Democracy's eternal vulnerability: Increasing resilience to disinformation by raising the two components of political truth literacy

System Dynamics Model Description

Introduction

The Dueling Loops of the Political Powerplace model (Figure 1) follows the long tradition of using small system dynamics models to find and communicate powerful insights (often counterintuitive) to the public and policy makers (Ghaffarzadegan et al. 2011), such as Jay Forrester's iconic World2 model with its 56 variables (Forrester 1971, pp.20–21). Because the Dueling Loops model (41 variables and about the same number of relationships) is easily understood and exhibits clear behavior, Ghaffarzadegan et al. argue that "important insights regarding the source of policy failures can be uncovered." The ultimate source is the unresolved main root cause.

Small insight models entail estimated parameters for exploratory use or as many measured parameters as feasible for actual cases. The Dueling Loops of model was based almost entirely on estimates. After the Truth Literacy Training study was complete, we were able to calibrate the LTQ and AAQ nodes. The values used are described later along with the simulation runs. Using modeler judgement based on system observation plus calibrating the high leverage points with study results, the Dueling Loops model was tuned to give realistic behavior over the full range of the high leverage points, false meme size, and influence per degenerate or rationalist.

System dynamics is a simulation modeling language that models the structure of a system in terms of its feedback loops and stocks, and how that structure causes behavior change over time. The goal of system dynamics is "to enhance learning in complex systems… to understand the source of policy resistance, and design more effective policies" (Sterman 2000, p.4).

A model is calibrated by measuring factors in the real world, using those values in the model, and running the model to compare its dynamic behavior to that in the real world. Model structure is then improved as needed to narrow the gap between model and real-world behavior to an acceptable level.

System dynamics models approximate a problem's behavior by comparing graphs of model behavior to graphs of collected data in the real world and refining the model until graph agreement is good enough. However, the main purpose is to generate useful insights by understanding the problem's feedback loop structure, so that the analyst knows WHY various problem behavior occurs. If a model's general behavior resembles problem behavior, and model structure makes complete sense and corresponds to the real world, then it is a useful and correctly predictive model, whether its variables are fully calibrated or not.

In system dynamics parlance, models with high calibration are said to be *quantitative models*. Those with low calibration (and hence high estimation) are *qualitative models* (Coyle 2000), as is ours. The 1971 World2 model used a mixture of estimated and measured parameters. Its successor, World3 in 1972, was fully calibrated, which increased the total number of variables from 56 to about 320. This estimate comes from

the about 320 variables in a Vensim reproduction of World3, which was originally written in the earlier modeling language of Dynamo, now obsolete.

How the model works

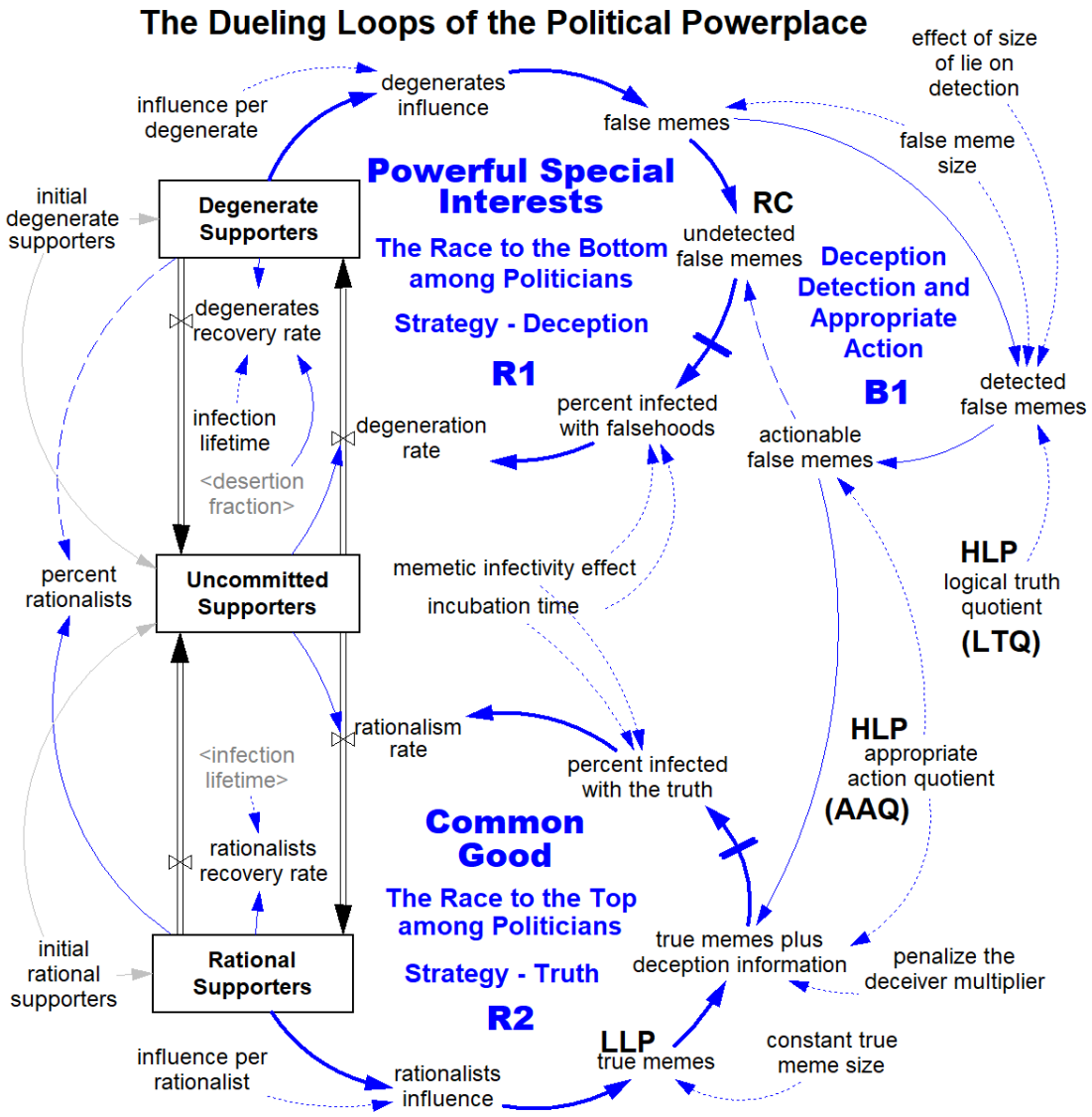


Figure 1. System dynamics model of The Dueling Loops of the Political Powerplace. Here "degenerate" is not pejorative, but signifies a person has fallen from the norm of rationality. They have degenerated into supporters of a false ideology.

