Responsible Mechanism Design

Pavel Naumov

University of Southampton, United Kingdom
p.naumov@soton.ac.uk

Abstract

Traditionally, the goal of mechanism design was to promote socially desirable behaviour of rational agents, to achieve fairness, or to promote efficiency. I would like to suggest a new subfield of mechanism design, Responsible Mechanism Design, focused on achieving individual accountability of agents for their contributions to the outcome of collective decisions.

Why Now?

From ancient democracies to the AAAI paper reviewing process, humans have long participated in collective decision-making, and just as long, they have assigned blame when outcomes go awry. Yet the mechanisms we use to make such decisions remain remarkably primitive. Most reduce to variants of majority rule, consensus, delegation, or fixed hierarchies of power. This simplicity is not accidental: humans tend to avoid mechanisms they cannot intuitively grasp, and our slow, error-prone communication makes complex coordination difficult. As a result, there has been little serious attempt to develop more sophisticated decision procedures. Compounding this is the fuzziness of responsibility itself. The final interpretation of legal responsibility is often left to courts, while moral responsibility has become the subject of conceptual studies in the humanities rather than mathematically grounded scientific analysis.

The situation is changing as AI agents start taking a significant part in collective decision-making. Not only are such agents easily capable of engaging in complex decision-making protocols, but they can communicate with each other much faster. As a result, for example, we should expect that simplistic traffic-sign-based road rules designed for humans will be replaced by sophisticated coordination protocols that would allow self-driving cars to avoid collisions while never stopping at intersections. Additionally, recent works in AI introduced rigorous, mathematically precise definitions of responsibility that can be applied to humans and artificial agents and can also be used by AI agents to reason about their responsibility. Those definitions can now be used to adjust group decision-making protocols in a way that enforces agents' individual accountability.

How to Define Responsibility?

To define responsibility with mathematical precision, I first capture multiagent interactions as games. For instance, consider the following very simple road situation depicted at the left of Figure 1:

Example 1 (two cars) *Two self-driving cars, a and b, are approaching a non-regulated intersection with the same speed. Car a is further away from the intersection than car b. Each of the cars has only two possible actions: to slow down ($-$) or to maintain the current speed (0). If car a maintains the current speed and car b decides to slow down, the cars collide. No collision happens under all other action profiles.*

Since I want to discuss responsibility for the crash, not for breaking contemporary traffic rules, let's ignore the latter when analysing this traffic situation. The table on the right of Figure 1 captures the example as a strategic game. Shaded cells of the table represent "undesirable" outcomes in which a collision happens. I call this game "strategic" because I assume that both cars must choose their actions at the same time (independently). A reader who is used to traditional games with utility functions can assume that both agents have utility -1 in the undesirable (shaded) outcome and utility 0 in all other outcomes.

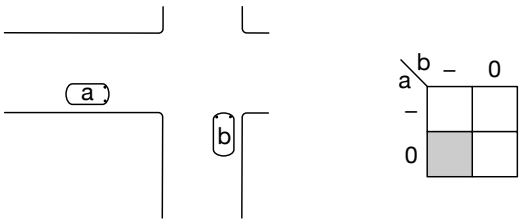

Figure 1: Two cars example.

Let us suppose that car a maintains the current speed (0) and car b decides to slow down ($-$). The cars collide. Who is to blame?

In philosophy, one of the most commonly discussed (Widerker and McKenna 2003) definitions of responsibility is based on the principle of alternative possibilities: *. . . a person is morally responsible for what he has done only if he could have done otherwise* (Frankfurt 1969). In recent

literature, "could have done otherwise" has been interpreted as having a strategy to prevent the undesired outcome (Yazdanpanah, Dastani, Jamroga, Alechina, and Logan 2019; Naumov and Tao 2019, 2020a; Baier, Funke, and Majumdar 2021; Shi 2024; Shi and Naumov 2025). In this paper, I refer to the notion of responsibility based on the principle of alternative possibilities as "counterfactual responsibility". I say that an agent is counterfactually responsible for an outcome if the outcome happened and the agent had a strategy that *guarantees* that the outcome is prevented no matter how the other agents act. In the two-car example, both cars had a strategy to prevent the collision: for car a, such a strategy is to slow down $(-)$ and for car b it is to maintain the current speed (0). Thus, *both cars are counterfactually responsible for the collision*.

