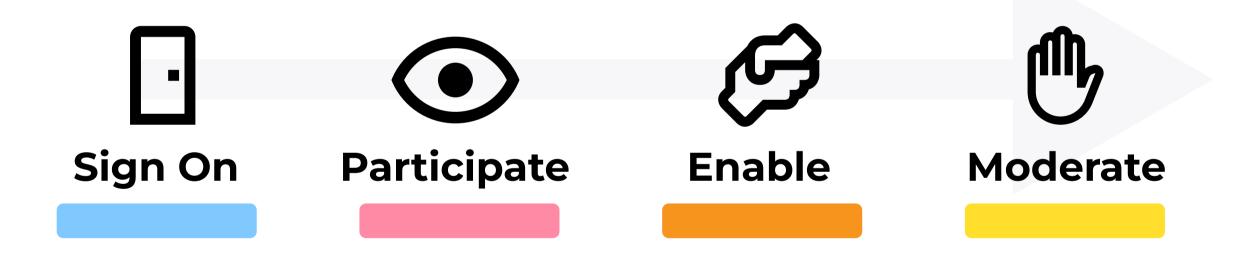
The Interactive Services Model (ISM)

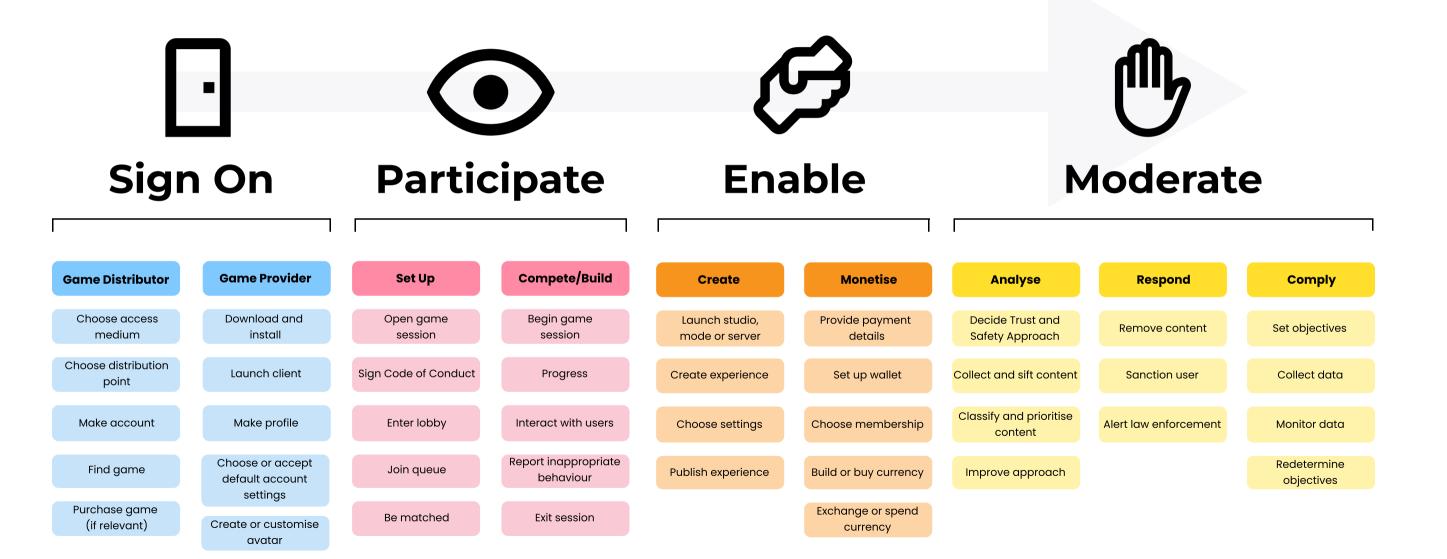
This Interactive Services Model (ISM) breaks down the user journey and platform workflows across interactive services, including social gaming experiences. It covers the sign on process, participation, additional functionalities that support the game experience, and moderation workflows.



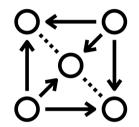


For each phase, from Sign On to Moderate, there are a set of functional models which show the functional and architectural processes that platforms take to achieve the objectives of each phase. The models are representative of four gaming genres, which are explained on the following page.





