values of the most important parameters in the sensitivity analysis to check for possible modeling flaws and to search for real-world policy levers.

5. Base case behavior

We now describe the behavior of this model in what we call the base case, with the parameter values that we judge to be a fair approximation of a hypothetical case of a man-made disaster, and one that is history-friendly (Capone et al., 2019; Malerba et al., 1999) to the Alaska Airways case. This base case is without additional policies and so without changes in parameter values as listed in Table 1. We simulate one decade (120 time units of months) of system behavior.

5.1. *Base case behavior of Loop B1, the normal learning loop*

We start with the straightforward behavior of the key constructs in loop B1. These are summarized in Figure 8:

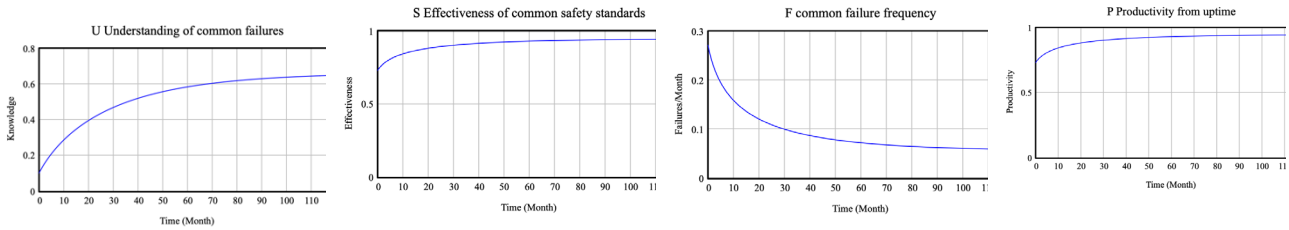


Figure 8: Key dynamics in loop B1, the normal learning loop

All these dynamics are driven by learning curve theory. The more failures have been analyzed, the more one learns, but less and less as knowledge matures. As mentioned in Section 1, the relatively straightforward curves from Figure 8 become inputs for much less straightforward dynamics in the other loops.

5.2. *Base case behavior of Loop B2, the regulator-operator cooperation loop*

The relative absence of failures leads to the mistaken belief in the industry, both with the operator and the regulator, that it is proof of the safety of the system. Consequently, the perceived risk reduces, as shown in the far-left of Figure 9. Later in the simulation, perceived risk increases twice. First, around month 52, as a result of whistleblower activity. Next, around month 64, the occurrence of a disaster increases risk perception massively, as this graph shows. However massive, this increase is only temporary. Lower perceived risks lead to more trust of the regulator in the competence of the operators (Connelly et al., 2018; Nooteboom, 1996), and hence in more cooperation between the two actors, as the mid-left graph of Figure 9 illustrates. This leads to a lower desired amount of disaster prevention capacity or erosion of standards to prevent dangerous events turning into disasters, as the bottom-left graph shows. At a certain time, this capacity becomes so low that a disastrous event becomes inevitable, as the graph for disaster occurrence in the far left corner of Figure 9 proves.

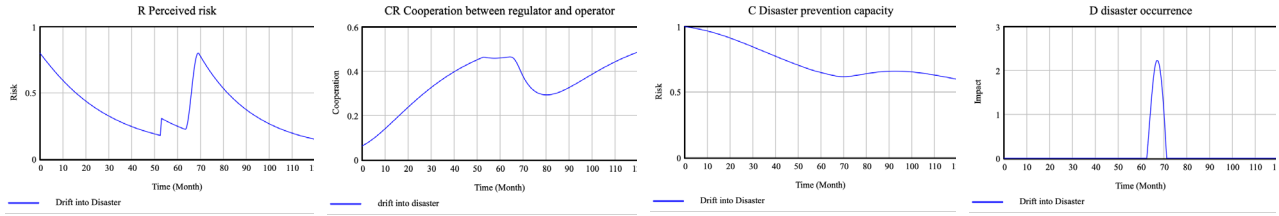


Figure 9: Key dynamics in loop B2, the operator-regulator cooperation loop

5.3. *Base case behavior of Loop B3, the drift into disaster loop*

The graphs in Figure 10 clearly show the two very distinct behaviors of the system over time, the pre-disaster and the post-disaster phase. During the pre-disaster phase, *O*, the safety level of the operations at the operator, steadily but slowly declines, as shown in the top-left graph. As a result, as shown in the mid left of Figure 10, the gap between how safe the operating conditions *should* be and how safe they actually *are* grows steadily, which makes the likelihood of a disastrous event ever greater as shown in the middle-right corner graph. The occurrence of an actual disaster, around month 63, which we already saw in Figure 9, then becomes inevitable. This is the moment when *E*, the industry reevaluation drive really takes off, as the far-right graph of Figure 10 indicates. Actually, there was already some movement in *E*, from month 50 onwards, as a result of (ineffective) whistleblower activity, which we will explain in more detail further on.