As in the causal loop diagram in the paper, an arrow from node X to node Y means X causes change in Y. Solid arrows are a direct relationship, meaning as X increases so does Y, or as X decreases so does Y. Dashed arrows are an inverse relationship, meaning as X increases Y decreases and vice versa. Dotted or gray arrows are constants, indicating X remains constant. R and B signify reinforcing and balancing loops. Unit conversion nodes like one dollar and one year are not shown, as these have no effect on model behavior.

The causal loop diagram in the paper is a high-level summary of the system dynamics simulation model with 14 variables (nodes). The simulation model (41 nodes) contains the many more nodes required to simulate the model.

The three boxes are what system dynamics calls "stocks." Stocks represent the most important factors whose behavior you are trying to understand and usually correspond to

physical objects, such as people, money, and pollution. Stocks form the backbone of a system dynamics model. Objects flow between stocks according to "rates." The model uses four rates (such as *degeneration rate*) to move supporters from one stock to another.

The model mimics the flow of information we see in the real world. People are exposed to a mixture of *true memes* and *false memes* via TV, social media, articles, books, conversations, etc. This is done using Dawkins' (1976) concept of memes and memetic replication. A meme is copied information capable of affecting behavior. All memes are learned from others, either directly from other people or indirectly through a transmission medium such as books, television, or social media. Replication occurs during memetic infection. A person is "infected" by a meme when it enters and is accepted by their mind. In the model a meme is a statement that is true or false.

The concept of memetic infection is used to determine the *degeneration rate* and the *rationalism rate*. Looking at the **Race to the Bottom**, *undetected false memes* is used to calculate *percent infected with falsehoods*. After a delay of 1 year,[1] which is *incubation time*, the infection matures enough to cause the *degeneration rate*. This causes *Uncommitted Supporters* to move to the *Degenerate Supporters* stock. The **Race to the Top** works in the same manner.

People don't stay infected forever. Some eventually recover. The model handles this with the two *recovery rate* nodes. Average length of infection is 30 years,[2] which is *infection lifetime*. This causes 3.3% (1/30 = .033) of those in the *Degenerate Supporters* and *Rational Supporters* stocks to recover each year and move back to the *Uncommitted Supporters* stock.

There is one slight difference. The *rationalists recovery rate* equals *Rational Supporters* / *infection lifetime*. But the *degenerates recovery rate* equals (*Degenerate Supporters* / *infection lifetime*) plus (*Degenerate Supporters* times the *desertion fraction*). The fraction is calculated in a separate feedback loop, **Desertion due to Repulsion to Deception** (Figure 2). Testing showed the fraction is needed to mimic the way people leave right-wing parties and politicians when they detect enough deception to be repulsed and have sufficient AAQ to take appropriate action.

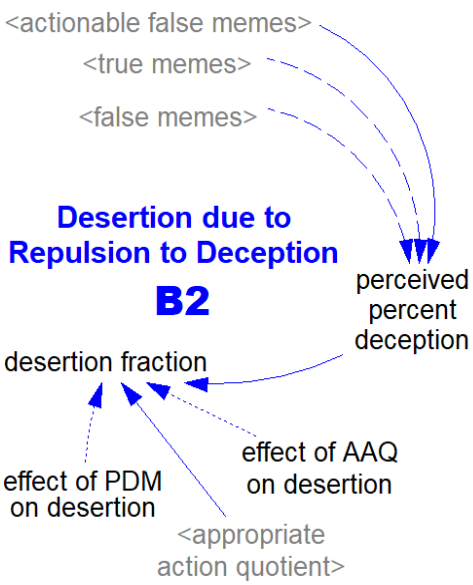


Figure 2. Variables used to calculate the *desertion fraction*. This forms part of a feedback loop that's not on the main model.

[1] We originally used a longer incubation time of 5 years. However, this caused the first five years of some of the model graphs to have distracting straight lines. Because of this we changed to 1 year. This had no effect on the model after 5 years.

[2] The PEW Research Center found that "People rarely switch parties." (Jones et al. 2020)

Let's trace how the **Race to the Bottom among Politicians** works in detail. The number of *Degenerate Supporters* time *influence per degenerate* equals *degenerates influence*. That times *false meme size* equals the number of *false memes*. Using a non-linear function in the *effect of size of lie on detection*, *false memes* times *logical truth quotient* equals *detected false memes*. That times *appropriate action quotient* equals *actionable false memes*. That times *false memes* equals *undetected false memes*. This is an important node since it represents the main root cause of democratic backsliding: the inherent advantage of the **Race to the Bottom**, aka low political truth literacy. Because it's low, *undetected false memes* is higher than it should be for the Voter Feedback Loop of democratic systems to work as designed.

The non-linear function in *memetic infectivity effect*, *incubation time*, and *undetected false memes* are used to calculate *percent infected with falsehoods*. That times the number of *Uncommitted Supporters* equals the *degeneration rate*. That causes supporters to move from the *Uncommitted Supporters* stock to the *Degenerate Supporters* stock. This closes the Race to the Bottom loop, which now begins all over again.

The **Race to the Top among Politicians** works in a similar manner with important differences. Starting at the bottom of the loop, *rationalists influence* times *constant true meme size* equals *true memes*. *Constant true meme size* is always one since the attractive power of the truth cannot be inflated.

Next, four variables are used to calculate *true memes plus deception information*. The deception information is calculated using *actionable false memes*, AAQ, and the *appropriate action quotient*. The higher any of these variables are, the higher *true memes plus deception information* is.

After that *percent infected with the truth* is calculated the same way *percent infected with falsehoods* is, except the former has *true memes plus deception information* as an input, while the latter has *undetected false memes*.

Regarding *constant true meme size* is always one: It's possible for one truth to have more attractive power than another. For example, one proposal may be more efficient than another or lead to a greater common good. Citizens may perceive this and be more attracted, all other factors being equal. But that's not because a statement, such as about a proposed policy on an important issue, is "more true." It's because the statement offers more benefits to citizens *and is true*.

Statements like this are how truth-based politicians constructively compete to see who can provide the most helpful truths (plans and positions) for optimizing the common good. For simplicity, this behavior is not modeled because its effect is dwarfed by political deception. Many other factors are also not modeled, such as demographics, the effect of wealth or education on political literacy, external events like war and large recessions, and specific deception strategies like identity politics and populist authoritarianism. An RCA-based approach to analysis focuses on what matters for the purpose of the analysis and ignores everything else.