Given the breadth of gaming experiences, PUBLIC decided to focus on certain game genres for the purpose of the ISM. We prioritised four genres, namely, Battle Arena, Battle Royale, Shooter, and Sandbox games. These genres were prioritised based on their popularity in UK, user demographics and interactive functionality. By interactive functionality we mean the level of user-to-user and user-to-user-generated content, interactions capable within a given game in these genres.



Battle Arena

Games where users control a character that has set abilities in a team of characters, usually controlled by other users. These games are played in symmetric maps that have a set base, often opposite the other team's base. The objective of the game is usually to destroy the other team's base.



Battle Royale

Games where users start from scratch every game and have to survive to be the last person or team standing to win the game. A significant share of gameplay revolves around scavenging random loot. The game forces users into confrontation (e.g., by a shrinking safe zone).



Shooter

Games in which the user controls a character which shoots enemies to defeat them.



Sandbox

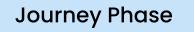
Games that give the user a large amount of creativity and freedom on how to play the game, what to do, or to create the game's content themselves with little to no predetermined objectives.



Sign On | Game Distribution Platforms

Game Distribution Platforms are digital platforms that distribute games to users.

The Sign On process for game distribution platforms covers choosing the access medium and distribution point by which the user will find and access the game. The user must then make an account with the game distributor, verifying any relevant details, choose their game, and make a purchase, if necessary. In addition, some game distribution platforms will provide their own user-to-user communication features or overlay with in-game accounts and communication features.



User Journey

User Journey

The steps and choices users make moving

through an interactive service experience.

Choose Access Medium

Choose Distribution Point

Cross-play

Previously, users on different distribution platforms couldn't play or communicate with one another, and if a user played a game on different types of hardware, they couldn't share progress. This has now become feasible for many games.

This is managed via users being assigned a developer ID, which can prevent conflicting usernames when crossplatform play is active on a game. For example a user with the username Ofcom123 on a console could be assigned a developer ID as Ofcom123#1 and a user on a PC with the same username could be assigned a developer ID as Ofcom123#2 when playing the same game.

Direct Sign Up

Direct relationship with the gaming distributor where the user adds the required data themselves.

Federated Sign On

User chooses from third-parties supported by the game distribution platform, consenting to share details to complete sign up for a service.

