Appendix C

Full Model Description (ODD)

The model description follows the ODD (Overview, Design concepts, Details) protocol for describing agent-based models (Grimm et al., 2020). The code, along with a video visualizing the simulation's dynamics, can also be found at: `https://github.com/AmalSalman/Theory-Change-in-Science-ABM`. The model was implemented in the GAMA agent-based modeling software, version 1.9.3 (GAMA-Platform, 2024; Taillandier et al., 2019). Deeper motivations and rationale behind the design decisions is covered in the model design section of the main text.

C.0.1 Purpose and Patterns

The ultimate purpose of this model is to demonstrate that BEN is translatable to the social epistemology of science by investigating whether an ABM utilizing the BEN cognitive architecture can plausibly represent the theory change process in science. To address this, we focus on the following question: How do the factors affecting (1) the level of interaction among scientists, (2) the level of peer influence, and (3) the difficulty of the anomaly contradicting the existing theory, influence the likelihood and speed with which a scientific community collectively converges on the correct consensus, namely, abandoning their current theory in favor of a new one? Therefore, we consider how factors describing the behavior of scientists (the level of interaction and peer influence) and factors describing the conditions of inquiry (the difficulty of the anomaly) impact consensus formation.

There are multiple ways to categorize a model's purpose. This model aligns with (Gelfert, 2020)'s definition of a proof-of-concept model, which "may establish, by way of its own example, that a certain type of approach or methodology is able to generate potential representations of a target phenomenon". This model also aligns with (Edmonds, 2019)'s definition of a descriptive model, which attempts "to partially represent what is important of a specific observed case (or small set of closely related cases)." This is because this model only considers cases of theory

change in science where the anomaly necessitates a complete replacement of the current, not cases where the theory can simply be modified or expanded to explain the anomaly. It also only considers cases where there is only one anomaly, or a set of closely related anomalies that can be interpreted as one, but not multiple anomalies at the same time.

The foundation of this model is designed by drawing analogies from a fire evacuation model that utilized the BEN cognitive architecture. However, that does not make this model an "analogous" model by (**?**)'s definition, because the analogy is not made here to illustrate or simplify the theory change process, but to close the gap between the typical social processes BEN was developed to model (e.g., evacuations and crowd behavior) and the social processes in science, which are fundamentally different and possibly more complex. This model's design aligns with the EROS principle (Jager, 2017)—which emphasizes the integration of psychological theories in defining social agents to enhance realism and explainability—, as was the design of BEN (see section **??** of the literature review).

To consider our model realistic enough for its purposes, we identified four general patterns that describe the expected collective behavior emerging from the simulations. If these patterns emerge, it would demonstrate that the model is capable of providing plausible representations of the theory change process, and therefore directly address our broader research question. These patterns can be considered baseline checks for the model and are formalized into concrete hypotheses to be tested in section **??**.

Due to the scarcity of empirical data on this topic, we had to rely on stylized facts to derive these patterns. Stylized facts "are broad, but not necessarily universal generalizations of empirical observations and describe the supposed essential characteristics of a phenomenon that requires an explanation" (Meyer, 2019). They are primarily used in economics and the social sciences, but have increasingly been used to validate ABMs and to design basic agent characteristics (Grimm and Railsback, Tue, 03/26/2019 - 12:00; Meyer, 2019). These stylized facts are derived from the literature discussed in section **??**.

Pattern 1: When a scientific field encounters a significant anomaly that contradicts the existing dominant theory, there are only three possible conclusive states of the system (i.e., when the community has reached a consensus): efficiently converging to the correct consensus, inefficiently converging to the correct consensus, or failing to converge on the correct consensus. This depends on how the investigation into the anomaly proceeds. The process of reaching consensus could be efficient, indicating a quick and consistent investigation, or inefficient, indicating that the investigation was consistent but slow or that it died out and started again after some time, or it can be false, indicating that the investigation was minor or too slow to be considered a significant attempt at theory change. This also implies that while fluctuations and oscillations in the rate of scientists adopting the new theory may occur, they are likely limited in number and magnitude, with only one major peak (tipping point) representing a significant shift in the community's consensus.

Pattern 2: There is always an initial state of resistance before any scientist decides to adopt the new theory. In the best cases, this is simply the investigation period which must always take place. In the worst cases, this signifies irrational collective behavior as discussed in section **??**. In other words, evidence about the anomaly needs to accumulate before scientists start to adopt the new theory, but their (collective) level of resistance can slow or hinder the accumulation of the evidence.

Pattern 3: A scientific community is more likely to converge on the correct consensus than not. The factors influencing this convergence (the level of interaction among scientists, level of peer influence, and difficulty of the anomaly), will likely not cause the community to fail to converge on the correct consensus, but will affect the speed at which they do.

Pattern 4: All the factors influencing the scientific community's convergence to the correct consensus (the level of interaction among scientists, level of peer influence, and difficulty of the anomaly) should play a role in the speed of that convergence. Characteristics of the anomaly (its difficulty) are not the only determinant of the likelihood or the speed of convergence to the correct consensus.

In addition to these general patterns, we hypothesize the following specific relationships between the modeled factors and the speed of convergence to the correct consensus:

Hypothesis 1: As the level of interaction among scientists increases, the speed of convergence to the correct consensus also increases. This is because increased interaction facilitates the dissemination and discussion of evidence, accelerating the evaluation and adoption of the new theory.

Hypothesis 2: As the level of peer influence increases, the speed of convergence to the correct consensus initially increases, but then decreases if the influence becomes too high. Moderate peer influence can promote the spread of new ideas, but excessive influence can lead to conformity pressures and hinder critical evaluation.

Hypothesis 3: As the difficulty of the anomaly increases, the speed of convergence to the correct consensus decreases. More challenging anomalies require more extensive investigation, which naturally slows down the process of reaching a consensus as more effort and time are needed to understand and validate the new theory.