The high leverage points work in this manner: The single high leverage point in the social force diagram in the paper, *raise political truth literacy components from low to at least medium*, becomes two high leverage points in the simulation model: LTQ and AAQ. As LTQ rises, so does *detected false memes*. That times AAQ equals *actionable false memes*. How that works to affect both loops was described above. The important behavior to understand is that raising the high leverage points affects BOTH loops. It reduces the power of the **Race to the Bottom** and strengthens the **Race to the Top**. Pushing on

the high leverage points hard enough causes loop dominance to shift from the presently dominant loop, the **Race to the Bottom**, to the **Race to the Top**.

The simulation model is a reasonable approximation of how the root cause, low leverage point, and high leverage point in the social force diagram work. Without the simulation model, it would have been *impossible* to correctly explain the fundamental layer of the social force diagram. We would have never found what appears to be the main root cause and its two high leverage points.

As simple as the Dueling Loops model looks, construction of the first version took about three years. When we began, we had no idea what would be found on the fundamental layer of the problem. Fortunately, we had the guiding hand of root cause analysis and eventually identified the basic feedback loop structure that appears to exist in all large political systems, both democratic and autocratic.

Simulation run results

This section shows how the model behaves using a series of simulation runs.

Model setting for the simulation runs

Model Settings	Simulation Runs Run 5 is the problem to solve. Run 10 is the solved mode.												
	1	2	3	4	5	6	7	8	9	10	11	12	13
False meme size	1	1.5	1.5	1.5	4.9	4.8	3.4	2	1.3	1	4.9	4.9	1
LTQ – Logical truth quotient	NA	0	NA	20%	20%	30%	20%	30%	40%	50%	50%	70%	50%
AAQ – Appropriate action quotient	NA	NA	0	20%	20%	20%	30%	30%	40%	50%	50%	70%	50%
DTQ – Democratic truth quotient	NA	0	0	4%	4%	6%	6%	9%	16%	25%	25%	49%	25%
Influence per degenerate	200	200	200	200	200	200	200	200	200	200	200	200	300
Results													
Percent rationalists at end of run	50%	41%	41%	42%	34%	38%	42%	43%	49%	50%	82%	94%	41%

Table 1. Simulation run settings and results. Before a run, the settings are set to the values shown. The model is then simulated. Results are then measured.

Each model simulation run is a logical experiment. The result tells us how certain things in the real world can be expected to behave, given particular starting conditions. This form of experimentation is orders of magnitude faster and cheaper than real-world experimentation. This explains why system dynamics modeling (and other types of modeling) is so useful, not just on business problems (where the tool was born), but in the social sciences where real-world experimentation is often slow, expensive, or impossible.

Experiment inputs are the constants changed. These are *false meme size,* LTQ and AAQ. While listed as a model setting, DTQ is not on the model but is calculated in the table for greater understanding. DTQ equals LTQ times AAQ.

The output of interest is *percent rationalists*, calculated by:

percent rationalists = Rational Supporters /
(Rational Supporters + Degenerate Supporters)

Percent rationalists measures dominance of the **Race to the Top** loop. The higher the percent, the more the system "wants" to solve common good problems, because fewer people have degenerated.

Table 1 has 13 simulation runs. All begin with 40 *Degenerate Supporters*, 40 *Rational Supporters*, and 20 *Uncommitted Supporters*, giving a total of 100 supporters. In all runs *influence per degenerate* or *rationalist* are equal and never changed. When a simulation run begins, neither side has an advantage except that provided by the model settings.

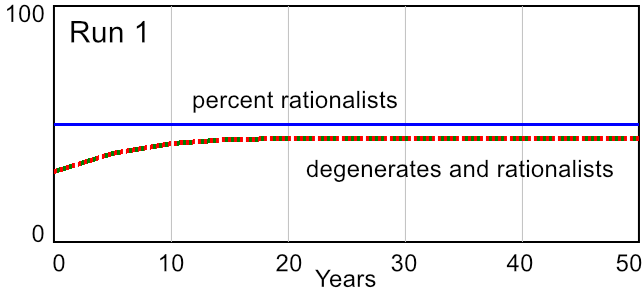
The model is so simple it's an equilibrium model. At the end of a simulation run, the graphs all show dynamic behavior has reached an equilibrium or is on its way there.

We developed a more complex Dueling Loops model that shows cyclic behavior, as **Race to the Bottom** and **Race to the Top** dominance changes. A cycle begins with the **Race to the Bottom** dominant. After a delay, this causes lowering of the common good, which causes faster desertions from the *Degenerate Supporters* stock and less movement to it, because citizens see though the deception used to fool them and their DTQ rises. This eventually causes a mode change to where the **Race to the Top** becomes dominant. The common good grows again. But citizens forget how they were

fooled and their DTQ drops, just as for example so many people have forgotten the horrors of fascism in World War Two and are moving to the far-right. This causes the **Race to the Bottom** to become dominant again, and another cycle starts. This behavior is not needed for this paper and would make the model much harder to understand.

Runs 1 to 9

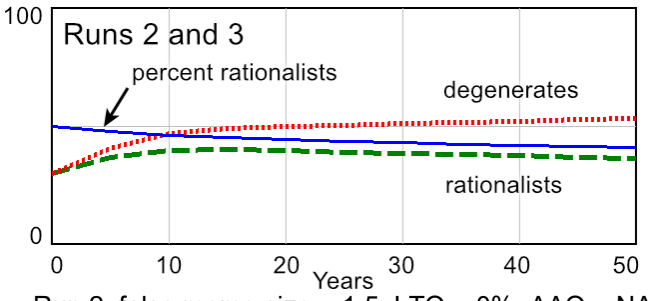
Run 1. The first run shows how when neither side (rationalists and degenerates) has an advantage, *percent rationalists* stays unchanged at 50%. The number of rationalists and degenerates rise evenly as some neutralists move to the other two stocks. Neither side has an advantage since *false meme size* = 1 and *logical truth quotient* (LTQ) or *appropriate action quotient* (AAQ) equals zero, causing DTQ to be zero.



false meme size = 1, LTQ = NA, AAQ = NA

Run 2. *False meme size* is raised from 1 to 1.5. While telling small lies offers only a small advantage, over time it accumulates into a large one. After 50 years, *percent rationalists* falls to 41%. This is enough for degenerates to win most elections.