The behavior of these graphs in the post-disaster period is radically different from the pre-disaster period. The industry reevaluation drive *E* peaks and dies out some 30 months periods later. This reevaluation leads to a strong improvement in *O*, the safety level at the operator level, and hence a much lower gap in safety of operating conditions, and a much lower likelihood of a disastrous event. Over time, this effect will erode, as the later stages of this graph suggest, but for the time being conditions are sufficiently safe.

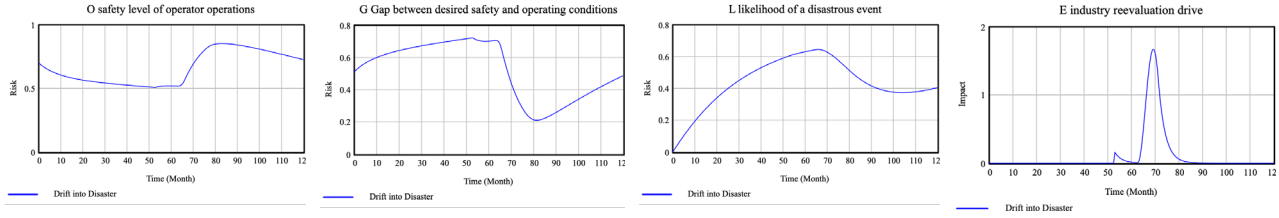


Figure 10: Key dynamics in loop B3, the drift into disaster loop

5.4. *Base case behavior of Loop B4, the standard freezing and unfreezing loop*

Loop B4 describes the dynamics of safety standards at the industry level. The core construct here is *I*, the safety level of industry standards, shown in the far-left corner of Figure 11. The value of *I* is determined by three separate rate of change equations, which also appear in Figure 11. Here, it becomes clear that the fundamental freezing of these standards during the first phase of the simulation is not really affected by a disaster occurrence, (since no disaster occurs) as the top-right graph shows. Also, the erosion of safety, in the mid-left corner, resembles an autonomous process, as a result of ongoing changes in the operational environment of the asset, its use and its components. Pre-disaster, these two autonomous effects lead to a gradual decline of the industry safety level. Post-disaster, the unfreezing rate *uf*, as shown in the mid-right of Figure 11 is dominant in driving the behavior of *I*. This is because of the significant unfreezing of the existing corpus of rules and regulations that is taking place, post-disaster.

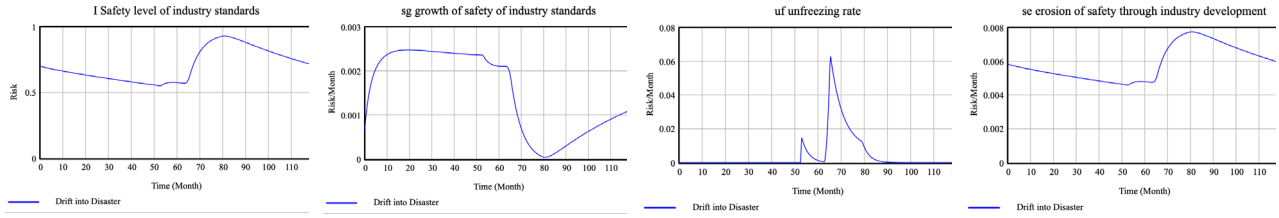


Figure 11: Key dynamics in loop B4, the standard freezing and unfreezing loop

5.5. *Base case behavior of Loop B5, the whistleblower loop*

As the left-hand graph in Figure 12 shows, there is one burst of whistleblower activity, starting in period 52, relatively shortly before the actual disaster. The right-hand graph shows why this is. It is at that time that the tipping point for such whistleblowing is passed. The tolerance for constructed ignorance is exceeded – the whistleblower is no longer able to maintain that one is better off not telling the outside world about what is happening.

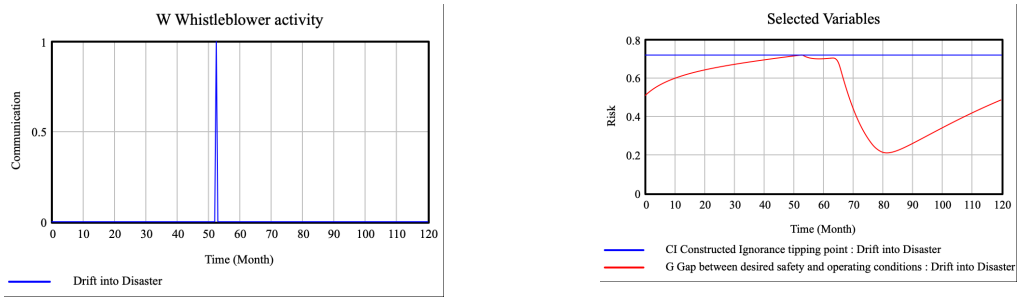


Figure 12: Key dynamics in loop B5, the whistleblower loop

As we saw earlier on, this communication has some effect, but not enough to prevent the disaster from happening. In other scenarios, with other parameter settings, this will be different.