These hypotheses will be tested by systematically varying the parameter value associated with each factor and analyzing the resulting model outcomes. The confirmation of these patterns and hypotheses will provide strong evidence for the translatability of BEN to the social epistemology of science.

C.0.2 Entities, state variables, and scales

The model has the following kinds of entities: agents representing scientists, a single agent representing a fire, grid cells, and the global environment. The agents use the BEN (Behavior with Emotions and Norms) cognitive architecture for decision-making (Bourgais et al., 2020). The space is defined by a 2-dimensional grid of (100 × 105) cells and the location of agents is defined by the spatial unit (grid cell) it is in. The space in this model is abstract and does not correspond to a real system, but is used to represent the dominant theory within a scientific field being modeled (e.g. Classical mechanics in Physics). Some grid cells are defined as walls which separate different rooms within this space, where rooms are used to represent different subfields or methodological approaches within the field modeled. The distance between cells, therefore, represents the metaphorical distance or division between subfields. Some grid cells define an exit door, where agents go to "evacuate" the space representing that they have abandoned the current dominant theory and adopted a new one, and the bottom (100 × 5) cells represent an evacuation space, where agents go after exiting the main space, i.e., the top (100 × 100) cells.

The fire agent is used to represent the anomaly and its only function is to indicate the location from which the evidence supporting the anomaly will start to spread, represented as smoke. The grid cells which do not represent a wall, an exit, nor an evacuation area, are considered free cells, where agents can roam and smoke can spread. Each free all holds information about the level of concentration of their smoke at that cell, which represents the strength of the published evidence supporting the anomaly. Figure C.1 shows a visualization of this space.

One time step in the model is approximated to represent half a year. This is a rough approximation based on the fact that in every time step, agents can decide to investigate the anomaly or abandon the current theory, and can go back and forth between these decisions, therefore, it makes sense that such decisions can only occur after some time has passed for deliberation. Furthermore, in every time step, agents can publish the results of their investigation into the anomaly or the new theory. While some research shows that scientists can publish up to a paper every five days (Ioannidis et al., 2018), our model only considers publications that result from an agent's own research, which further supports the need for the time step to not be too long nor too short. Half a year, or an academic semester, is therefore considered a sufficient approximation for one time step.

The state variables characterizing these four entities are listed in table C.1. How these entities and their state variables are initialized and the processes that causes them to change are described in sections **??**.

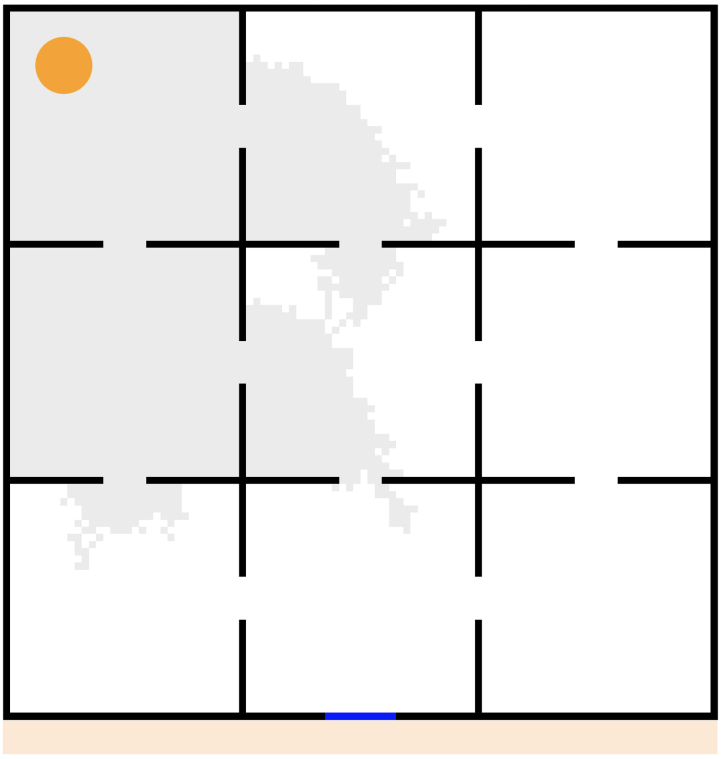

FIGURE C.1: Visualization of the model environment

TABLE C.1: Summary of model state variables.

Entity	**Variable name**	**Description**	**Possible values**	**Type / Units**
Grid cells	`grid_value`	Indicates which grid cell represents a wall, an exit, or an evacuation area	1, 2, or 3 (respectively representing wall, exit, or evacuation area cell)	Integer

Continued on next page

Table C.1 – continued from previous page

Entity	Variable name	Description	Possible values	Type / Units
	`smoke_level`	Indicates the level of concentration of smoke in a cell, which represents the strength of evidence supporting anomaly.	[0-1]. 0 signifies no evidence, 1 signifies that every conceivable control experiment, credibility, and reproduciblity test has been made and every conceivable alternative explanation for the anomaly has been considered. Thus, representing that anomaly is unexplainable by current theory.	Float
Global environment	`theory_level`	Indicates how well-developed the new alternative theory is. Global variable, thus, accessible by any agent at any time.	[0 - 1]. 0 signifies no development of a new theory. 1 signifies that the new theory is sufficiently well-developed for any agent to join.	Float
Fire	`center_location`	The grid cell where the center of the fire is and consequently the grid cell from where `smoke_level` propagates.	A grid cell index	Tuple of integers
Scientist Agent	`location`	The grid cell the agent occupies	A grid cell index	Tuple of integers
	`is_proposer`	Indicates where the agent is one who initially discovered the anomaly and/or proposed a new theory in response	True/False	Boolean