In this run LTQ equals zero, so AAQ doesn't matter. If no *false memes* are detected, there can never be any *actionable false memes*, since *false memes* minus *detected false memes* equals *actionable false memes*.

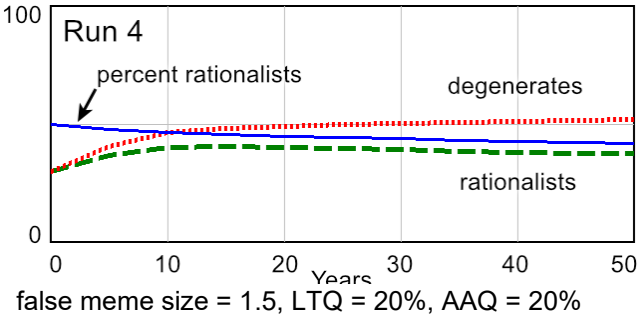


Run 2. false meme size = 1.5, LTQ = 0%, AAQ = NA
Run 3. false meme size = 1.5, LTQ = NA, AAQ = 0%

Run 3. This run keeps *false meme size* at 1.5, and instead has AAQ equal to zero. Because of this, LTQ doesn't matter, since AAQ times *detected false memes* equals *actionable false memes*.

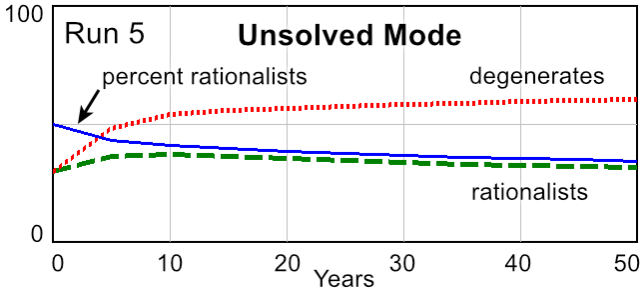
As in run 2, DTQ is effectively zero and graph behavior is identical to run 2. Runs 2 and 3 demonstrate that *both* high leverage points must be pushed on for political truth literacy solutions to work. Let's do that in the next run.

Run 4. Like runs 2 and 3, this run keeps false meme size at 1.5. However, both LTQ and AAQ are set to 20%, causing DTQ to equal 4%. While DTQ is quite low, this is enough to make a small difference. At the end of the run, *percent rationalists* has risen from 41% to 42%.



false meme size = 1.5, LTQ = 20%, AAQ = 20%

Run 5 – Unsolved Mode. This is the problem to solve.



false meme size = 4.9, LTQ = 20%, AAQ = 20%

Social agents are adaptive. Degenerate politicians are clever enough to adjust the size of lies to the optimum size: not too big and not too small. The *effect of size of lie on detection* is a lookup table (Figure 4) whose curve reflects how as the size of a lie grows, it's more easily detected and diminishing returns begin. As size rises from 1 to 50, the percent detected rises from zero to 100%. The second curve, *memetic infectivity effect*, is used to calculate *percent infected with falsehoods* and *the truth*.

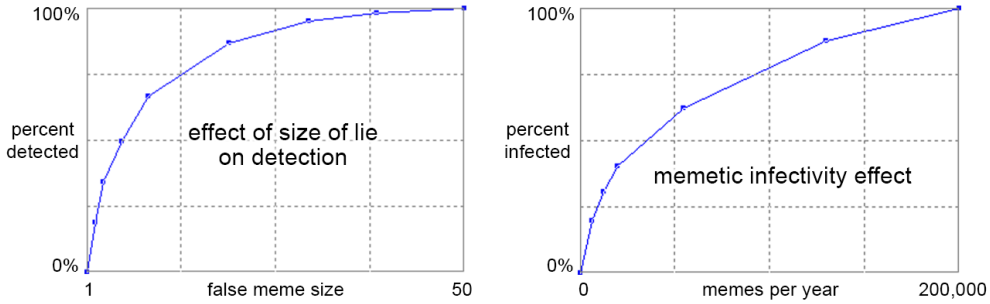


Figure 4. Lookup table values for two curves. Input is the x axis; output is the y axis. Both are non-threshold model curves (non-S curves), very similar to those observed in respiratory disease infection (Sze To and Chao 2010). The disease curves use probability of infection instead of percent detected or percent infected.

Visual experimentation with the running model shows the optimum *false meme size* is 4.9. Compared to runs 2 and 3 this gives the degenerates a much larger advantage of 34% *percent rationalists*, which is 66% degenerates.

Based on the Truth Literacy Training study and system observation, we hypothesize that in most political systems both high leverage points (LTQ and AAQ) are low, at about 20%. Run 5 thus reflects approximate real-world behavior and is what system dynamics calls the reference mode. It is the problem to solve, the run we keep referring back to as we iterate the model and design the other runs.

In run 5 the system is in the unsolved mode. Powerful feedback loop forces are locking the system into a usually dominant **Race to the Bottom among Politicians**. What's driving these forces is the unresolved main root cause. This is force R in the social force diagram.

The values of 20% for LTQ and AAQ are calibrated from group 1 (control group) in the study. Group 1 LTQ was 8% in the first study, 22% before the follow up refresh training, and 20% after the refresh training. These average roughly 20%. The AAQ scores were 25%, 59%, and 20%. The 59% is an aberration due to the large confidence intervals and the (accidentally) easier follow up statements. Thus, these also average about 20%.

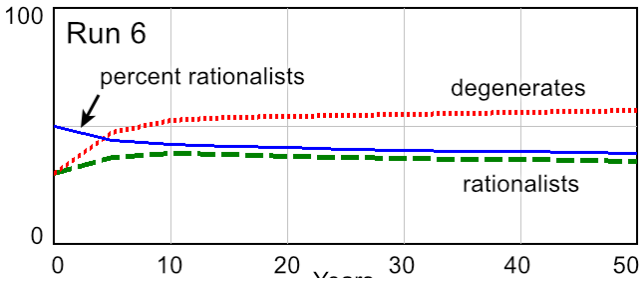
Using the World Values Survey, Noel and Therien (Noel and Therueb 2008, p.34) report that 24.7% of respondents place themselves on the left, 30.3% on the center, and 45.2% on the right. This gives a *percent rationalists* of 24.7% / (24.7% + 45.2%) = 35%, which compares favorably to the 34% rationalists in run 5. This is another form of model calibration.