6. *Sensitivity analysis*

6.1. *Overview*

Where are the levers for improvement? First of all, we are interested in the impact of different values for the constants chosen for the base case simulation, as shown in Table 2. For which of these values does a relatively small change have a significant impact on model behavior? More specifically, we are interested in preventing disasters, in minimizing the value of the variable *D*, the variable denoting Disaster occurrence. To what changes in parameter values is this variable especially sensitive?
Here, we leverage the analytical toolset of Vensim, the software package this model has been coded in. Figure 13 shows the results of applying Vensim's "Sensitivity2All" analysis to variable *D* (https://vensim.com/documentation/sensitivity2all.html). Figure 13 shows the mean absolute deviations (MAD) of *D* in descending order.

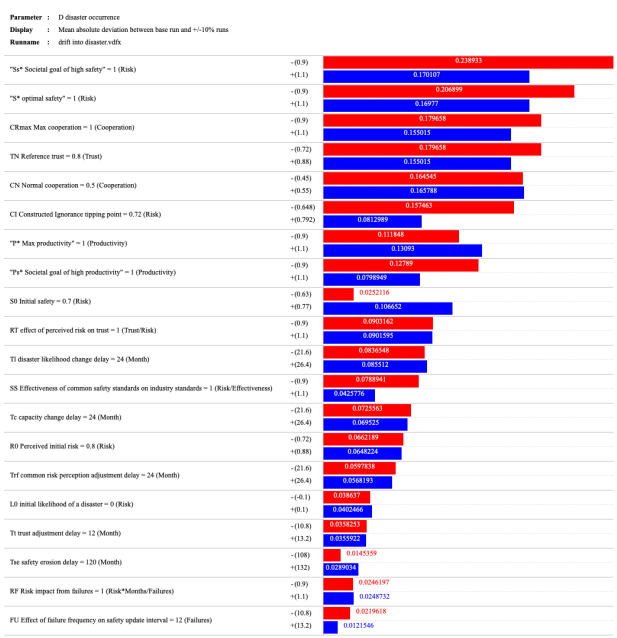


Figure 13: Mean absolute deviations of D between base runs and +/- 10% runs

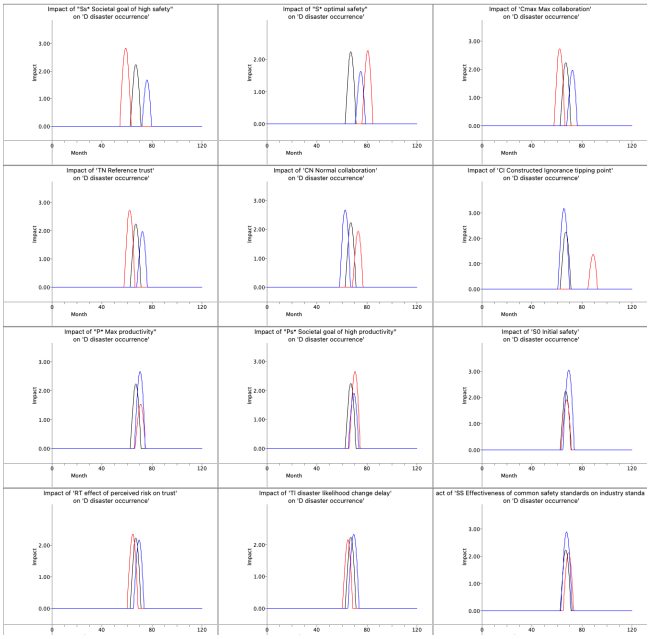


Figure 14: Graphs of behavior of D under + or - 10% changes in the top 12 constants

A first observation to make from Figure 13 is that, out of the 43 constants, only changes in at best a dozen of those appear to be significant for *D*. Zooming in on those 12, Figure 14 makes it even clearer that preventing disasters in this model through changes in constant values is not easy. In all the 24 alternate scenarios shown in the behavior-over-time graphs of the top-12 parameters, a disaster still occurs. The disaster may occur later or earlier, may be smaller or bigger in magnitude, but still it occurs.

If we look at the nature of these parameters, then several of the initially more promising ones can, by mathematical definition, not be the subject of real-world policy changes. The top-3 consists of limit values: S_s^* S^*, and C_{max}. Suggesting that these could be greater than 1, than 100%, is like suggesting that we can travel faster than the speed of light. The three next constants on the list are more helpful, as these are still susceptible to change: *TN* (reference trust), *CN* (normal cooperation), and *CI* (constructed ignorance tipping point). We will investigate at these in more detail in the next section. What about the remaining six parameters out of the top 12 in Figure 14? Parameters P^* and P_s^* are again theoretical limits, for perfect productivity and for the societal goal of high productivity, respectively, so we will ignore those as well. We will also investigate S_0 (initial safety), *RT (*effect of perceived risk on trust), and *SS* (effect of common safety standards on industry standards) in the next section.