Let us now consider a slightly modified example depicted in Figure 2.

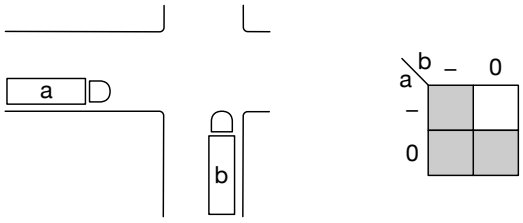

Figure 2: Two lorries example.

Example 2 (two lorries) *Same setting as in Example 1 except the two vehicles are now lorries, which are longer than cars. The two vehicles avoid collision only if lorry a slows down and lorry b maintains the current speed.*

Note that, in Example 2, to avoid collision, the vehicle must coordinate their actions. None of them has a strategy that guarantees avoiding the collision no matter how the other vehicle acts. Hence, if a collision happens, *neither of the self-driving lorries is counterfactually responsible for it*.

Let me further analyse Example 2. Suppose that, approaching the intersection, both vehicles decided to slow down $(-)$. In this case, the collision happens, but the two vehicles contributed to it differently. While the action $-$ of lorry a left the possibility that the collision will not happen, the action $-$ of lorry b *guaranteed that the collision happens no matter what the action of the other vehicle is*. In this situation, intuitively, it is only vehicle b that should be held responsible. This type of responsibility is known as *responsibility for seeing to it that* (Broersen 2011a,b; Naumov and Tao 2021, 2023). Its properties have been studied in STIT ("seeing-to-it that") logic (Belnap and Perloff 1990; Horty 2001; Horty and Belnap 1995; Horty and Pacuit 2017; Olkhovikov and Wansing 2019). In our example, under action profile $(-,-)$, lorry b is responsible for seeing to it that the collision happened. At the same time, in Example 1, neither of the two cars is responsible for seeing to it that the collision occurs.

Counterfactual responsibility and responsibility for seeing to it that are the two most commonly studied modes of responsibility in philosophy, logic, and AI literature. I mention alternative definitions of responsibility at the end of this paper. In general, different forms of responsibility might be appropriate in different contexts. In this paper, to keep the presentation simple, I say that *an agent is responsible for an outcome if she is responsible either counterfactually or for seeing to it*.

Figure 3: Responsibility allocation in Example 1 (left) and Example 2 (right).

Figure 3 shows which agent is responsible for each undesirable (greyed out) outcome. Label "ab" means that both agents are responsible.

To further illustrate the definition of responsibility, let us consider the example originally suggested by Halpern (2016).

Example 3 (fish) *Factories a, b, and c have accumulated 20kg, 10kg, and 10kg of a pollutant, respectively. They must decide whether they drop this pollutant into a nearby river today. If at least 15kg of pollutant is simultaneously dropped into the river, the fish in the river dies.*

Let us again consider the case when the decisions are made independently. Note that neither of the three factories has a strategy to prevent the death of the fish because the other two factories, together, have accumulated enough pollutants to kill the fish. Thus, if the fish dies, none of the factories is responsible for this counterfactually. At the same time, only factory a's action $+$ (drop the pollutant) guarantees that the fish is dead, no matter what the other factories do. Thus, only factory a, if it chooses to drop the pollutant, is responsible for seeing to it that the fish is dead. In particular, if only factories b and c drop the pollutant, then the fish dies, and nobody is responsible for this. Figure 4 shows responsibility allocation in this example. Action 0 consists of not dropping the pollutant.

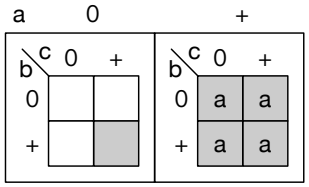

Figure 4: Responsibility allocation in Example 3.

So far, we have only considered strategic games–the type of games where all agents act simultaneously and just once. My next example will be modelled using *extensive form* game in which agents act one at a time. The example is based on the real-life story: the death of Ms. Elaine Herzberg, the

first pedestrian killed by a self-driving vehicle. The investigation of this tragedy has revealed that the Uber self-driving car is designed to pass control to a human driver a few seconds before the collision (NTSB 2019).