Because political truth literacy is low, the **Race to the Bottom** is the dominant loop most of the time, where "Special interests now take precedence over the common

good. …we now live in a diminished democracy …with ordinary citizens squeezed out of the public sphere by partisan ideologues and professional propogandists" (Dillard and Shen 2013, p.16). The **Race to the Bottom** is not dominant all the time, because loop dominance changes back and forth due to a variety of reasons beyond the scope of this paper.

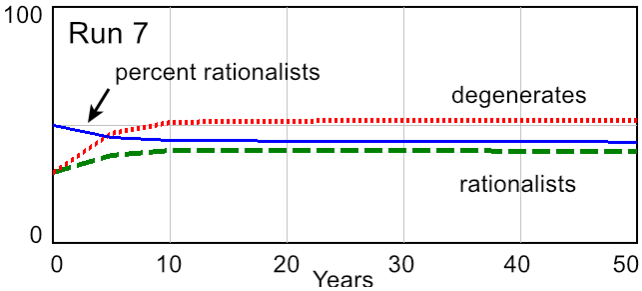
The question arises, which of the two high leverage points (LTQ and AAQ) has the highest leverage? Runs 6 and 7 use experimentation to answer this question.

Run 6. This raises LTQ from 20% to 30% and leaves AAQ at 20%. The result is an optimum *false meme size* of 4.8 and 38% *percent rationalists*.



false meme size = 4.8, LTQ = 30%, AAQ = 20%

Run 7. This sets LTQ back to 20% and raises AAQ from 20% to 30%. The result is an optimum *false meme size* of 3.4 and 42% *percent rationalists*.



false meme size = 3.4, LTQ = 20%, AAQ = 30%

Comparing the two runs, results show that pushing on the LTQ high leverage point raised *percent rationalists* 4 percentage points. But pushing on the AAQ point raised it 8 points. AAQ has roughly twice the leverage of LTQ. This is fortunate, since AAQ training is much easier than LTQ training. Let's review why.

AAQ training (aka vote training) consists of learning two simple rules:

1. **Penalize the Deceiver** – If you discover a politician has attempted to deceive you, then when you vote or take action you should strongly oppose the politician or the source of the deception. This will have the effect of reducing attempted deception. The training says: "If the claim is *false*, apply the **Penalize the Deceiver** rule and *strongly oppose* the deceiver. For example, this would have a *Very large impact on voting* ***against*** *them.*"
2. **Support the Truth Teller** – If you discover a politician has told the truth, then when you vote or take action you should strongly support the politician or the source of the truth. In this manner we encourage more truth tellers. The training says: "If the claim is *true*, apply the **Reward the Truth Teller** rule and *strongly support* the truth teller. For example, this would have a *Very large impact on voting* ***for*** *them.*"

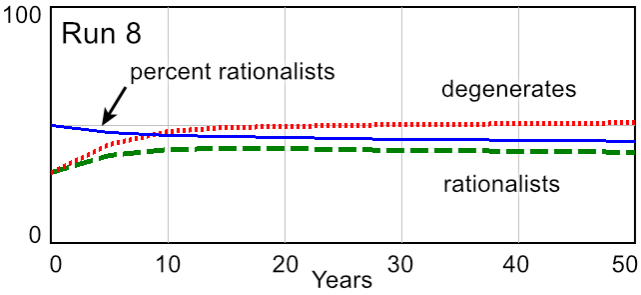
LTQ training in the Truth Literacy Training study consists of learning how to spot 6 fallacies and flawed application of the Strong Evidence Rule. This requires learning dozens of rules and the procedure of how to apply the Personal Truth Test, which is a much larger task than the two rules of AAQ training.

Simulation runs 8, 9, and 10 push on the two high leverage points equally to determine how much LTQ and AAQ must be raised to solve the problem.

The pattern is as LTQ and AAQ rise, more lies are detected and acted upon. This causes optimum *false meme size* to fall. Politicians are adaptive. Here they adapt by

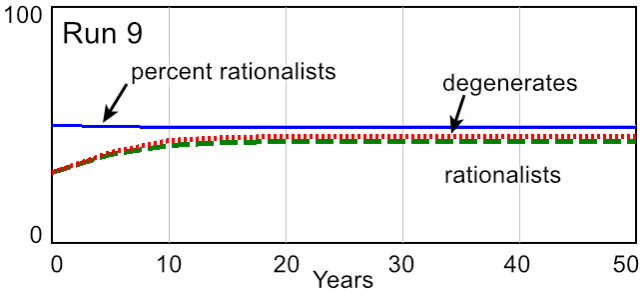
reducing the size of their lies to reduce chance of detection. They are trying to slip smaller lies through people's defenses, and it works. But it works only up to a certain point:

Run 8. As LTQ and AAQ rise from 20% in run 5 to 30% in this run, optimum *false meme size* falls from 4.9 to 2. *Percent rationalists* rises from 34% to 43%.



false meme size = 2, LTQ = 30%, AAQ = 30%

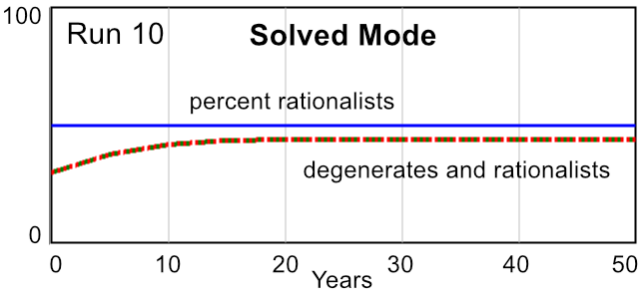
Run 9. As LTQ and AAQ rise still further to 40%, optimum *false meme size* falls to 1.3. *Percent rationalists* rises to 49%.



false meme size = 1.3, LTQ = 40%, AAQ = 40%

Run 10. The solved mode.

Finally, as LTQ and AAQ rise to 50%, optimum *false meme size* falls to 1, its lower limit. *Percent rationalists* rises to 50%, its upper limit. Raising LTQ and AAQ further has no effect, since *false meme size* cannot be less than one.



false meme size = 1, LTQ = 50%, AAQ = 50%

Run 10 represents the solved mode and was a counterintuitive discovery offering deep insight, a frequent occurrence when using system dynamics modeling to reveal a problem's feedback loop structure. We expected that as LTQ and AAQ rose, more degenerates would move to the **Race to the Top** and that loop would become dominant because it contained the most supporters.

But that's not what the model predicts will happen. The model shows the two loops will end their perpetual duel and effectively merge into one loop, because now both loops compete for supporters by telling the truth. The two feedback loops behave as a single **Race to the Top**, because they have the same meme size of one. This is what the social force diagram in the paper means when it says in fundamental solutions, that "This shifts loop dominance from the **Race to the Bottom** to the **Race to the Top**."