Continued on next page

Table C.1 – continued from previous page

Entity	Variable name	Description	Possible values	Type / Units
	`expert_level`	Indicates agent's level of expertise relative to the expertise required for anomaly investigation	1- 5	Integer
	`openness`	One of the five personality traits.	[0-1]. 0 indicates narrow-minded agent, 1 indicates open-minded agent.	Float
	`conscientiousness`	One of the five personality traits.	[0-1]. 0 indicates impulsive agent, 1 indicates cautious agent.	Float
	`extraversion`	One of the five personality traits.	[0-1]. 0 indicates introverted agent, 1 indicates extroverted agent.	Float
	`agreeableness`	One of the five personality traits.	[0-1]. 0 indicates cooperative agent, 1 indicates hostile agent.	Float
	`neuroticism`	One of the five personality traits.	[0-1]. 0 indicates neurotic agent, 1 indicates calm agent.	Float
	`target`	The grid cell the agent is heading to	A grid cell index	Tuple of integers

Continued on next page

Table C.1 – continued from previous page

Entity	Variable name	Description	Possible values	Type / Units
	`current_beliefs`	Indicates which beliefs the agent currently has.	One or more of: `smoke_Serious`: when agent believes that smoke (anomaly) is serious, i.e., unexplainable by current theory. `must_investigate`: when agent believes the anomaly is serious enough to warrant further investigation.	Belief Predicate
	`current_desires`	Indicates which desires the agent currently has.	One or more of: `to_stay`: agent has desire to stay, i.e., to keep working with current theory. `to_exit`: agent has desire to exit, i.e., abandon current theory and adopt a new theory. `to_investigate`: agent has desire to investigate the anomaly. `to_influence_others`: agent has a desire to intentionally influence others' beliefs. `to_publish`: agent has a desire to publish results of their research.	Desire Predicate

Continued on next page

Table C.1 – continued from previous page

Entity	Variable name	Description	Possible values	Type / Units
	`current_plan`	Indicates which action plan or social norm the agent is currently employing	One of: Plan: `Staying` Plan: `Exiting` Plan: `Investigating` Plan: `Advocating_to_Exit` Plan: `Advocating_to_Stay` Norm: `Follow_Others`	Action Plan
	`disengagement_level`	Indicates how disengaged an agent is from the discourse surrounding the anomaly and the new theory. The more an agent comes in contact with others who are disengaged (not investigating nor have an opinion on whether anomaly is serious or not), the more disengaged they become.	[0-1], 0 indicates no disengagement and 1 indicates maximum disengagement.	Float

C.0.3 Process overview and scheduling

Every time step (time step), the following processes are executed in the following order, after initialization. The submodels implementing these processes are described in detail in Section C.0.7 The initialization is described in detail in Section C.0.5 and includes how the `grid_value` of grid cells is defined and how agents are assigned their `expert_level` and five personality traits.

1. The grid cells execute their "update `smoke_level`" submodel, which adds the effects of the agents' publications from the previous time step to `smoke_level` and defines how the smoke spreads to neighboring cells.

2. The global environment executes its "update `theory_level`" submodel, which adds the effects of the agents' publications from the previous time step to `theory_level`.

3. The agents each execute the following sub models in the following order, and they do this in parallel, i.e., all agents execute their submodels at the same time:

   (a) The "perception & interpretation" submodel defines how an agent perceives surrounding cells (the `smoke_level` they hold) and surrounding agents. Characteristics of the perceived agents are used to update the agent's `disengagement_level`. Agents here access the global variable `theory_level` as well. `disengagement_level` is used to possibly distort the agent's perception of `smoke_level` and `theory_level` which represents the agent's interpretation of the evidence. The more disengaged an agent is, the more they interpret `smoke_level` and `theory_level`, to be less than they actually are.

   (b) The "update beliefs and desires" submodel uses agent's information about the world and the agent's personality traits to update the agent's `current_beliefs` and `current_desires`. This is done through a set of defined inference rules.

   (c) The "action plan selection" submodel defines the agent's `current_plan` based on their current desires and context. Action plans define an agent's behavior. Specifically, they define an agent's new target if they have reached their last assigned target, and they define if the agent will publish.

   (d) The "moving" submodel defines how the agent will move to its new target grid cell or if it will not move. The target cell is defined by the action plan selected in the previous step. The sub model executes the movement action such that the agent walks towards their target. Depending on speed and other variables controlled by parameters, the agent might reach its target in one or more time steps.

   (e) The "publishing" submodel defines when and how the agent will publish depending on the action plan selected. Under certain conditions the agent can publish, and then the impact of their publication (i.e., how much it can increase `smoke_level` or `theory_level`) is calculated based on parameters. The impact of publications of all agents in one time step is aggregated into a variable for `smoke_level` or `theory_level` respectively that is send to the "update `smoke_level`" and "update `theory_level`" submodels.

C.0.4 Design Concepts

C.0.4.1 Basic Principles

At the system level, the model aims to reproduce the basic characteristics of the theory change process in science as emergent properties of the interactions between highly cognitive agents. While agent-based models (ABMs) have been used extensively in the social epistemology of science, none have explored the mechanisms underlying theory change in science in the specific case we focus on in this model, where the current, dominant theory in a scientific field is challenged by an unexplainable anomaly and needs to be replaced by a new theory.

ABMs in the social epistemology of science typically model idealistic cases where there are many alternative theories and/or methodologies for scientists to choose from as they explore their field and look at questions like how many of the agents need to be mavericks (agents who like to investigate the unexplored) for the community to reach a certain truth. In most scientific explorations, this is indeed true, and scientists do not need to choose between two alternatives but can modify and expand their working theory as they investigate and uncover new things. However, the cases this model focuses on, where the dominant theory needs to be completely replaced, while less frequent, are crucial as they change the course of scientific discourse. They are interesting to study because in most, if not all, of these cases, there is always a period where scientists resist this change. Sometimes this resistance is good, as scientists spend time making sure that the anomaly is indeed unexplainable and find what kind of new theory is best to pursue to explain all the evidence. Other times, however, this resistance takes much longer than needed and becomes irrational. Examples of such cases are reviewed in the literature review of the main thesis text.