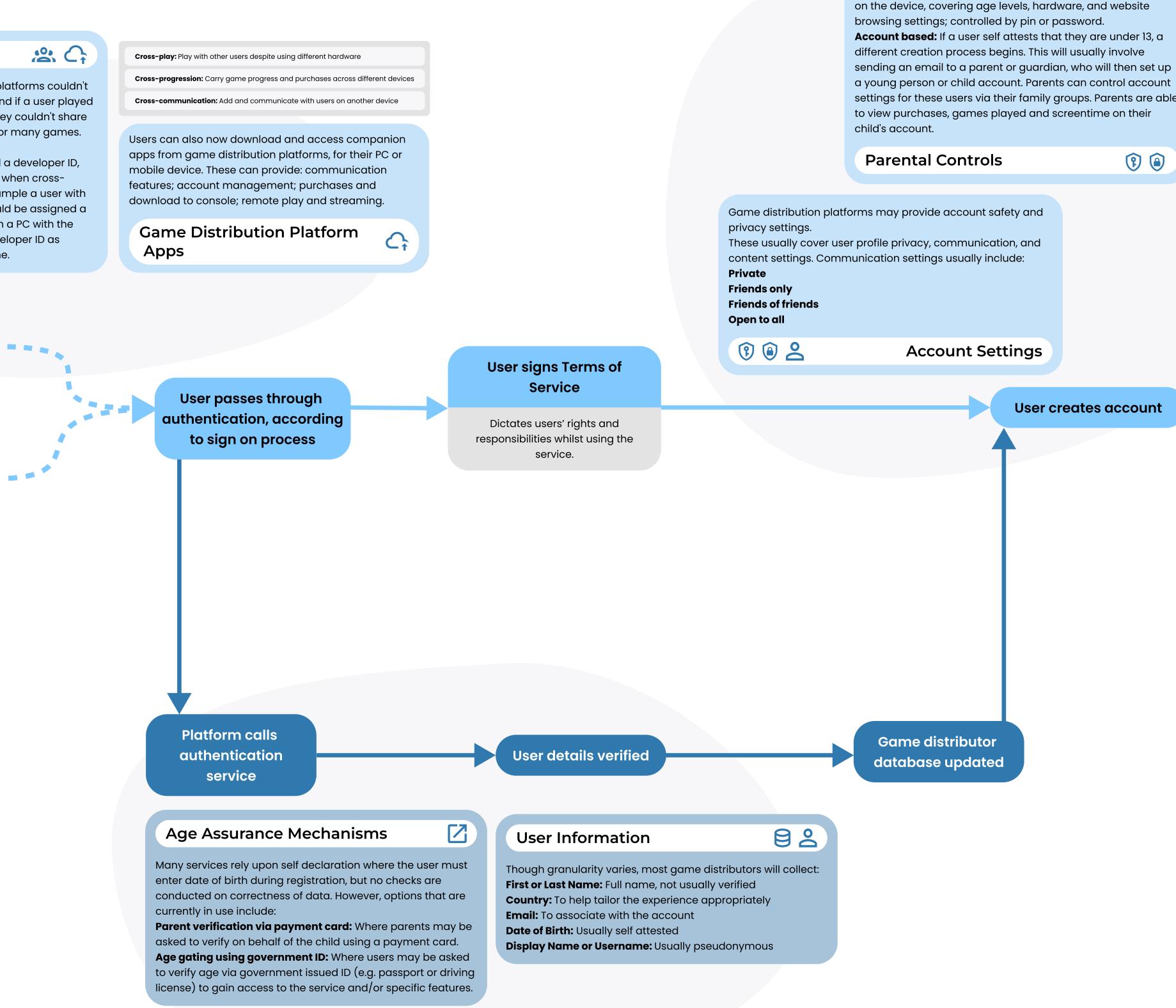
Q Search Server Customisation 🙆 Security Community Profile Comms Feature Monetisation 💆 Game Client Prust and Safety Database Cloud External Search Service Database Automation



Platform Architecture

The platform workflows and backend processes that support the user journey and safety policies.





Find Game

inappropriate behaviour.

Malware and App Store Protection 🤅

users are made aware of. Backend functionality: Such as content moderation mechanisms.

Game Distributor Safety Requirements

User Content Controls V Ó

Platform matches user to games

User searches for game

Finding Games

Jsers find games through a number of avenues: **Staff picks:** A set of games recommended by

Search: By name, genre or other descriptive

Subscriptions: If a user is on a subscription membership, the game distributor may promote games that are part of, or have recently been added to, the subscription catalogue.

Network recommendations: Some game distributors will show users what their friends are playing or, depending on user preferences, may let them view if their friends own a particular

Some game distributors provide out-of-game communication features where different media types can be shared. Their Trust and Safety measures will usually be applied to their communication features, but may not apply to communication features within third party games accessed via their distribution platform.

Direct messages: Between two individual users, often restricted to friends or friend of friend interactions.

Party chat: These can be open or closed, where open means that a user can join, or request to join, a friend's or friend of friend's party, and closed means that users can only discover or join the party if they have been invited to do so. **Communities and clubs:** Online meeting places created by users. These can be visible, accepting requests to join, or

_ _ _ _ _ _ _ _

Q 🙁

private, only accepting invited users.

Communication Features

User opens communication features

User creates account

• • • • • • •

(?)

Parental controls can be applied in one of two ways:

On device: Where settings are applied to all users and guests

Find more information on payments

User sets up wallet

in the 'Monetise' stage

User adds friends

Adding Friends

Users can find friends by:

Name, Username, Gamertag or ID search: Users can find friends by using a search bar; this will not show private profiles. Suggested friends and 'Players Met': Game providers and distributors may cultivate lists of suggested friends for users. This can include 'Players met' from in-game experiences. Users can 'quick add' from this list.

.inking social media accounts: To share updates, users can add friends who play the same games or belong to a common

game distribution platform. Transference from existing networks: Where one network interconnects with another, automatically syncing friends across. For example, a game distribution platform may sync friends from their platform into the individual games a user

Community search: Users can find friends by searching for clubs or communities and adding users from these networks.