Optimum *false meme size* falls from 4.9 in run 5 to 1 in run 10. This corresponds to the right moving from an extreme far-right false ideology to a truth-based moderate position, one so moderate that like the rationalists, they too pursue the common good.

In run 10 the tendency of politicians to tell lies does not just fall to a low level. *It disappears altogether, because now the only winning strategy for politicians is telling the truth*. Those who do not tell the truth die out because degenerates switch from right-

wing partisan extremism (which requires lying) to truth-based moderate behavior. They are now moderates, not degenerates. There will be differences of opinion among political parties and politicians. Different interpretations of what common good goals are paramount and how to achieve them will persist. But if LTQ and AAQ are high enough, citizens will elect politicians who can now work together in harmony. [3]

Movement from partisan extremes to the moderate center eliminates support of authoritarianism. In his examination of the left and right in political systems, Bobbio (1996, pp.xvi–xvii) explains why:

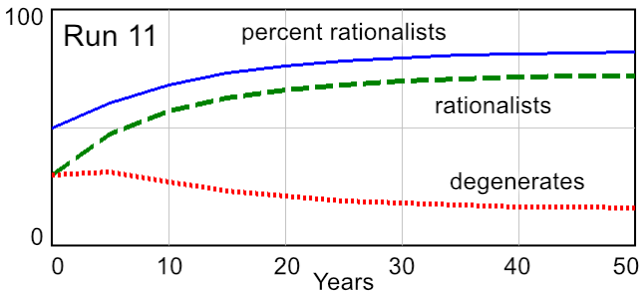
> …. the modern political universe is made up of two entirely separate axes: left/right and liberty /authoritarianism. …the two axes in politics combine to produce four categories: the extreme right, the moderate right, the moderate left, and the extreme left. The extremists are authoritarian, and do not accept the rules of democracy, and although the moderate left and moderate right disagree over the question of equality, they accept the same rules for the political game.

The Dueling Loops model employs a single axis, consisting of Bobbio's "the extreme right, the moderate right, the moderate left, and the extreme left." When supporters move from extreme to moderate positions, "they accept the same rules for the political game," and reject authoritarian pressure to backslide. In Bobbio's words, when the "inclusive middle" dominates, "left and right cease to be two mutually exclusive totalities like two sides of a coin which cannot both be seen at the same time; they become two parts of a whole, a dialectic totality" (Bobbio 1996, p.7).

Runs 11 and 12

Some **Race to the Bottom** factions will adapt to rising DTQ by moderating toward the political center. But we expect others will not. They will continue to promote their deception-based far-right ideology. How would the model behave if instead of moderating, the right choses to continue a strategy of deception? Runs 11 and 12 are used to answer this question.

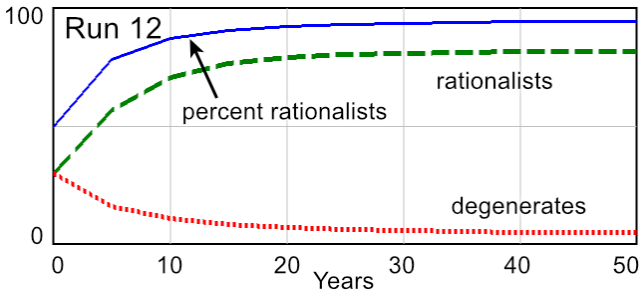
Run 11. This shows what happens when instead of choosing an optimum *false meme size* of one that maximizes the number of their supporters, a group feels so locked into their own false ideology that they do not adapt at all from run 5. Even though LTQ and AAQ have risen to 50%, the degenerates stick with the same level of lies used in run 5. The result is 82% rationalists, which means 18% degenerates. 82% well above 50%, so it's easily high enough for the system to solve common good problems.



false meme size = 4.9, LTQ = 50%, AAQ = 50%

[3] That supporters in the model stay in the Race to the Bottom rather than switching to the Race to the Top agrees with the PEW Research Center's finding that "People rarely switch parties." (Jones et al. 2020)

Run 12. Here LTQ and AAQ are raised still further to 70%. The result is 94% rationalists and 6% degenerates. If LTQ and AAQ were raised still more to 80% (not shown), only 3% degenerates would remain.



false meme size = 4.9, LTQ = 70%, AAQ = 70%

We feel anything over about 80% is not realistic in the immediate future. Even though the study was able to raise LTQ and AAQ to about 80% and 90%, raising political truth literacy to that high a level in a large population will probably take generations. However, we do expect that raising LTQ and AAQ to medium levels of about 50% for swing voters and the young is quite practical in the short term, in less than ten years, if a state is strongly committed to preserving democracy.

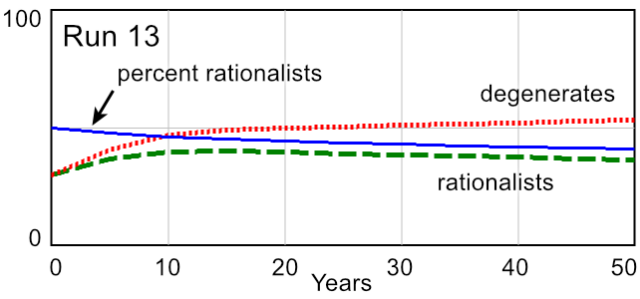
Runs 11 and 12 deal with the problem of far-right minorities who despite their small percentage of the population, have a significant effect on the political system. In a democracy the rights and desires of minorities must be respected and addressed. If 18% of voters (run 11) promote a far-right ideology, a nation will be too distracted to focus efficiently on highly demanding problems. This may be seen in the disproportionate influence far-right groups can have, such as the authoritarian populist wave in Europe of Le Pen in France, the Austrian Freedom Party in Austria, the Sweden Democrats in Sweden (who emerged from violent neo-Nazi groups in the late 1980s), and the mis-named Center Party and Forum for Democracy parties in the Netherlands. In Germany, Denny (2021) found that the far-right Alternative for Germany (AfD), even though national polls show only 10% support:

> …poses a significant and complex threat to the German constitutional order. Highly organized and openly hostile to the rules binding other political actors, the German far right has outperformed its electoral support in shaping German society. In 2020, [one of Germany's intelligence agencies] reported that the number of right-wing extremists in Germany has increased to 33,300, of whom 13,300 are thought to be willing to commit violence.