It is important to understand why this irrational resistance occurs so that it can be prevented, but it is not very clear yet why this happens. The system is very complex, and many factors are involved, including not only details about the specific anomaly and the methods scientists use to address it but also social dynamics, power structures, and external influences like funding availability. Recently, this behavior has been explained through the psychological concept of groupthink (**?**). Groupthink is a psychological phenomenon where groups reach an irrational consensus because they are driven by a psychological need for social harmony. Previously, this problem has been famously explained through the concept of paradigm shifts, where periods of "normal science" (working within an established paradigm) are disrupted due to anomalies by revolutionary "paradigm shifts" that make fundamental changes to the theories or "paradigms" scientists work with (see more in the literature review of the main thesis text). This has been modeled several times but only once with an ABM (De Langhe, 2018), and the purpose of that ABM was to evaluate and better understand Kuhn's theory and so significantly differs from our purpose. Furthermore, no ABMs in the social epistemology of science have employed cognitive architectures.

This all poses several challenges in designing this model, as it differs significantly from previous models, and there is little quantitative research to base the design on. To address this problem, we aimed to make the model as simple as possible, use logical processes that make intuitive sense, and limit our assumptions to the agent rather than the system level (so little to no top-down control). Moreover, since this model attempts to be analogous to the fire evacuation model used to evaluate BEN, we used that as the starting point for assumptions at the systems level. The parallels drawn between the fire evacuation model and this model, as well as the conceptual decision-making model for individual agents, are described in detail in the model design section of the main thesis text.

C.0.4.2 Emergence

The model's primary results are how long it takes for the majority of agents to exit (if they do), and how many agents exit in total. These two results emerge from the interactions between agents and between the agents and the environment. While these results are significantly affected by parameter settings, they are not imposed by any top-level rules or mechanisms. The parameter settings were chosen in such a way that they make sense conceptually (see more in section C.0.7) and there was no calibration (inverse determination of parameters) as we view that as imposing too much control over the model. The model is considerably abstract and attempts to capture the dynamics of a highly complex system with relatively simple processes.

Of the four patterns used to evaluate the model (see section C.0.1), only patterns 2 and 4 can be said to be somewhat imposed by design decisions. For pattern 2, this is more clear as we have defined the agent model in such a way that it is very unlikely, but still possible, that an agent can immediately decide to exit; they have to go through an investigative step before that happens (see details in section C.0.7). Therefore, it is also unlikely that as a collective, the agents can immediately start exiting.

For pattern 4, this is slightly more complex because, in a way, we have ensured that social factors play a role, for example, by introducing `disengagement_level`, and the impact of such factors is to some extent defined by parameter settings. However, all design decisions were made at the individual agent level to be conceptually sound and not to impose any collective behavior, and there is sufficient randomness in the model for the collective behavior to be unpredictable. Patterns 1 and 3 are high-level collective behaviors that emerge and are not as straightforward as 2 and 4.

Finally, the submodel "update `smoke_level`" which defines how the smoke spreads across the grid is dependent on parameter settings, and the smoke spreading mechanism cannot be said to be an emergent property of the model. We considered making this mechanism more fine grained such that smoke spread is also an emergent property, but it would have significantly increased the complexity of the model and was determined to be unnecessary.

C.0.4.3 Adaptation

As defined by the BEN cognitive architecture, at the end of every time step, each agent decides if they will change or keep their current plan or social norm. Agent behavior is modeled via indirect objective-seeking, meaning that agents follow predetermined rules to make their decisions and do not for example have choose between alternative plans according to some thresholds or probabilities (which is also possible to do in BEN). This was done to keep the model simple. The process that defines how agents make this decision is the "action plan selection" submodel, which is heavily influenced by the "update beliefs and desires" and "perception and interpretation" submodels, and is described in detail in section C.0.7.

C.0.4.4 Objectives

There is no direct objective-seeking in this model. The indirect objective-seeking in this model is that agents selecting an action plan every simulation time step based on their desires. The way agents select between alternative action plans is outlined in the flowchart in figure **??** and the conceptual model it is based on is outlined in the flowchart in figure **??**. This process is described in detail in section **??**.

C.0.4.5 Learning

There is no learning in this model. This is partly the reason why we are mainly interested in results up until the majority of agents exit. We expect that after a consensus starts to form (when the majority of agents have exited), the agents still deciding would have a different decision-making process that involves learning from their past behavior and the behavior of other agents. This does not have a direct impact on our main research question and so was not modeled for simplicity.

C.0.4.6 Prediction

The aim in designing this model was to be as conceptually realistic as possible while keeping the model simple (see more in section C.0.4.1). Therefore, while we cannot say that the model was designed to represent how the agents actually make predictions or decisions, because there is a lack of empirical/quantitative studies to rely on about how individual scientists behave in these situations, we can say that the model was designed to represent a simple, logical, and realistic process for how the agents make predictions or decisions. The agents do not make any assumptions or predictions about the possible consequences of their actions, as they only do what they think is right, i.e., what they think would lead to the truth, with possible (unconscious) biases, with the exception of agents who have a desire to intentionally influence

others (defined in their initialization), in which case the implicit assumption is that more agents will be influenced to copy their beliefs (this is described in detail in the "action plan selection" submodel in section C.0.7).