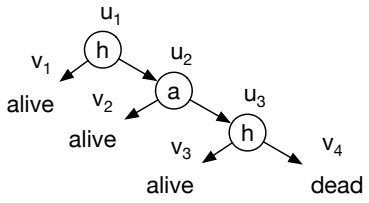

Figure 5: Shared vehicle control between human driver (h) and an autopilot (a).

Example 4 (autopilot) *After an object appears in front of the vehicle, the human driver (h) can stop the vehicle or pass control to the autopilot (a). Autopilot can stop the vehicle or return control to the human driver. In the latter case, the human can stop the vehicle or let it hit the pedestrian, see Figure 5.*

Suppose that neither the human driver nor the autopilot decides to push the brakes, and the pedestrian dies. This scenario corresponds to *decision path* u_1, u_2, u_3, v_4. Note that the human had a strategy (break) to prevent death at nodes u_1 and u_3. Autopilot had such a strategy at node u_2. Thus, both of them are *counterfactually* responsible for the death. At the same time, only the human is responsible for seeing to it that the pedestrian is dead. This is because the human took the action (do not break) at node u_3 that made the death unavoidable.

My next example is modelled by a *concurrent* game where some decisions are made sequentially and some simultaneously. The example models the launch procedure of US Minutemen II intercontinental ballistic missiles (United States Air Force 2024).

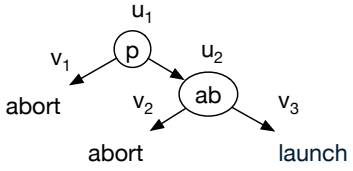

Figure 6: Two Keys Protocol example.

Example 5 (two keys) *Only the President of the United States can authorise the launch of nuclear missiles. If such an authorisation is issued, a secret authorisation code is transmitted to the launch facility and entered into the system. Then, two on-duty officers, a and b, must simultaneously turn two keys located 6 feet from each other to initiate the launch. If one of the officers does not turn the key, missiles are not launched, see Figure 6.*

First, imagine that the President authorises to start a nuclear war, both officers turn the keys, and ... half of the world is destroyed. This corresponds to the decision path u_1, u_2, v_3. Who is to be blamed? It is easy to see that all three parties had strategies to prevent the launch. Thus, all three of them are responsible counterfactually. At the same time, none of them is responsible for seeing-to-it that the missiles are launched, because none of them took an action that alone guarantees the launch.

What is a Responsible Mechanism?

I call the mechanism design "responsible" if it has good responsibility-related properties. What such properties are depends on the intended application. Below, I list examples of such properties.

Gap

Responsibility gap of a decision-making mechanism is the set of all undesirable outcomes in which no single party is individually responsible. For example, recall from Example 3 that if only factories b and c drop the pollutant, then the fish dies, and nobody is responsible for this. We capture this fact by not labelling with any agent the greyed-out cell in Figure 4 that corresponds to this outcome (action profile). Thus, this outcome belongs to the responsibility gap of the simultaneous pollutant dropping mechanism.

At the same time, the mechanisms from Example 1, Example 2, Example 4, and Example 5 are *gap-free*. Indeed, under each of these mechanisms, if an undesirable outcome (vehicle collision, death of a pedestrian, world distraction) happens, then at least one agent is individually responsible for it (either counterfactually or for seeing to it).

In most situations, a well-designed group decision mechanism must be gap-free. Responsibility gap has been extensively discussed in the literature (Braham and van Hees 2011; Duijf 2018; List 2021; Duijf 2022; Dastani and Yazdanpanah 2023; Shi and Naumov 2025).

Diffusion

The term *diffusion of responsibility* refers to a situation when more than one agent is simultaneously responsible for the undesirable outcome. For instance, in Figure 3, the greyed-out cells labelled with "ab" represent the outcomes from Example 1 and Example 2 in which the diffusion happens.

The diffusion of responsibility also happens in outcome v_4 of Example 4, see Figure 5. This is because, as we discussed earlier, the autopilot and the human driver are both responsible for that outcome. The diffusion of responsibility between all three agents, the President, officer a, and officer b, happens in outcome v_3 of Example 5, see Figure 6.