The vote share for populist parties in 32 European democracies with a lower or single house of parliament rose from 5.3% in the 1950s to 12.4% in the 2010s (Norris and Inglehart 2019, p.9). Historically, the far right is where authoritarian support begins. But it cannot begin if the political truth literacy components are medium or high.

Run 13

The paper states: "The key model insight is that the size (and hence the attractive power) of a lie (*false memes* on the model) can be inflated, while the size of the truth (*true memes*) cannot." However, this inflation only gives the **Race to the Bottom** an advantage because of the unresolved root cause of low political truth literacy. Once the root cause is resolved, inflation to increase *false meme* attractiveness no longer works.



false meme size = 1, LTQ = 50%, AAQ = 50%,
influence per degenerate raised from 200 to 300

A major factor in political systems is the way inflated *false memes* are amplified by social media. For example, Fisher (2022, pp.5, 9) describes how Facebook's recommendation engine amplifies false memes and how large the effect has become. "The more incendiary the post, the more widely the platforms spread it. … Our algorithms exploit the human brain's attraction to divisiveness".

Run 13 models this amplification by increasing *influence per degenerate* from 200 to 300. *Influence per rationalist* remains at 200. This approximates the amplification effect, since it causes the **Race to the Bottom** to inherently have 50% more memes than the **Race to the Top**. That's a huge advantage.

The result is that optimum *false meme size* is still one. *Percent rationalists* falls from 50% in run 10 to 41%. But this doesn't matter, because degenerate politicians are now telling the truth. In the real world, run 13 corresponds to a democracy that has swung to the right due to amplification, but is not dysfunctional. Healthy cooperative political debate among moderates about differences of opinion prevails, rather than dysfunctional polarization. The right are moderates rather than far-right extremists.

Experimentation shows that starting with run 13, if *false meme size* is raised from 1 to 4.9 and LTQ and AAQ are raised to 70%, the result is about the same as run 12.

Far-right behavior

While much further research is required, the analysis and model tell us that by raising LTQ and AAQ from low to medium, the main root cause can be resolved. As that occurs, most on the right will move the center, where healthy political discourse will prevail instead of debilitating polarization. The rest of the right (the far right) will cling to their false beliefs due to the deceptive power of motivated reasoning, a well-established theory explaining how biased decision-making works (Kunda 1990). The theory explains why once a person is fooled into strong false political beliefs, the person becomes highly partisan and their false beliefs are unshakable.

The far-right's beliefs are so immune to truth-based arguments that they mostly cannot be changed. However, raising a nation's LTQ and AAQ reduces the ability of the far-right to recruit new members via deception. In theory, over the long term the far right will gradually diminish to such a low percentage that they will have little political impact. Or as run 10 demonstrates, they will die out altogether. How various right-wing groups can best be moved to the center or eliminated altogether is an important area for further research.

Motivation threat

At one point in Truth Literacy Training development, we found many subjects were not taking the training seriously, resulting in widely varying and mostly low scores. The training was not working.

The problem was solved by discovery of an insightful set of experiments on "dispelling the illusions of invulnerability" to deceptive persuasion (Sagarin et al. 2002). The study found it was not enough to expose subjects to deceptive statements and explanations of why they were deceptive. This failed to work because "our participants' sense of unique invulnerability to deceptive ads left them unmotivated to use defenses against such ads." This illusion of invulnerability caused subjects to believe they were not susceptible to deception, with the result that "they did not resist the ads containing illegitimate authorities [a form of deception] more effectively than did controls." This was corrected by "demonstrating in an undeniable fashion that participants can be fooled

by ads containing counterfeit authorities." This correction is similar to Linden's motivation threat against being fooled.

To dramatically demonstrate to people they are not invulnerable to deception, we changed the initial part of the training. After subjects in groups 2 and 3 answer questions for the first three statements in the pretraining section of the questionnaire, before any training has occurred, they read an educational item on *The concept of truth literacy*. There they are shown their own answers for the first three statements (the first two are almost always wrong) versus the correct answers. The item then says:

> If you got all the answers right, congratulations. However, here's how other people did. In a past survey with 34 participants, none got the answer to the first question right. Three got the answer to the second question right. On the third question 19 people got the answer right.
>
> Why are the first two questions so hard? It's because they use clever forms of deception, which makes it terribly difficult to determine how true the claims are.
>
> The reason so many people got the third question right is it's not deceptive. Generally, it's much easier to spot the truth as opposed to deception, because we are so used to processing true statements from people we talk to, books we read, and so on.

This shocks people into realizing they are vulnerable to deception. From this point on, almost all take the questionnaire seriously. From the viewpoint of the elaboration likelihood model of persuasion (Dillard and Shen 2013, pp.137–149), elaboration *motivation* has increased from low to high. Subsequent training increases their elaboration *ability*, with the result that when training is complete, most deceptive persuasion attempts will be processed (elaborated) correctly and they will not be fooled.

Raising political truth literacy is a preventive solution. Linden (2022) theorizes that:

> The most common framework for preventing unwanted persuasion is psychological inoculation theory. The theory of psychological inoculation follows the biomedical analogy and posits that, just as vaccines trigger the production of antibodies to help confer immunity against future infection, the same can be achieved with information. By pre-emptively forewarning and exposing people to severely weakened doses of misinformation (coupled with strong refutations), people can cultivate cognitive resistance against future misinformation.

"Pre-emptively forewarning" refers to "demonstrating in an undeniable fashion that participants can be fooled by ads containing counterfeit authorities." From Linden's observation on inoculation theory and our solution to the "not taking the training seriously" problem we reached an important conclusion concerning motivation:

> Preventive solutions, such as Truth Literacy Training, must first raise a person's *motivation threat level* before they will take preventive training seriously and then use that training to protect themselves from deception attacks.

The motivational threat level is part of the LTQ and AAQ variables:

$$\text{LTQ} = \text{LTQ skill} \times \text{LTQ motivation to use the skill}$$

$$\text{AAQ} = \text{AAQ skill} \times \text{AAQ motivation to use the skill}$$

The study measured only LTQ and DTQ, so we do not know what the approximate skill and motivation variable values are. For this reason we did not model them, since that would involve too much speculation with estimates. What their values are, what their causal structure is, and how they respond to solution elements are interesting topics for further research.

Why didn't backsliding start long ago?

The question arises: If political truth literacy (DTQ) has always been low, why didn't backsliding begin long ago? This could have been modeled, but as stated in the paper:

> We wish to make only a small solid advance, which necessarily limits research scope. We do not seek to explain the timing or causes of the three waves of autocratization, nor the ascent of democracy. We address only the case where free and fair elections still prevail.