C.0.4.7 Sensing

Sensing is called perception in the BEN architecture, and it is a skill inherent to all agents. For agents to perceive their surroundings, a geometrical shape has to be defined which constrains their perception area. This is typically a cone, as shown in figure C.2. The agent's heading (where it is pointing to) is defined by its current location and the shortest path to its target location. The `vision_amplitude` parameter defines the angle for how wide the cone is, at maximum, this angle is 180° and the cone becomes a circle and the agent can see all around them. The other important parameter defining the cone is `perceive_distance` which define the radius of the cone. In this model, `perceive_distance` depends on the agent's `expert_level` such that agents with higher `expert_level` are able to perceive things from further away. This cone is then called the agent's `perception_area` because the agent can perceive anything inside that cone. If the agent perceives another agent, they can access any of their state variables such as their `expert_level` as well as their beliefs (without permission from the perceived agents). If the agent perceive a grid cell, they can also access any of its state variables such as `smoke_level`. Note that while agents perception is modeled as 100% accurate (without noise), we added an extra step in this process where agents interpret the evidence they perceive. This is to add a level of realism to this process, which is described in detail in section C.0.7.3.

We considered having two perception cones, one for perceiving other agents and another for perceiving smoke, to allow for different levels of perceiving ability between the two, for example, an agent might be able to perceive other agents close to them but their `expert_level` is too low for them to be able to perceive any smoke. However, we found this unnecessary as it adds complexity to the model while not being directly relevant to our research question, but it could be useful in future modification of the model.

C.0.4.8 Interaction

The BEN architecture provides the ability to use a social module which allows agents to have social links between one another, GAMA also allows for agents to communicate with each other in different ways, such as sending messages to one another. In this model, we decided it was unnecessary to use either of those features, with the goal to keep the model as simple as possible. Therefore, the interaction between agents in this model is limited to their sensing or perception ability described in the previous subsection. We define a simple process to model the effect of social influence on agents, which, using the state variable `disengagement_level`, represents

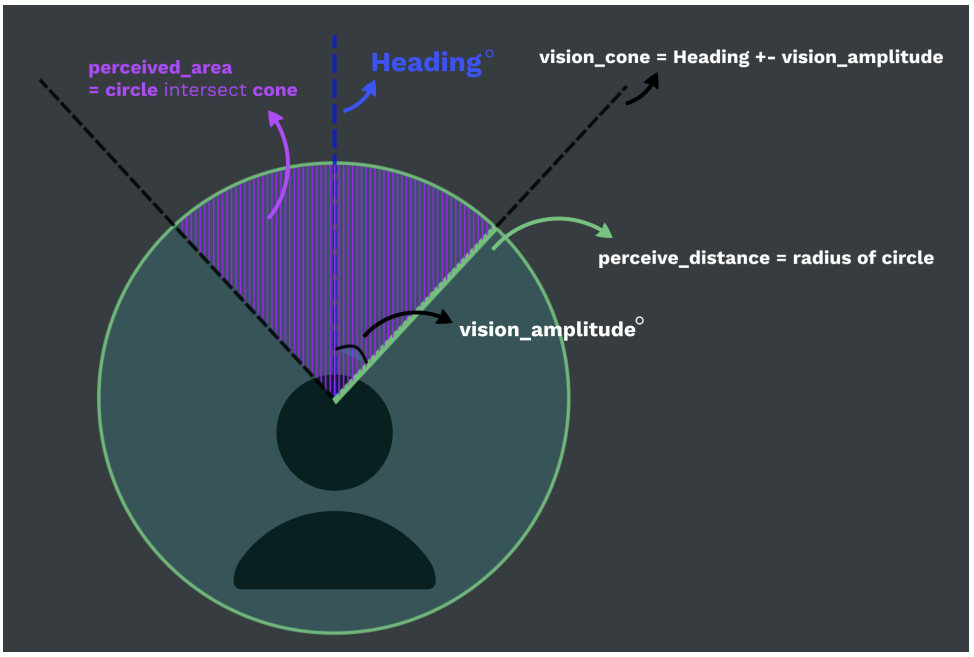

FIGURE C.2: Description of an agent's perception cone.

how agents can get more disengaged the more there are agents around them who are disengaged and how this affects how they interpret the evidence they perceive. This process is described in detail in section C.0.7.3.

The agent's action plans define their target, which they move towards every time step, and this has the biggest effect on which agents end up interacting with. This is described in detail in section **??** The interactions in this model are mediated rather than direct interactions, meaning that agents only affect each other indirectly. Agents affect each other by 1) affecting their `disengagement_level`, 2) publishing.

C.0.4.9 Stochasticity

Most of the stochasticity in this model occurs in the initialization of the model. The agents are given an initial random location. With a big enough number of agents, they are sufficiently uniformly distributed across the grid. The agent's `expert_level` is assigned in the initialization and depends on the agent's location such that the agents closer to the fire (anomaly's initial location) have a higher `expert_level`. Since we divide the grid into six rooms in the initialization, the agents' `expert_level` are approximately normally distributed. The five personality traits for each agent are approximated with a normal distribution with the mean and standard deviation taken from empirical data. In the initialization, agents are given a desire to intentionally influence others based on their extroversion personality trait, such that the more extroverted an agent is, the more likely they are to want to influence others. This is all explained in more detail in the initialization process in section C.0.5.

Furthermore, the "publishing" submodel has some stochasticity. This process is divided into two main steps; the first defines when an agent gets a desire to publish, and the second defines if their publication gets accepted by a publication journal. The agent's chances for getting a desire to publish increase the less impulsive they are (the `conscientiousness` personality trait) and the more time they spend investigating. The probability of their publication getting accepted after they have a desire to publish is defined by a parameter. This submodel is defined in more detail in section **??**.

The reason stochasticity is used in the initialization process is to introduce variability between the agents with specified distributions that to some extent exist in real life. On the other hand, stochasticity was used in the publication process to make the process variable without having to model the causes of variability, which would have added unnecessary complexity to the model.