At the same time, no diffusion of responsibility happens in any of the undesirable outcomes of Example 3. Each time the fish dies, it is only factory a who is responsible, see Figure 4.

Diffusion of responsibility has been studied in social sciences (Mynatt and Sherman 1975; Forsyth, Zyzniewski, and Giammanco 2002; Liu, Liu, and Wu 2022), law (Iusmen

2020; Rowan, Kan, Frick, and Cauffman 2022), ethics (Bleher and Braun 2022), and neuroscience (Feng, Deshpande, Liu, Gu, Luo, and Krueger 2016). Generally speaking, diffusion is an undesirable property of a decision-making mechanism that leads to "circle of blame" and "bystander effect".

Fragmentation

Example 6 (postdoc hiring) *I finally got funding to hire a postdoc to do research in AI Ethics. Eight candidates applied for the position, many of them have a background in Ethics and AI. Alice and Wendy have a background in both, see Table 1. Not having enough time, I asked my assistant to pre-screen the candidates and give me a short list of five applications to choose from. From the short list, see the table, I selected Jim.*

As one can see from the table, Jim is one of the least qualified candidates. Who is to be blamed in this situation?

	Ethics	AI	Short listed?	Selected?
Mike		✓	✓	
Alice	✓	✓		
Bob			✓	
Cathy	✓			
Jim			✓	✓
Tom		✓	✓	
Linda		✓	✓	
Wendy	✓	✓		

Table 1: Postdoc selection

One might choose to blame my assistant. After all, by not shortlisting Alice and Wendy, my assistant have *seen to it* that the candidates who have a background in both Ethics as well as AI, are not selected.

A more sensible approach, perhaps, is to notice that, by dropping all candidates with a background in Ethics, my assistant has seen to it that the selected postdoc has no background in Ethics. At the same time, I have seen to it (and also responsible counterfactually) that the selected candidate has no background in AI. In other words, we both are responsible, but for *different aspects* of the outcome. Braham and van Hees (2018) refer to such situations as *fragmentation of responsibility*. Similarly to diffusion, fragmentation can lead to circles of blame and create the bystander effect. One of the goals of responsible mechanism design could be the elimination of responsibility fragmentation.

Dictatorship

Example 7 (US Consitution) *The US Constitution defines the mechanism for passing new laws in the United States. The Congress (treated in this example as a single agent c) can reject a bill outright. If Congress passes the bill, the President (agent p) can sign or veto the bill. If the President vetoes the bill, the Congress can override the veto, see Figure 7.*

It is a matter of personal opinion if passing or rejecting the bill is an "undesirable" outcome. In such situations, it makes sense to consider responsibility in all outcomes (for either passing or rejecting the bill, depending on the outcome).

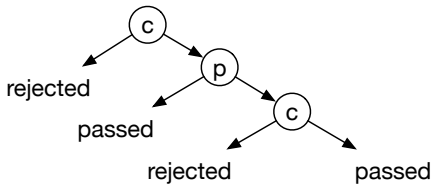

Figure 7: US Constitution mechanism.

It is interesting to note that no matter what happens, the Congress is always counterfactually responsible for the outcome. This is because the Congress has an upfront strategy to guarantee that the bill is rejected and an upfront strategy to guarantee that the bill is passed, see Figure 7. In general, I call a decision mechanism a *dictatorship* if an agent (a "dictator") has an upfront strategy to achieve both outcomes.

Any dictatorship is gap-free because the dictator can always prevent any outcome. However, many people would probably find a dictatorship to be an undesirable property of a mechanism design. In many cases, one would expect responsible mechanism design to eliminate responsibility gaps without using a dictatorship. In (Naumov and Tao 2025), we introduced a notion of an "elected dictatorship" as a class of mechanisms where the group decision process, essentially, comes down to an election of a dictator who might unilaterally make the decision. We have shown that, for some class of mechanisms, elected dictatorships are the only way to eliminate the responsibility gap.

Distributed Responsibility

In some settings, responsible mechanism design should ensure that every agent bears responsibility in at least one possible outcome—so that all participants have a genuine stake in the game. This fosters not only accountability but also a sense of inclusion and ownership over the collective decision-making process. I suggest using the term "distributed responsibility" to refer to situations when each agent is responsible in at least one outcome. The two diagrams in Figure 3 show that the responsibility is distributed in Example 1 and Example 2. At the same time, the responsibility is not distributed in Example 3. As one can see in Figure 3, factories b and c are never responsible in that setting.