Still, this is an important question. Failure to be able to offer a strong explanation would invalidate the analysis. Examination shows there are several causes of why backsliding didn't start long ago:

(1) *The shift from illegal to legal takeovers*. At first, backsliding was due to illegal takeovers. In an illegal takeover, such as military coups and invasions, DTQ matters little compared to a dictator's goals and military strength. But with the rise of the popularity of democracy, anti-pluralists switched to legal takeovers via elections characterized by deceptive "anti-pluralist rhetoric" (Luhrmann et al. 2023). This began in the 1990s, and since 2000 has accounted for about 80% of takeovers (Svolik 2019).

(2) *Democracy delivered economically for a long time but then failed to deliver* "the promised broad-based growth" beginning in the 1970s and accelerating in the 1990s. The trend greatly accelerated in the 1990s and "dramatically intensified" during and after the Great Recession of 2008 (McCarthy 2019). Instead of broad-based growth, "the top 1% took 38% of all additional wealth accumulated since the mid-1990s, whereas the bottom 50% captured just 2% of it" (WIL Staff 2022).

(3) *Displacement of traditional gateway news organizations by internet blogs and social media*, which are far less reliable. With the rise of social media in the late 2000s, "conspiracy stories and fabricated information are often rewarded with more clicks than the truth receives. Among the hierarchies disrupted by the internet was the one formed by the 'legacy media' in democratic countries—media organizations working in print, radio, and television that had over time developed journalistic standards for vetting and verifying information" (Fukuyama 2020).

(4) *Injection of misinformation by international actors*, particularly Russia, into traditional news and especially social media. This began in the 2000s with Putin's perfection of "modern authoritarianism," now widely copied by China and others (Puddington 2017). Russia Today TV, a combination of legitimate and propaganda news in four languages, launched in 2005 and reaches about 700 million households in over 100 countries (RT Staff 2023). Russia is globally adept at social media digital propaganda (Sanovich 2017).

Conclusions

Referring to social force diagram in the paper, the purpose of the simulation model was to find the deeper cause of intermediate cause 4, *successful political deception*. The deeper cause was found to be *low political truth literacy*, which is the main root cause. *Political truth literacy* (DTQ) was found to consist of two components, LTQ and AAQ. Raising them from low to medium resolves the root cause, so they are the high leverage points.

References

Bobbio, N. 1996. *Left and right: the significance of a political distinction*. The University of Chicago Press.

Coyle, G. 2000. Qualitative and quantitative modelling in system dynamics: some research questions. *System Dynamics Review* 16(3), pp. 225–244. doi: 10.1002/1099-1727(200023)16:3<225::AID-SDR195>3.0.CO;2-D.

Dawkins, R. 1976. *The Selfish Gene*. Oxford University Press.

Denney, S. 2021. The German Far Right Doesn't Need to Win Elections to Be Dangerous. *Lawfare*. Available at: https://www.lawfareblog.com/german-far-right-doesnt-need-win-elections-be-dangerous.

Dillard, J. and Shen, L. 2013. *The SAGE Handbook of Persuasion: Developments in Theory and Practice*. SAGE.

Fisher, M. 2022. *The Chaos Machine: The Inside Story of How Social Media Rewired Our Minds and Our World*. Little, Brown and Company.

Forrester, J. 1971. *World Dynamics*. Wright-Allen Press.

Fukuyama, F. 2020. 30 Years of World Politics: What Has Changed? *Journal of Democracy* 31(1), pp. 11–21. doi: 10.1353/jod.2020.0001.

Ghaffarzadegan, N., Lyneis, J. and Richardson, G.P. 2011. How small system dynamics models can help the public policy process. *System Dynamics Review*.

Jones, B., Doherty, C., Daniller, A., Dunn, A. and Green, T. 2020. *Voters Rarely Switch Parties, but Recent Shifts Further Educational and Racial Divergence*. PEW Research Center. Available at: https://www.pewresearch.org/politics/2020/08/04/voters-rarely-switch-parties-but-recent-shifts-further-educational-racial-divergence/.

Kunda, Z. 1990. The Case for Motivated Reasoning. *Psychological Bulletin* 108, pp. 480–498.

Linden, S. 2022. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*.

Luhrmann, A., Medzihorsky, J. and Lindberg, S.I. 2023. Walking the Talk: How to Identify Anti-Pluralist Parties. *Party Politics*. Available at: https://journals.sagepub.com/doi/10.1177/13540688231153092.

McCarthy, J. 2019. Authoritarianism, Populism, and the Environment: Comparative Experiences, Insights, and Perspectives. *Annals of the American Association of Geographers* 109(2), pp. 301–313. doi: 10.1080/24694452.2018.1554393.

Noel, A. and Therueb, J.-P. 2008. *Left and Right in Global Politics*. Cambridge University Press.

Norris, P. and Inglehart, R. 2019. *Cultural Backlash: Trump, Brexit, and the Rise of Authoritarian Populism*. Cambridge University Press.

Puddington, A. 2017. *Breaking Down Democracy: Goals, Strategies, and Methods of Modern Authoritarians*. FreedomHouse.org. p. 64. Available at: https://freedomhouse.org/sites/default/files/2020-02/June2017_FH_Report_Breaking_Down_Democracy.pdf.

RT Staff. 2023. *Distribution of Russia Today*. Available at: https://www.rt.com/about-us/distribution/ [Accessed: 21 January 2025].

Sagarin, B.J., Cialdini, R.B., Rice, W.E. and Serna, S.B. 2002. Dispelling the illusion of invulnerability: The motivations and mechanisms of resistance to persuasion. *Journal of Personality and Social Psychology* 83(3), pp. 526–541. doi: 10.1037/0022-3514.83.3.526.

Sanovich, S. 2017. *Computational Propaganda in Russia: The Origins of Digital Misinformation*. Programme on Democracy and Technology. Available at: https://demtech.oii.ox.ac.uk/research/posts/computational-propaganda-in-russia-the-origins-of-digital-misinformation/.

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

Svolik, M.W. 2019. Polarization versus Democracy. *Journal of Democracy* 30(3), pp. 20–32. doi: 10.1353/jod.2019.0039.

Sze To, G.N. and Chao, C.Y.H. 2010. Review and comparison between the Wells–Riley and dose-response approaches to risk assessment of infectious respiratory diseases. *Indoor Air*.

WIL Staff. 2022. *World Inequality Report 2022*. World Inequality Lab. Available at: https://wir2022.wid.world/executive-summary/.