Purchase (optional)

Game software can be a target for malware. Distributor platforms will often vet the applications that are listed on their store, and sometimes alongside the game provider will check applications and user devices and their settings for unusual or

Game providers may also provide guidance to users, such as FAQs and blogs about targeted attacks or use of known

App Stores and Game distributor platforms sometimes require game providers to meet safety requirements in order to appear on their distribution platforms. Such safety requirements can, for example, include the obligation to have: Community standards: A set of behavioural standards that

User facing features: Such as reporting, or blocking.

Points of contact: For escalations around serious offences.



Users can set controls for what games they can see in game and app stores, filtering out for factors such as graphic content, profanity, for certain themes, and for age restrictions.

User finds game

User signs up for subscription

Pay to Play Games

Pay to Play Games require a subscription payment on an ongoing basis in order to use a

Online Multiplayer Subscription: Users may be equired to pay a recurring fee to access online multiplayer for pay-to-play titles. In additior users may be able to redeem from a small selection of different games each month. Catalogue Subscription: User pays on a recurrine basis for access to a larger catalogue of games, which can change over time.

Single Game Subscription: User pays on a recurring basis for access to a single online game. Users can sometimes also pay additiona fees to receive the most recent game updates

User makes purchase

Buy to Play Games

Buy to Play Games can be played after a one time purchase. These games may also include expansions and downloadable content that are made available free of charge or for a fee. One off: User pays once for permanent access to

User downloads game (if free)

Free to Play Games

Though many games are accessible for free they may derive value in other ways:

Microtransactions: Where users can spend money on in-game items, content or in exchange for virtual currency to make purchases. **Premium memberships:** Provide monthly benefits such as access, discounts, or perks relating to in-game content, currency, or

Advertising for revenue: Some free to play games, often mobile, rely on rewarded video advertisements, offer walls, and banner ads for revenue.

Recommender Algorithms

Content based filtering: Uses metadata on the game, such as genre, to make basic recommendations.

Collaborative filtering: Uses data collected from a user's transactions and activity to map user interests. Makes recommendations and predicts user preferences based on the preferences of

Hybrid filtering: Offers suggestions to the user by combining content based and collaborative filtering. Generates complex insights about user preferences and game popularity and trends.

Sign On | Game Provider

Game Providers may design, publish, and/or support games.

The Sign on process for game providers often represents an optional secondary layer of account creation beginning from the launch of the game. A user may opt to use a supported federated sign on process, use an existing game distribution platform account, or opt to sign up directly with the game provider. In some instances, signing up with the game provider is required before a user can play the game or access certain in-game features. The method in which the game is accessed may have implications on the user registration process.

User Journey

The steps and choices users make moving through an interactive service experience.



User installs game and launches client

Platform Architecture

The platform workflows and backend processes that support the user journey and safety policies.