6.2. *Specific changes in parameter values for policy design*

TN (reference trust) from 0.8 to 1.0

Figure 15 shows 2 scenarios for *TN*. The differences between these two is significant.

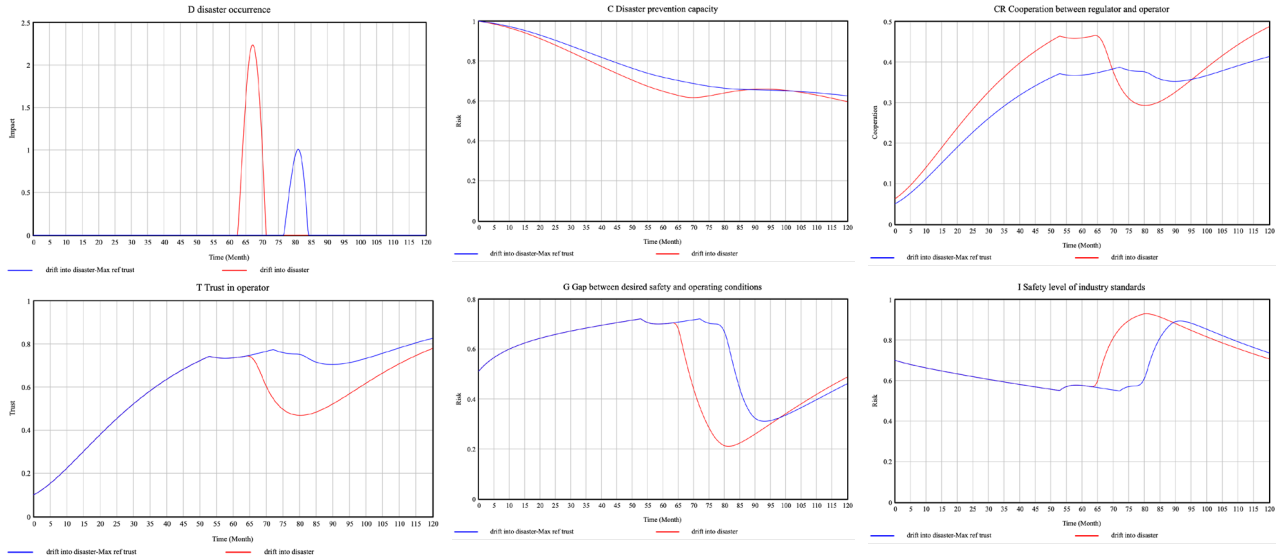


Figure 15: Key dynamics with TN reference trust from .8 to 1.0

TN operates directly on CR, the cooperation between regulator and operator, as specified in Equation 12: $CR = \frac{T}{TN} * CN$. TN cannot become greater than 1.0, but in the base case a value of 0.8 is chosen. When this is increased to 1.0, the rate of cooperation is significantly changed. In the simulation model, this has beneficial effects, as the graphs in Figure 15 show. Moving from top left to bottom right, we trace back logic. First, in the top left, a disaster still occurs but much later and less severe. Why is that? Mainly because C, the disaster prevention capacity, has eroded much slower, as the top right graph shows. And that slowing down of erosion is accomplished by a lower level of *CR*, the cooperation between regulator and operator. Different from the base case, trust *T* continues to grow, because there is no disaster happening yet. Nevertheless, this higher trust is translated into less intense forms of cooperation, because of the higher value of *TN* in Equation 12. Meanwhile, elsewhere in the system, decay of performance continues, as the two bottom graphs in in Figure 15 illustrate. The gap between desired and actual operational safety conditions keeps growing, and indeed *I*, the safety level keeps declining. Not surprisingly, there is some whistleblower activity, from month 52 onwards, which has some modest improvement effects. So, we are looking at an industry setting where failures and incidents occur quite frequently, but there are sufficient buffers in the system to absorb these and to prevent them from turning into disasters.

Apart from the limited room for improvement in this calibration of the model, the more fundamental problem with increasing *TN* is not in the model, but in the real world. What does it mean if it is normally so that operators are *completely* trusted (*TN*=1.0), just not yet this specific one? This is certainly not the world we live in today. We need to take this off the list of candidates for real-world policy interventions. Perhaps we face fewer problems with the next potential policy lever, *CN*.