How Can We Design Responsible Mechanisms?

Order

One of the most effective ways to close the responsibility gap is to design mechanisms in which agents act in a certain order rather than concurrently.

Example 8 *Two self-driving cars, a and b, are approaching the intersection with the same distance; each of them can either slow down (action $-$) or maintain the current speed*

(action 0). If both cars choose the same action, they will collide. See Figure 8, left.

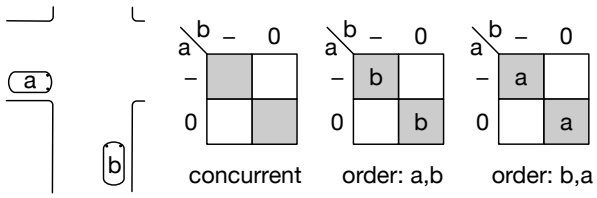

Figure 8: Symmetric road situation.

Note that if the cars choose actions concurrently (independently), then neither of them can prevent the collision. Also, neither of the actions by itself guarantees the collision. Thus, if the cars act concurrently and a collision happens, then neither of them is responsible for the collision, see the "concurrent" diagram in Figure 8. However, if one of the cars chooses the action first and, say, communicates this action wirelessly to the other car, then the second car has a strategy to avoid a collision. For example, if the first car decides to slow down, then the second car will avoid collision by maintaining the current speed, see the last two diagrams in Figure 8. Thus, in this example, *order eliminates the responsibility gap without creating diffusion of responsibility* between agents. However, in this example, the order *does not distribute* responsibility—if a collision happens, it is always the second car that is responsible, see Figure 8.

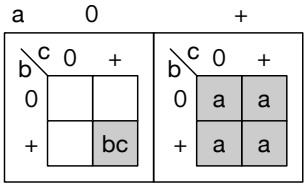

Figure 9: Allocation of responsibility in Example 3 if the factories act in order a, b, c.

Let us now see how order affects responsibility in Example 3. Recall that Figure 4 shows responsibility allocation in this example under the assumptions that all three factories make their choices concurrently. Figure 9 depicts responsibility assuming the order a, b, c. As the figure shows, this order *eliminates the gap and distributes the responsibility between all three agents*. However, it introduces diffusion of responsibility between b and c in one of the outcomes.

In general, *requiring agents to take actions in an order, rather than concurrently, eliminates the responsibility gap and increases the distribution of responsibility, but it can also diffuse the responsibility.* The choice of the specific order is important. Figure 10 shows that both gap and diffusion are eliminated if the order b, c, a is used instead of a, b, c. At the same time, responsibility is no longer fully distributed: factory b is never responsible for the outcome.

The next example shows that just changing the order might not be enough to close the gap without introducing

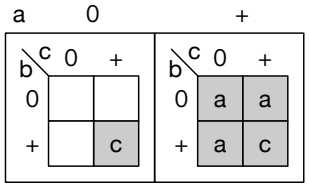

Figure 10: Responsibility in Example 3 if the factories act in order b, c, a.

diffusion.

Example 9 (three cars) *Three self-driving cars, a, b, and c, are approaching an intersection. Each of them can either slow down (action $-$) or maintain the current speed (action 0). If all three cars choose the same action, they will collide. Otherwise, the collision will be avoided, see Figure 11.*

In the second from the right diagram in Figure 11, I show the allocation of responsibility in this example using order a, b, c. This order eliminates the gap but introduces diffusion between two agents. Change of the order will not eliminate the diffusion because this setting is symmetric. However, the diffusion can be eliminated if, first, cars a and b choose their actions *concurrently*, then car c chooses its action, see rightmost diagram in Figure 11.

Imperfect Information

Observe that to achieve the allocation from the right-most diagram in Figure 11, it is not actually strictly necessary that cars a and b choose their actions concurrently. The same allocation can be achieved if the cars make the choices in order a, b, c, but the mechanism *hides* from b the choice made by a. I assume that, as before, car c gets the information about choices made by a as well as b. This is an example of a decision mechanism with *imperfect information*. For such mechanisms, the definitions of counterfactual and seeing-to-it responsibilities need an adjustment. I will explain the adjustment using the Drawing Straws mechanism. In 2017, this mechanism was used to decide who gets a seat in the Northumberland County Council (England) after votes were evenly divided (Elgot 2017).