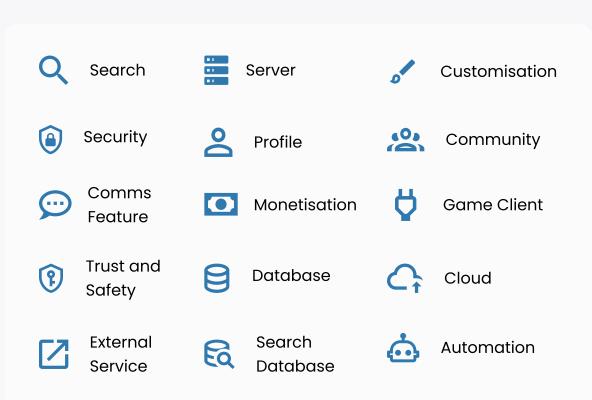


Game client installe and launched on user device

Game Client

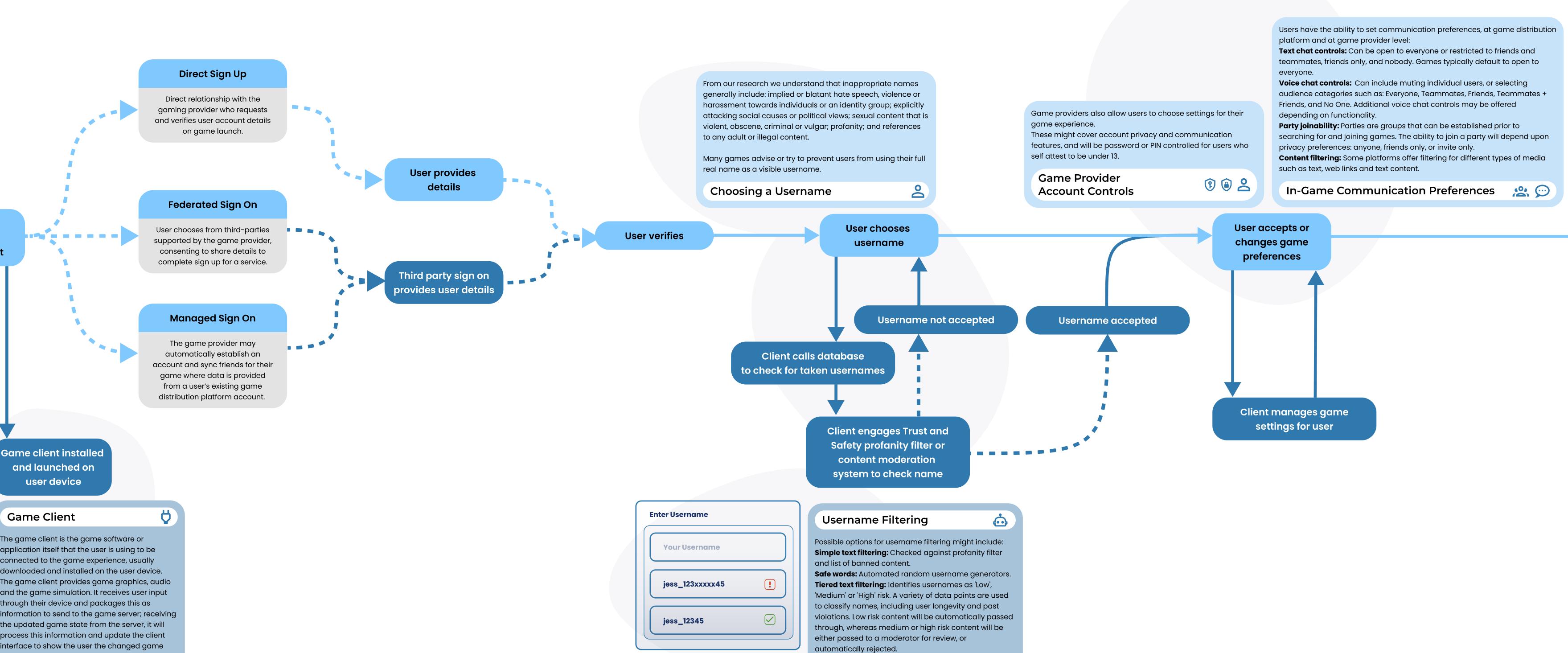
ne game client is the game software or application itself that the user is using to be connected to the game experience, usually lownloaded and installed on the user device. hrough their device and packages this as

The game client will also process any changes to layout or settings made by the client.