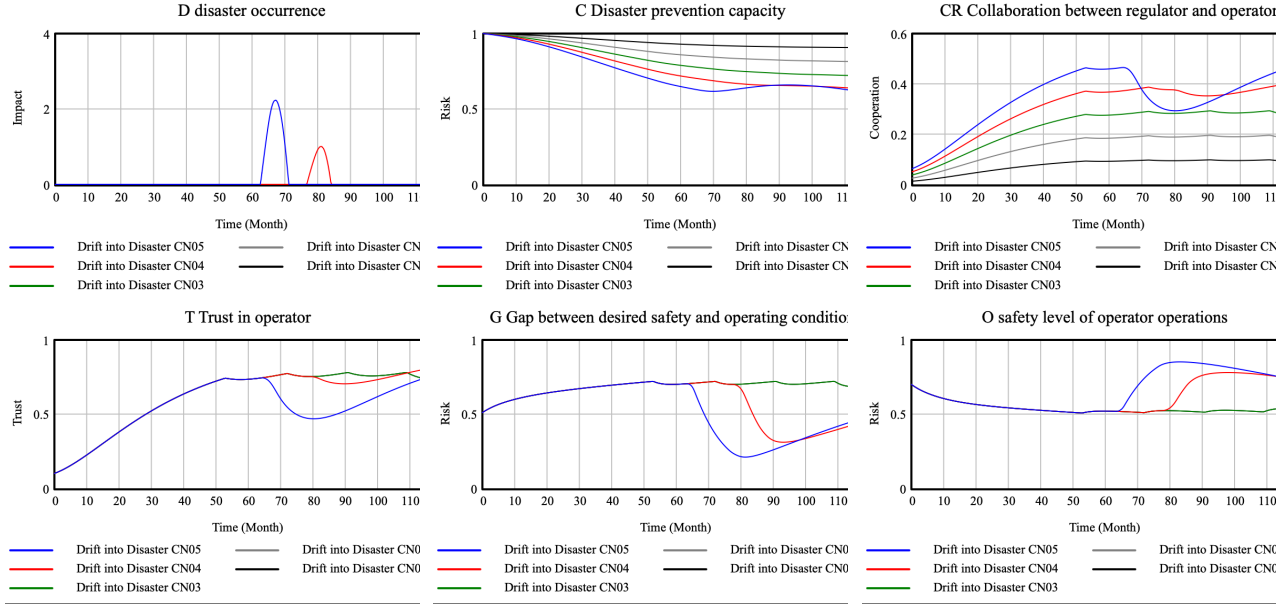


Figure 16: Key dynamics with CN normal cooperation varying from 0.5 to 0.1

CN (normal cooperation) from 0.5 to 0.1

Also *CN* operates directly on *CR*, the cooperation between regulator and operator. *CN* stands for the level of cooperation between regulator and operator that is seen as normal in the industry. In the base case, it has a value of 0.5, suggesting that some level of cooperation is seen as normal. *CN* is also active in Equation 11. This means that, in the sensitivity analysis, reducing *CN* has a similar effect to increasing *TN*, as this can, just like *TN*, dampen the amount of cooperation taking place. However, whereas we could maximally change *TN* by 25% from 0.8 to 1.0, we can change *CN* stronger, with for instance 80%, from 0.5 to 0.1. Such greater changes also lead to more greater improvements.

Figure 16 above shows five of scenarios, with values for *CN* varying from 0.5 (the base case) to 0.4., 0.3, 0.2 and 0.1. As can be distilled from the top left graph, *the three scenarios with values of CN from 0.3. to 0.1. are all successful in preventing a disaster from happening* during the ten-year simulation period. The causal logic behind this is similar from the logic behind *TN*, which after all operates in a similar manner on the same variable *CR*, as equation 11 shows. Cooperation is less, so buffer capacity declines less rapidly, which makes it possible for the operator to absorb an increasing number of incidents, as the gap between desired safety and actual operating condition remains stable, which the operator's safety level *O* also remains more or less stable.

CI (constructed ignorance tipping point) from 0.72 to 0.56

The third potential policy lever we investigate, based on the preliminary *Vensim Sensitivity2all* analysis shown in Figure 14, is *CI*, the constructed ignorance tipping point. This parameter, and the variable it feeds into, *W,* whistleblower activity, are included in the base scenario. This is because, typically, in the real world, some degree of whistleblower activity has taken place prior to the occurrence of a major man-made disaster. By definition, this activity has been ineffective, as its explicit goal at the time was to succeed in preventing a disaster. Under what conditions could whistleblower be effective? As Figure 17 shows, this depends on how quickly potential whistleblowers within the operator find themselves unable to persist in constructed ignorance. In formal modeling terms, this means on how low the tipping point *CI* is set.

The graph for *D* in the top-left of Figure 17 shows that, the lower *CI* is set (in steps of 0.04 downwards, starting from the base case value of 0.72), the later and the smaller the disaster occurs. *At a level of CI=0.56, no disaster occurs at all* in the ten-year simulation period. The top-right graph shows the tradeoff to this. In order to achieve this, the 0.56 tipping point scenario leads to 6 instances of whistleblower activity in 10 years, which seems like quite often indeed.

The reason why this high frequency of whistleblowing is needed becomes evident from the graph for *E* in the top right graph of Figure 17: every whistleblowing activity leads to some, but not to a great deal, of industry reevaluation drive. Indeed, the changes are just enough to keep *O*, the operator safety level, above the threshold level, as the mid-right graph makes clear. Meanwhile, the amount of cooperation between regulator and operator remains more or less stable. The whistleblower activities keep *R*, the perceived risk, at a reasonable level, but *T* trust remains relatively high and so *C,* the disaster prevention capacity, keeps sliding.