C.0.4.10 Collectives

The collectives in this model correspond to the action plans that agents follow, and therefore are emergent properties of the agents. Initially, all agents are using the Staying action plan and thus form one collective, except one agent who is the "proposer" and is using the Investigating plan. As time passes, agents start selecting different plans and form five more collectives, as the collective of agents staying grows smaller: a collective of agents exiting or have exited, ones advocating for other agents to exit, advocating for them to stay, investigating, or using the social norm to follow others. These collectives differ from each other not only because the behavior of agents is different in each one, but because they directly represent the main different ways agents can respond to the anomaly. In other words, knowing an agent's action plan or the corresponding collective they are in means knowing where they stand with regards to the anomaly. This is described in detail in section **??**.

C.0.4.11 Observation

There are three main model output types. In the first, the model can produce a graphical interface, where each agent is represented by a triangle and their color represents the action plan they're following. Each grid `smoke_level` is represented by increasing shades of gray. See figure C.1 for an example of what this looks like. The graphical interface updates every time step to show how agents behave and how the smoke increases. In the second, it is possible to track variable changes per time step in the model. Of particular interest to our analysis (in order to evaluate the expected patterns outlined in section C.0.1) are the number of agents investigating as well as the number of agents exiting or having exited per time step. In the third , it is possible to track outputs at the simulation level rather than per time step. Of particular interest to our analysis is the time step at which the first agent started investigating,

the time step at which the first agent exited, the time step at which the simulation stopped (either because the majority have exited or because the simulation reached the defined maximum number of time steps without the majority exiting), and the total number of people who exited. More observations and measurements might be useful for some analyses, here we focus only on the ones which help us evaluate the patterns described in section C.0.1.

C.0.5 Initialization

Environment initialization: The model environment is initialized with a 100×100 grid cells. To initialize the `grid_value` state variable per cell, a CSV file is imported, which contains a singular integer value per cell. After importing the CSV file, we map each integer value to its assigned meaning in the model. Specifically, 1 indicates a wall cell, 2 indicates an exit door cell, and 3 indicates an evacuation area cell. Every other cell is then considered a free cell, where agents can roam and the smoke can spread. Furthermore, to aid in specifying where one room ends and another begins, we make use of the same CSV file to map a unique integer value to each of the free cells in unique rooms, such that they can be more easily identified in the model. The map is divided evenly into six rooms with wall thickness equal to one cell. There are doors (free cells) between every room and the next of length of six cells. The exit door is located at the center of the bottom wall and is of length 10 cells. The location of the exit and the layout of the rooms can be adjusted in the initialization if needed for analysis. Furthermore, a single fire entity is created, with its center location determined by the parameter `center_location`. It is located in the upper left room. One of the corner rooms is the ideal location for the fire since it would be a room that is the farthest away from all other rooms, which corresponds with anomalies typically being discovered within subfields or methodologies that differ from most others. See figure... for a visualization of the initial environment.

Agent initialization: The number of agents initially created is determined by the parameter `nb_scientists`. Each agent is given a random initial location. Using this location, the model identifies which room they were randomly placed in, and assigns their `expert_level` based on how far away that room is from the fire location. If they're located in the upper left room where the fire is located, they are assigned an `expert_level` of 5. If they're located in one of the two diagonal rooms below this room, they are assigned an `expert_level` of 4, and so on until agents in the bottom right room are assigned an `expert_level` of 1. See figure... for a visualization of this. This allows for the agents' `expert_level` to be approximately normally distributed.

One of the agents is initialized as the proposer of the new theory and the discoverer of the anomaly (assigned a true value to `is_propose`). The proposer is initialized with the belief `must_investigate` and the desire `to_investigate`. The proposer is given an initial random location within the upper left room where the fire is located, to make sure they have an `ex`

`pert_level` of 5. This is done as a more intuitive way to start the model rather than for example adding some initial `smoke_level`. The rest of the agents are initialized with the desire `to_stay`.The proposer is therefore one step ahead of all the other agents. We chose to have only one proposer because that is more realistic and in what typically happens in cases where there is only one anomaly (fire) to be investigated which are the cases this model focuses on. One scientist typically discovers the anomaly and publishes about it, while offering an alternative explanation for it, which could be the start of a new theory or not, and this is what we try to capture in the model, see more in section **??**. It is also common that it is not a singular scientist that makes the discovery, but a small group of scientists collaborating with each other, but we do not include group collaborations in this model for simplicity.

The agents are assigned a value between 0 and 1 for each of their five personality traits. These are truncated normal probability distributions, with the mean and standard deviation for each trait taken from the data of the (**?**) paper. This data was collected from 2,015 scientists to compare scientists with non-scientists personality and other psychological traits. In this paper each personality trait was scored between values 1 and 5 while in the BEN architecture each personality trait takes values between 0 and 1, so the values from the paper were normalized for use in the model. We chose to base the agents' personalities on data to decrease the number of parameters in the model and also to The agent's extraversion personality value is then used as a probability to determine if they have a desire to intentionally influence others, such that the more extroverted they are, the more likely they get the desire `to_influence_others`. Extraversion is chosen to determine whether agents have a desire to influence others because psychological studies have shown that... Furthermore, all agents are initialized with `diseng agement_level` equal to 0, i.e., with no disengagement. This is to make sure that there is no top-level control about how biased agents are, but that they only start to influence each other's `disengagement_level` after the simulation starts. Since all agents except the one proposer start with the desire to stay, agents will immediately start becoming disengaged as the simulation starts, but this conceptually make sense, because in the beginning of the simulation agents do not yet know about the anomaly and therefore are not interested, i.e., are disengaged from it.