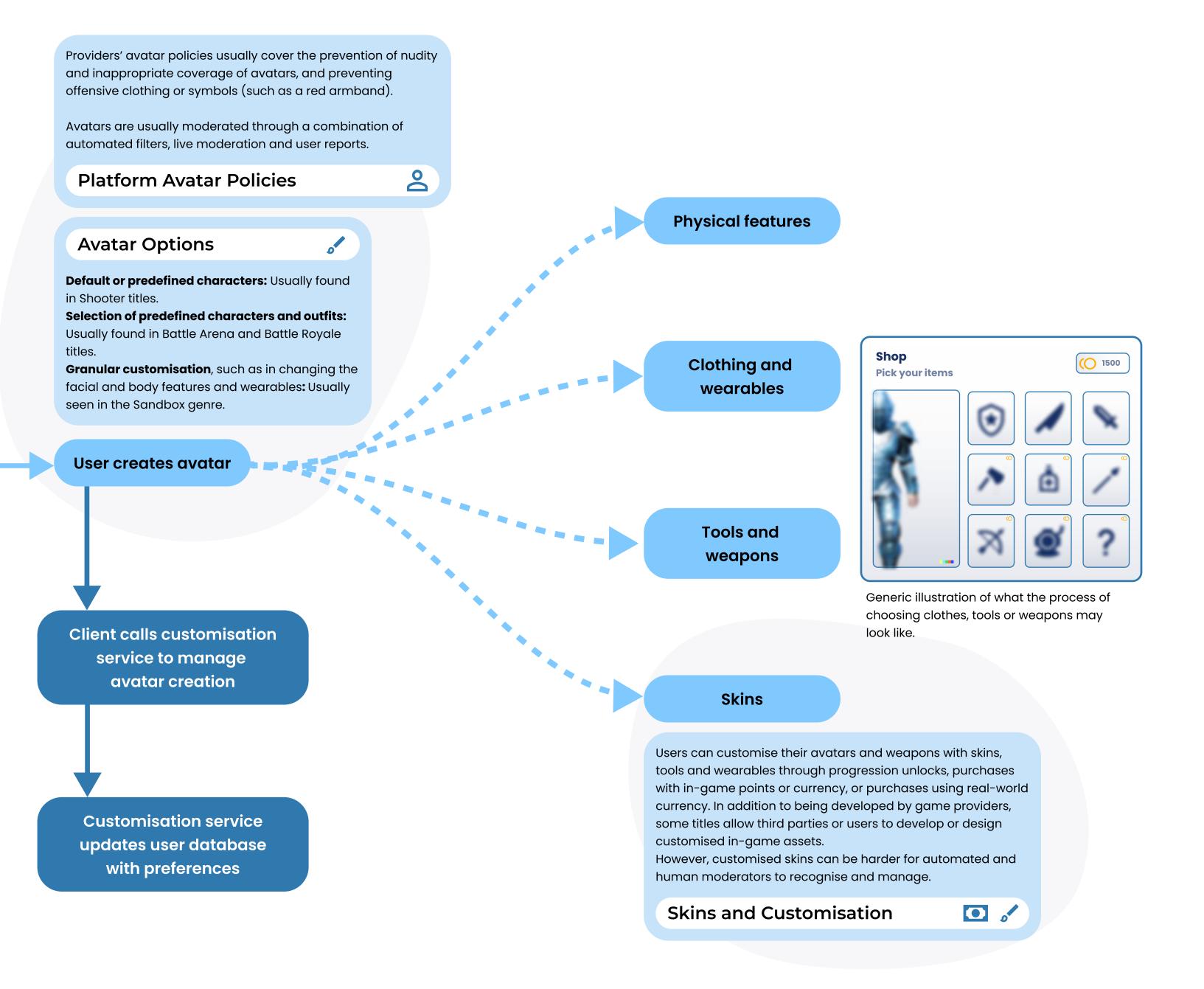








More complex systems will also convert username text into plain text, removing punctuation, Unicode, emojis and other unnatural language characters.



Participate | Set Up

The Set Up process covers actions taken just before the launch of a game session. These include lobbies, choosing teams and matchmaking.

User Journey

Journey Phase

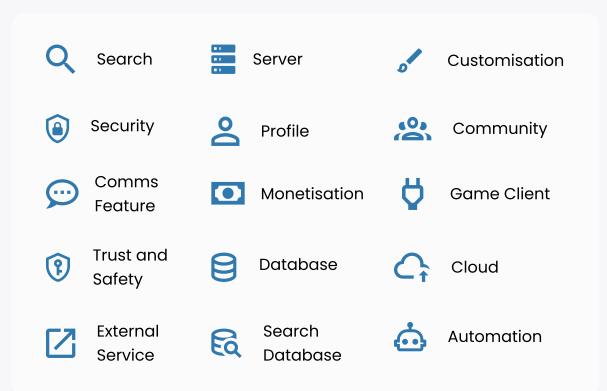
The steps and choices users make moving through an interactive service experience.



Client sends reques to game server

Game Server

per second is known as 'Tickrate'.