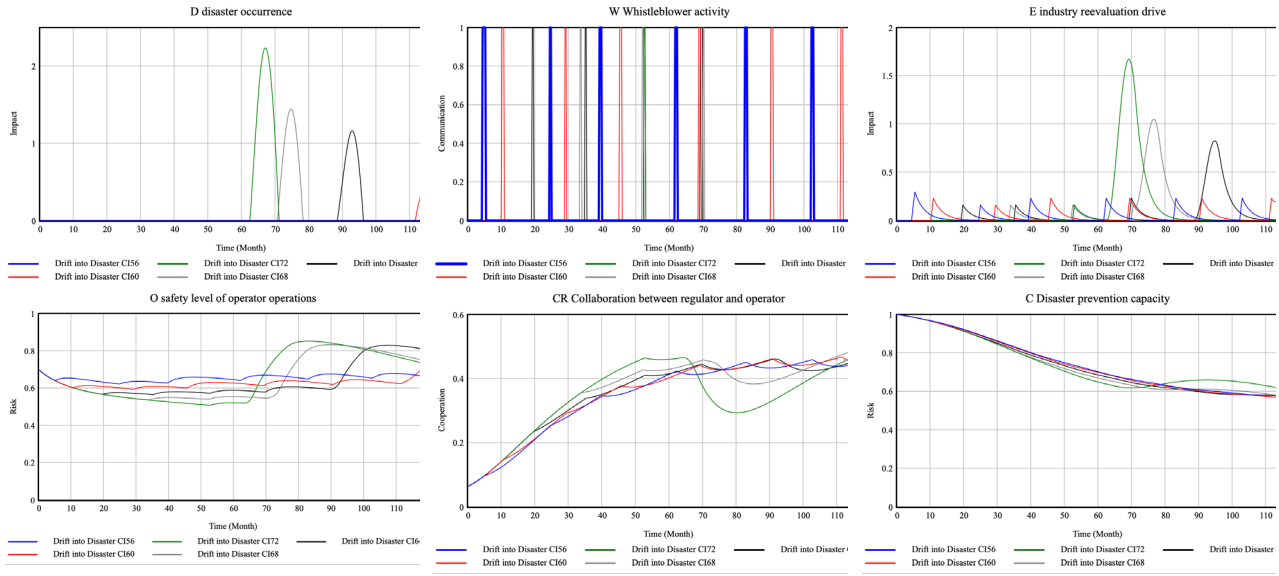


Figure 17: Key dynamics with CI constructed ignorance tipping point varying from 0.72 to 0.56

S_0 (initial safety) from 0.7 to 1.0

S_0 refers to the initial level of safety. This is set in the base case at 0.7. Even if we increase this to a theoretical maximum 1.0, then, as Figure 18 makes clear, a similar decay in safety will only slow down the drift into disaster without any fundamental further change. So, *S_0* is not suited as real-world policy lever.

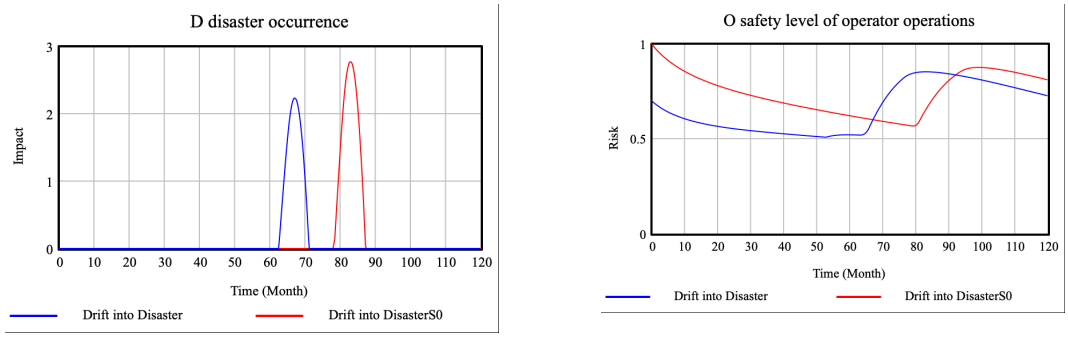


Figure 18: Key dynamics with Initial safety S0 varying from 0.7 to 1.0

RT (effect of perceived risk on trust) from 1.0 to 1.5

RT refers to the effect of perceived risk on trust. This parameter is located in loop *B2*, the regulator-operator loop. *RF* is set in the base case at 1.0. An increase in *RT* makes *T* respond more aggressively to changes in *R*, the perceived risk. If we increase *RT's* value by 50% to 1.50, then the results are mildly, as Figure 18 shows. The disaster occurs later. On the other hand, its magnitude is larger. The later occurrence is mainly because trust starts at a lower level, leading to less cooperation, and hence to a slower decline in buffer capacity.