Therefore, if the agent is the proposer, their `current_beliefs` would include `must_investi gate` and their `current_desires` would include `to_investigate`. for all other agents, their `current_beliefs` would be empty and their `current_desires` would include `to_stay` and with some probability, would also include `to_influence_others`. The agents' action plans are not explicitly defined in the initialization, but it follows from the initialization that all agents'`current_plan` in the first simulation time step will be Staying, except for the proposer's, which will be Investigating.

C.0.6 Input data

The model does not use input data to represent time-varying processes.

C.0.7 Submodels

A full list of all model parameters is presented in table C.2 and the details of each submodel's implementation are presented below.

TABLE C.2: Summary of model Parameters.

Parameter	Default value	Minimum	Maximum	Description
Global / environment parameters				
`grid_matrix`	CSV file	-	-	This matrix is used in the initialization - see section ref, and is used to define the size of the grid (100 x 100 cells) and to determine which cells are wall, exit, or evacuation area cells. The size of the grid, the layout of the rooms, the thickness of the walls, and the location and width of the exit can all be adjusted by modifying this CSV file.
`center_location`	(8, 8)	(0, 0)	(100, 100)	This determines which cell over which the center of the fire is located. This is also the cell where smoke starts spreading.
`exit_location`	8	1	12	This determines the location of the exit door on the outer edge of the grid. See figure ref which shows what location each integer indicates. More locations than these 12 are possible but this range sufficiently covers the possibilities.

Continued on next page

Table C.2 – continued from previous page

Parameter	Default value	Minimum	Maximum	Description
`nb_scientists`	500	250	1500	This determines the number of agents created at the model initialization. Since agents are initially randomly located across the grid and we want them to be sufficiently uniformly distributed, this number cannot be too small, otherwise their distribution might be too concentrated or too spread. It also does not serve to have too many agents, since then they will be too heavily influenced by one another. If there is one agent per grid cell, then we would have 1000 agents. Using a much higher number than that would be undesirable.
`acceptance_rate`	0.32	-	-	This determines the probability that an agent would be able to publish, and is estimated as the average acceptance rates of scientific journals. According to Elsevier, the average acceptance rate of over 2,300 journals is 32% (Herbert, 2020) so we use that as an estimate for this parameter.
`openness_mean`	0.710	-	-	This determines the mean value for the truncated normal probability distribution from which the agents' openness personality trait is drawn. It is derived from research data from 2,015 scientists (**?**).

Continued on next page

Table C.2 – continued from previous page

Parameter	Default value	Minimum	Maximum	Description
`openness_std`	0.178	-	-	This determines the standard deviation for the truncated normal probability distribution from which the agents' openness personality trait is drawn. It is derived from research data from 2,015 scientists (**?**).
`conscientiousness_mean`	0.568	-	-	This determines the mean value for the truncated normal probability distribution from which the agents' conscientiousness personality trait is drawn. It is derived from research data from 2,015 scientists (**?**).
`conscientiousness_std`	0.180	-	-	This determines the standard deviation for the truncated normal probability distribution from which the agents' conscientiousness personality trait is drawn. It is derived from research data from 2,015 scientists (**?**).
`extraversion_mean`	0.668	-	-	This determines the mean value for the truncated normal probability distribution from which the agents' extraversion personality trait is drawn. It is derived from research data from 2,015 scientists (**?**).
`extraversion_std`	0.195	-	-	This determines the standard deviation for the truncated normal probability distribution from which the agents' extraversion personality trait is drawn. It is derived from research data from 2,015 scientists (**?**).

Continued on next page

Table C.2 – continued from previous page

Parameter	Default value	Minimum	Maximum	Description
`agreeableness_mean`	0.620	-	-	This determines the mean value for the truncated normal probability distribution from which the agents' agreeableness personality trait is drawn. It is derived from research data from 2,015 scientists (**?**).
`agreeableness_std`	0.190	-	-	This determines the standard deviation for the truncated normal probability distribution from which the agents' agreeableness personality trait is drawn. It is derived from research data from 2,015 scientists (**?**).
`neuroticism_mean`	0.593	-	-	This determines the mean value for the truncated normal probability distribution from which the agents' neuroticism personality trait is drawn. It is derived from research data from 2,015 scientists (**?**).
`neuroticism_std`	0.183	-	-	This determines the standard deviation for the truncated normal probability distribution from which the agents' neuroticism personality trait is drawn. It is derived from research data from 2,015 scientists (**?**).
Agent meta-parameters				
level of interaction	0.5	0	1	Describes how well connected the agents in the scientific field are. The more connected they are, the more they come in contact with each other and the faster the smoke spreads.

Continued on next page

Table C.2 – continued from previous page

Parameter	Default value	Minimum	Maximum	Description
peer influence level	0.5	0	1	Describes how much influence agents have on each other's perception of `smoke_level` and `theory_level`, and consequently, on each other's beliefs and actions.
anomaly difficulty level	0.5	0	1	Describes how difficult an anomaly is to study. The more difficult the anomaly is, the longer it takes for it to be understood and the less contribution an individual publication can have to its understanding.
Agent parameters derived from level of interaction				
`vision_amplitude`	-	18	180	Defines the angle width of the agents' perception cones. See figure ref.. At maximum it is 180 degrees and the agents can see all around them (the cone becomes a circle). Agents have to be able to perceive agents and cells around them, so we estimated the minimum as 10% of the maximum.
`max_perceived_distance`	-	3.2	32	Defines the maximum radius of agents' perception cones. The agents with lower `expert_level` than 5 get a fraction of this parameter as their perceived-distance. The maximum was estimated as the length of one room (agents can see everything in the room they are in), and the minimum was estimated as 10% of the maximum.