Example 10 (drawing straw) *Candidate a take two straws, a short and a long, and holds them between her fingers, showing only the ends of the straws. Candidate b pulls one of the straws. If he pulls the long straw, he wins the elections.*

Suppose that b draws a short string and loses the elections. Should we say that agent b is responsible for losing the elections? Technically, b had a strategy (pull long string) to win the elections. Also, the action that b took guaranteed that he was going to lose. Thus, technically, under my original definition, b is responsible for his loss. Of course, this is a counterintuitive conclusion. As it has been argued in the literature, in an imperfect information case, the agent must *know* which action prevents the undesirable outcome to be counterfactually responsible for it (Yazdanpanah, Dastani, Jamroga, Alechina, and Logan 2019; Naumov and Tao

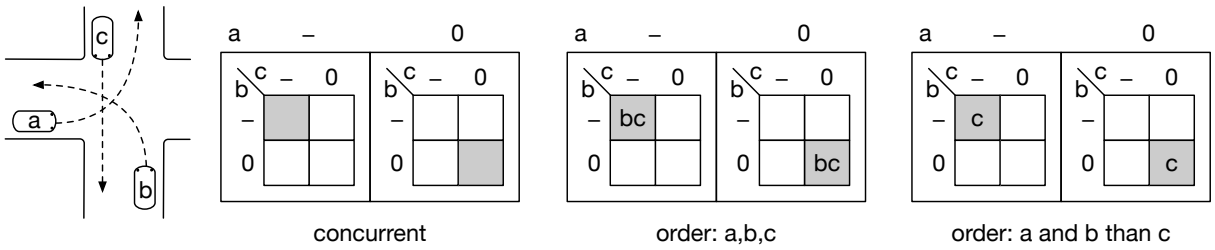

Figure 11: Three cars example.

2020b). Similarly, to be responsible for seeing to it, the agent must *know* that his or her action makes the outcome unavoidable (Naumov and Tao 2023). After these adjustments, mechanisms with imperfect information become a powerful tool for controlling the diffusion of responsibility. Indeed, *by denying selected agents access to some of the information, we potentially can make them not responsible for the outcomes*.

Control Redistribution

Another way to guarantee that a decision-making mechanism has the right responsibility-related properties is to *redistribute control* between agents. To see how this technique works, let me return to Example 4. The left diagram in Figure 12 captures this example as a *transition system*, where autopilot (a) and human driver (h) control different states. In a self-driving mode, under normal conditions, the system

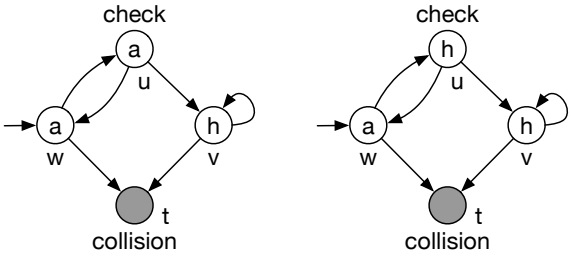

Figure 12: Actual Uber mechanism that allows diffusion (left) and a diffusion-free modified mechanism (right).

alternates between states w and u. If the autopilot makes a mistake, the system might transition from state w to "collision" state t. In state u, the autopilot a might decide to pass the control to the human driver h by transitioning to state v. The human driver can either keep driving using a loop transition at state v or make a mistake and end in the collision state t.

In the case of Ms Elaine Herzberg's tragic death, after cycling between states w and u, the autopilot passed control to a human. The human made a mistake that led to the collision. The autopilot and the human driver are both responsible for the collision counterfactually. Autopilot could have prevented the collision by looping between states w and u. The human driver could have prevented it by looping at state v. The human driver is also responsible for seeing-to-it that Ms Herzberg died because the human action of transitioning the system from state v to state t made the death unavoidable. This diffusion of responsibility between the autopilot and the human driver could be avoided by making a change in the way Uber's vehicle control mechanism is designed. If a human, not an autopilot, controls state u, as shown on the right of Figure 12, then the autopilot does not have a strategy to prevent collision by looping between states w and u no matter what the human does. In practice, this change means that it will be up to the human driver to take over the control of the vehicle if the driver finds it necessary. Such a change would eliminate the diffusion by making the human solely responsible for the collision resulting from transitioning from state v to state t.