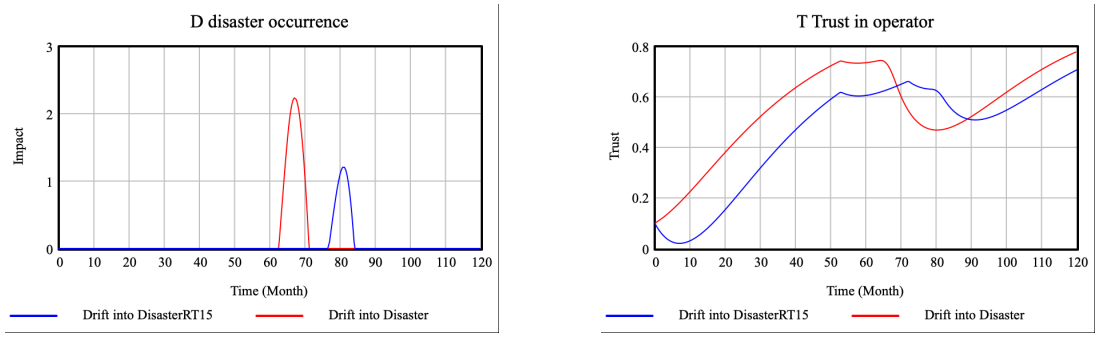


Figure 19: Key dynamics with RT (effect of perceived risk on trust) varying from 1.0 to 1.5

SS (Effect of common safety standards on industry standards) from 1.0 to 2.0

This parameter primarily works on loop B4, where it influences I, the safety level of industry standards. In the base run, *SS* is set at 1.0. If we increase this value with 50% and then with 100%, the results are beneficial indeed. This is especially true for a value of 2.0: with a value for *SS* of 1.50 or 1.75, a disaster still occurs during the 10-year simulation period, as the top-left graph of Figure 20 illustrates.

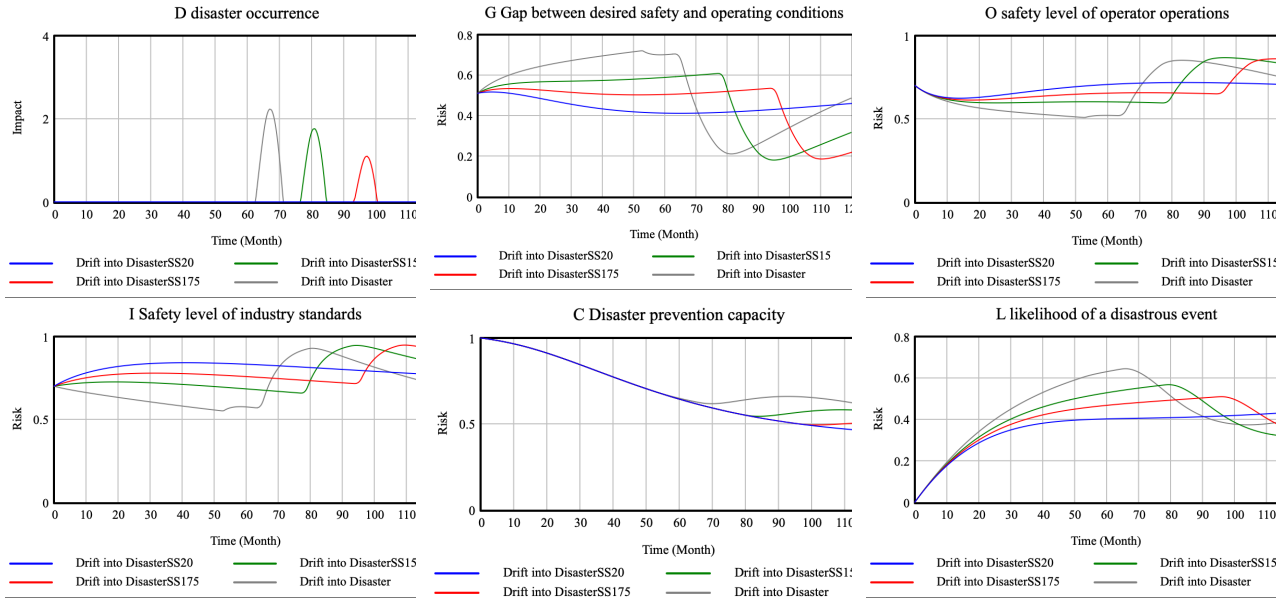


Figure 20: Key dynamics with SS effectiveness of safety standards on industry standards from 1.0 to 2.0

What is happening? *G*, the gap between desired safety and operating conditions, becomes smaller in the *SS*=2.0 scenario, as the top-middle graph shows. Why is this? Because the two variables that drive the value of *G, O* and *S*, are both much higher. There is much more learning taking place from the more reliability research in loop *B1*, the normal learning loop, when *SS*=2.0, rather than 1.0. As a result, safety measures become much more effective.