Continued on next page

Table C.2 – continued from previous page

Parameter	Default value	Minimum	Maximum	Description
`average_speed`	-	0	10	Defines the speed at which the agents more between their location and target. If it gets too high, the agents' movements become frantic and more random, so the speed is capped at a maximum of 10. At minimum, the agents are static and do not move.
`wandering_distance`	-	0	50	Defines how far away the agents' target can be. The maximum is estimated as roughly twice the `max_perceived_distance` (length of two rooms). At minimum, the agents are static and do not move.
`agent_size`	-	1	4	Defines how big the agents' triangular representation's size is. Estimated roughly as 1 to 4 cells' lengths.
`smoke_spread_rate`	-	0	0.5	Defines how fast the smoke spreads. Mechanisms explained in the update `smoke_level` submodel below.
Agent parameters derived from peer influence level				
`basic_effect`	-	0	1	Defines how much any agent can influence another agent.
`higher_expertise_effect`	-	0	1	Defines how much extra influence an agent over another agent because they have a higher `expert_level`.
`influence_desire_effect`	-	0	1	Defines how much extra influence an agent over another agent because they have a desire to intentionally influence others.
Agent parameters derived from anomaly difficulty level				

Continued on next page

Table C.2 – continued from previous page

Parameter	Default value	Minimum	Maximum	Description
`max_impact`	-	0.01	0.10	Defines the maximum impact a single publication can have over `smoke_level` and `theory_level`. Publication cannot have no impact, so the minimum is set as 1%. Also, the impact cannot be too high, otherwise it would be too unrealistic, and so the maximum is set as 10% (e.g., `smoke_level` can increase 10% because of a single publication).
`partial_impact`	-	0	1	When an agent publishes the results of their investigation of the anomaly, they have an indirect, partial impact on the development of the new theory, and vice versa, when an agent publishes the results of their investigation into the new theory, they have an indirect, partial impact on the evidence about the anomaly. This parameter defines what percentage of the impact of the direct investigation is added to the indirect investigation.
`timesteps_for_full _understanding`	-	2	20	Defines how many time steps (how much time) are needed for an agent to have a full understanding of the anomaly (or at least of the part of the anomaly they are investigating) such that they can publish with confidence.

C.0.7.1 update smoke_level

This process defines how the evidence supporting the anomaly (the smoke) accumulates at the fire location and spread to neighboring cells. This spatial representation of information diffusion

helps in drawing a parallel between the fire evacuation model and this model, while allowing to intuitively simulate how information about the anomaly might reach and influence scientists in different parts of the scientific field. This submodel starts by taking the accumulation of agents publication impacts from the previous time step (see Publishing submodel below), and adding it to the `smoke_level` of the grid cell located at `center_location`. This is the grid all located at the center of the fire. Then every grid cell which has a higher `smoke_level` than its neighboring cells, shares some of the `smoke_level` with them and losing some itself, simulating the gradual dissipation of smoke. Specifically, when a cell has a neighboring cell with less `smoke_level`, the neighboring cell copies its `smoke_level` and the cell itself loses some `smoke_level` defined by the parameter `smoke_spread_rate`. The smoke spread is therefore at its fastest when `smoke_spread_rate` is zero, and slows down as `smoke_spread_rate` increases.

C.0.7.2 update theory_level

This process defines how `theory_level` accumulates, which represent how well- developed the new theory is. `theory_level` is a global variable, and it is updated every time step simply by adding the accumulation of agents publication impacts from the previous time step (see Publishing submodel below). `theory_level` is a global variable rather than a variable assigned to each grid cell like `smoke_level` to simplify the model's mechanisms. The assumption here is that agents who are considering abandoning the current theory in favor of a new one will seek out publications about the new theory's development (represented by `theory_level`) and it is less important that they are exposed to it without seeking it out like `smoke_level` is.

C.0.7.3 perception & interpretation

Agents' perception ability is described in detail in section C.0.4.7. It occurs through a geometrical cone, called the perception are, shown in figure **??**, which agents can perceive everything inside of. The cone's angle is defined by the parameter `vision_amplitude` and its diameter is defined by the variable `perceived_distance`. Each agent's `perceived_distance` is calculated by multiplying the parameter `max_perceived_distance` by `expert_level` / 5. This way the agents with highest expertise (`expert_level` = 5) have the highest `perceived_distance`, for every lower `expert_level`, `max_perceived_distance` is decreased by 20%. This is to simulate a simple variability between the agents, where agents with higher expertise can perceive smoke from farther away, simulating that they understand the anomaly the most, and can also perceive other agents from farther away, simulating that they have a bigger reach over other agents because of their relevant expertise. Agent's `expert_level` is assigned in the model's initialization.

Every time an agent perceives another, their `disengagement_level` is updated. `disengagemen t_level` is an agent variable used to represent how disengaged an agent is from the discourse

surrounding the anomaly and the new theory, and takes values between zero and one, zero representing no disengagement and one representing maximum disengagement. Agents who are disengaged are ones who do not have the belief `smoke_serious`. The assumption here is that the more an agent comes in contact with others who are disengaged, the more disengaged they become, and the more disengaged they become, the less serious they will interpret the evidence (`smoke_level`) to be. This indirectly implies that the group becomes more resistant the more time passes while only a small number of agents are investigating the anomaly.

Not all agents are affected by others the same way. Research shows that conscientiousness (one of the five personality traits) is closely linked with social conformity values (Roccas et al., 2002). So the more conscientiousness an agent is, the less affected they are by others. How much an agent influences another is defined by the parameters `basic_effect`, `higher_expertise_effect`, and `influence_desire_effect`, which are derived from the meta-parameter `peerinfluencelevel`. Furthermore, because `disengagement_level` takes values between 0 and 1, we normalize the influence of a single agent by dividing the effect of their influence by an estimated variable for the maximum number of agents an agent is expected to perceive. This variable, `max_perceived_agents` is calculated as follows and relies on the assumption that agents are uniformly distributed across the six rooms, and on the estimated length of one room.