What's Next?

I see Responsible Mechanism Design not as a single problem, but as a new field focused on designing collective decision-making processes that enforce individual accountability. The precise meaning of accountability will depend on context. In some applications, it may mean ensuring mechanisms are gap-free, diffusion-free, or fragmentation-free. In others, it may involve distributed responsibility or the exclusion of dictatorships. In certain domains, entirely different notions of responsibility may be required, such as higher-order responsibility (Shi 2024), probabilistic responsibility (Duijf and van De Putte 2022), or best-effort responsibility, where agents are credited for attempting to prevent undesirable outcomes (Braham and van Hees 2018). Constraints on the cost of prevention strategies may also matter (Kagan 1991; Cao and Naumov 2017). And in systems involving children, animals, or AI agents, we may need to restrict accountability to proper moral agents, or else explore forms of collective or institutional responsibility.

Beyond design itself, the field must explore which responsibility criteria are jointly satisfiable, and under what conditions. It must also engage with complexity: can we efficiently verify whether a mechanism has desirable responsibility properties? Can agents compute strategies that guarantee avoidance of harm? And where perfect mechanisms are impossible, can we still minimise the responsibility gap or diffusion?

Mechanism design has taught us how to allocate goods, effort, and risk. Responsible Mechanism Design teaches us how to allocate accountability. That is the next frontier.

Acknowledgments

I would like to express my gratitude to my students and co–authors who worked with me on responsibility-related topics and without whom this paper would have never been written: Rui Cao, Xiulin Cui, Kaya Deuser, Junli Jiang, Sophia Knight, Anna Ovchinnikova, Alexandra Pavlova, Bahar Rastegari, Qi Shi, Jia Tao, and Rui-Jie Yew.

References

Baier, C.; Funke, F.; and Majumdar, R. 2021. A Game-Theoretic Account of Responsibility Allocation. In *30th International Joint Conference on Artificial Intelligence (IJCAI-21)*.

Belnap, N.; and Perloff, M. 1990. Seeing to it that: A canonical form for agentives. In *Knowledge representation and defeasible reasoning*, 167–190. Springer.

Bleher, H.; and Braun, M. 2022. Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems. *AI and Ethics*, 2(4): 747–761.

Braham, M.; and van Hees, M. 2011. Responsibility Voids. *The Philosophical Quarterly*, 61(242): 6–15.

Braham, M.; and van Hees, M. 2018. Voids or fragmentation: Moral responsibility for collective outcomes. *The Economic Journal*, 128(612): F95–F113.

Broersen, J. 2011a. Deontic epistemic STIT logic distinguishing modes of mens rea. *Journal of Applied Logic*, 9(2): 137–152.

Broersen, J. M. 2011b. Making a start with the STIT logic analysis of intentional action. *Journal of Philosophical Logic*, 40(4): 499–530.

Cao, R.; and Naumov, P. 2017. Budget-Constrained Dynamics in Multiagent Systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 915–921.

Dastani, M.; and Yazdanpanah, V. 2023. Responsibility of AI systems. *Ai & Society*, 38(2): 843–852.

Duijf, H. 2018. Responsibility voids and cooperation. *Philosophy of the social sciences*, 48(4): 434–460.

Duijf, H. 2022. *The logic of responsibility voids*. Springer.

Duijf, H.; and van De Putte, F. 2022. The problem of no hands: responsibility voids in collective decisions. *Social Choice and Welfare*, 58(4): 753–790.

Elgot, J. 2017. Lib Dem and Tory candidates draw straws in Northumberland vote. https://www.theguardian.com/politics/2017/may/05/lib-dem-and-tory-candidates-draw-straws-in-northumberland-vote. Accessed: 2025-01-15.