In loop *B2*, the cooperation loop, the confidence trap is still fully active, though. Failures remain rare, so trust remains high and so *C*, the disaster prevention capacity, keeps shrinking. At the end of the 10-year period, at t=120, *C* is at 0.45 while *L* is at 0.43, and so still *C*>*L* and so no disaster occurs – just yet.

6.3. *Combined scenarios for policy design*

There are many options for combining these three parameters. An integral analysis of the entire hypercube of permutations is beyond the scope of this paper. Here is one plausible combination of improvement in all three policy parameters:

SS=1.5, CI=0.65, CN=0.2

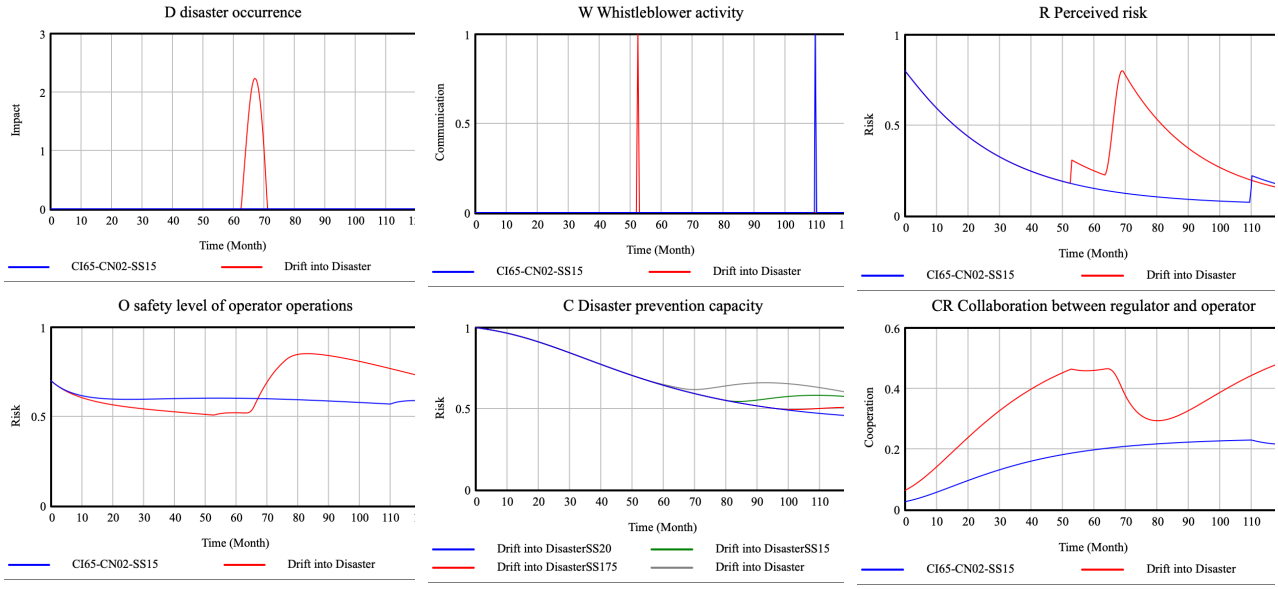


Figure 21: Key dynamics with SS effectiveness of safety standards on industry standards from 1.0 to 2.0

Clearly, this leads to good results in the ten-year period simulated. Indeed, only one burst of whistleblower activity is triggered, all the way at the end of the simulation period. The two other policy levers perform their beneficial work up to that time.

References

Connelly, B. L., Crook, T. R., Combs, J. G., Ketchen Jr, D. J., & Aguinis, H. (2018). Competence-and integrity-based trust in interorganizational relationships: which matters more?. *Journal of Management*, *44*(3), 919-945

Bond, T. C. (1999). The role of performance measurement in continuous improvement. *International Journal of Operations & Production Management*, *19*(12), 1318-1334.

Capone, G., Malerba, F., Nelson, R. R., Orsenigo, L., & Winter, S. G. (2019). History friendly models: retrospective and future perspectives. *Eurasian Business Review*, *9*, 1-23.

Holan, P. M. D., & Phillips, N. (2004). Remembrance of things past? The dynamics of organizational forgetting. *Management Science*, *50*(11), 1603-1613.

Johnson, R. A. (2003). *Whistleblowing: When it Works—and why*. London: Lynne Rienner Publishers

Malerba, F., Nelson, R., Orsenigo, L., & Winter, S. (1999). 'History-friendly'models of industry evolution: the computer industry. *Industrial and Corporate change*, *8*(1), 3-40.

Nooteboom, B. (1996). Trust, opportunism and governance: A process and control model. *Organization Studies* 17, 985-1010.

Simon, H. A. (1957). *Models of Man; Social and Rational*. New York: Wiley.

Sterman, J. D. (2000). *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Boston: Irwin/McGraw-Hill.

Turner, B. & N. F. Pidgeon. (1997). *Man-Made Disasters*. Oxford, UK: Butterworth-Heinemann.