Feng, C.; Deshpande, G.; Liu, C.; Gu, R.; Luo, Y.-J.; and Krueger, F. 2016. Diffusion of responsibility attenuates altruistic punishment: A functional magnetic resonance imaging effective connectivity study. *Human brain mapping*, 37(2): 663–677.

Forsyth, D. R.; Zyzniewski, L. E.; and Giammanco, C. A. 2002. Responsibility diffusion in cooperative collectives. *Personality and Social Psychology Bulletin*, 28(1): 54–65.

Frankfurt, H. G. 1969. Alternate possibilities and moral responsibility. *The Journal of Philosophy*, 66(23): 829–839.

Halpern, J. Y. 2016. *Actual causality*. MIT Press.

Horty, J.; and Pacuit, E. 2017. Action types in STIT semantics. *The Review of Symbolic Logic*, 10(4): 617–637.

Horty, J. F. 2001. *Agency and deontic logic*. Oxford, England: Oxford University Press.

Horty, J. F.; and Belnap, N. 1995. The deliberative STIT: A study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24(6): 583–644.

Iusmen, I. 2020. Whose children? Protecting unaccompanied migrant children in Europe: A case of diffused responsibility? *The International Journal of Children's Rights*, 28(4): 925–949.

Kagan, S. 1991. *The Limits of Morality*. Oxford Ethics Series. Clarendon Press.

List, C. 2021. Group agency and artificial intelligence. *Philosophy & technology*, 34(4): 1213–1242.

Liu, D.; Liu, X.; and Wu, S. 2022. A Literature Review of Diffusion of Responsibility Phenomenon. In *2022 8th International Conference on Humanities and Social Science Research (ICHSSR 2022)*, 1806–1810. Atlantis Press.

Mynatt, C.; and Sherman, S. J. 1975. Responsibility attribution in groups and individuals: A direct test of the diffusion of responsibility hypothesis. *Journal of Personality and Social Psychology*, 32(6): 1111.

Naumov, P.; and Tao, J. 2019. Blameworthiness in Strategic Games. In *Proceedings of Thirty-third AAAI Conference on Artificial Intelligence (AAAI-19)*.

Naumov, P.; and Tao, J. 2020a. Blameworthiness in Security Games. In *Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*.

Naumov, P.; and Tao, J. 2020b. An Epistemic Logic of Blameworthiness. *Artificial Intelligence*, 283. 103269.

Naumov, P.; and Tao, J. 2021. Two Forms of Responsibility in Strategic Games. In *30th International Joint Conference on Artificial Intelligence (IJCAI-21)*.

Naumov, P.; and Tao, J. 2023. Counterfactual and seeing-to-it responsibilities in strategic games. *Annals of Pure and Applied Logic*, 174(10): 103353.

Naumov, P.; and Tao, J. 2025. Responsibility Gap in Collective Decision Making. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*.

NTSB. 2019. Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian Tempe, Arizona March 18, 2018. National Transportation Safety Board, Report NTSB/HAR-19/03. Washington, DC.

Olkhovikov, G. K.; and Wansing, H. 2019. Inference as doxastic agency. Part I: The basics of justification STIT logic. *Studia Logica*, 107(1): 167–194.

Rowan, Z. R.; Kan, E.; Frick, P. J.; and Cauffman, E. 2022. Not (entirely) guilty: The role of co-offenders in diffusing responsibility for crime. *Journal of Research in Crime and Delinquency*, 59(4): 415–448.

Shi, Q. 2024. Responsibility in Extensive Form Games. In *Proceedings of 38th AAAI Conference on Artificial Intelligence (AAAI-24)*.

Shi, Q.; and Naumov, P. 2025. Responsibility in Multi-Step Decision Schemes. *Journal of Philosophical Logic*.

United States Air Force. 2024. Launching Missiles. https://www.nationalmuseum.af.mil/Visit/Museum-Exhibits/Fact-Sheets/Display/Article/197675/. Accessed: 2024-09-24.

Widerker, D.; and McKenna, M., eds. 2003. *Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities*. Burlington, VT: Ashgate.

Yazdanpanah, V.; Dastani, M.; Jamroga, W.; Alechina, N.; and Logan, B. 2019. Strategic responsibility under imperfect information. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 592–600. International Foundation for Autonomous Agents and Multiagent Systems.