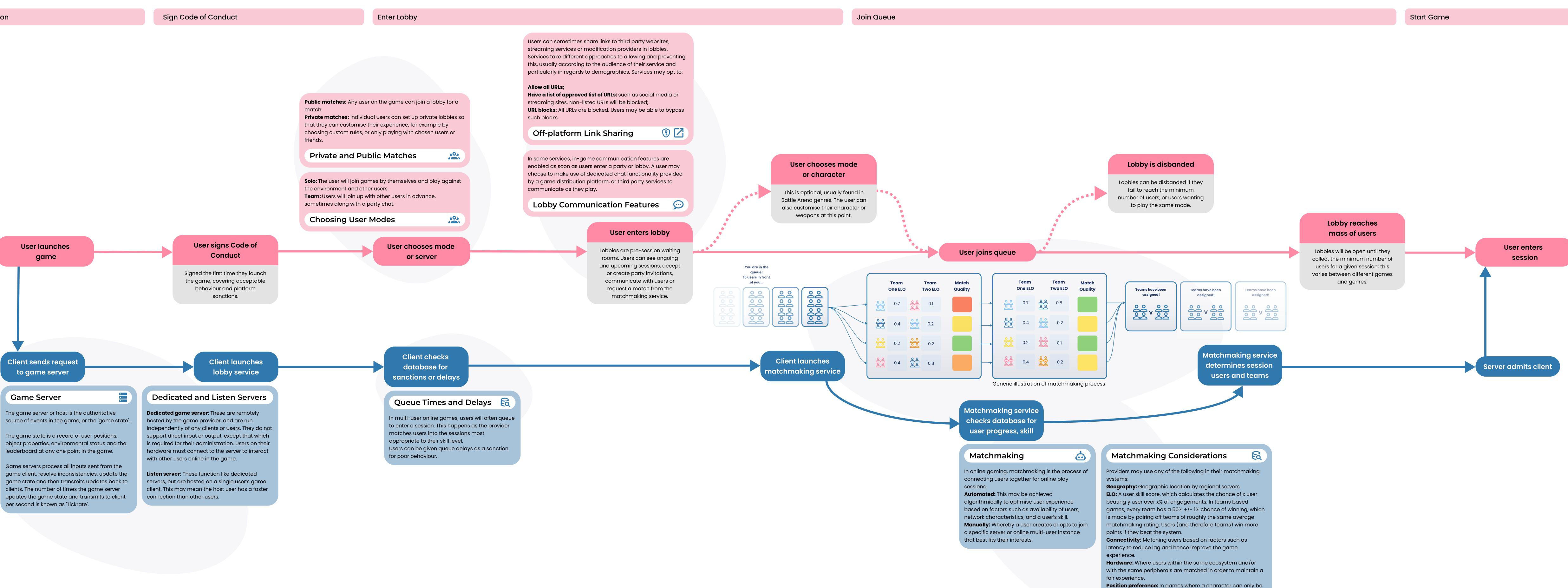




Platform Architecture

The platform workflows and backend processes that support the user journey and safety policies.

Platform Architecture



The game server or host is the authoritative

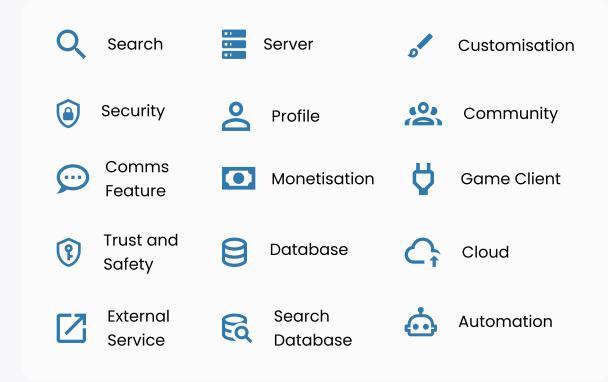
The game state is a record of user positions, object properties, environmental status and the leaderboard at any one point in the game.

Game servers process all inputs sent from the game client, resolve inconsistencies, update the game state and then transmits updates back to clients. The number of times the game server updates the game state and transmits to client

- Position preference: In games where a character can only be selected by one user (mainly team based, such as Battle Arena), lobbies will be made up of a sufficient diversity of users so that users have a reasonable chance of attaining the user thev want.
- Queue time: How long a user has been waiting in the queue, ninimising wait times.
- Group size: Making groups up to different sizes in order to displace any advantage gained by being in a premade team.

Participate | Compete

Compete covers competitive user versus user games. This stage describes the basic actions of a game session, and the process a user might follow if they experience inappropriate behaviour.

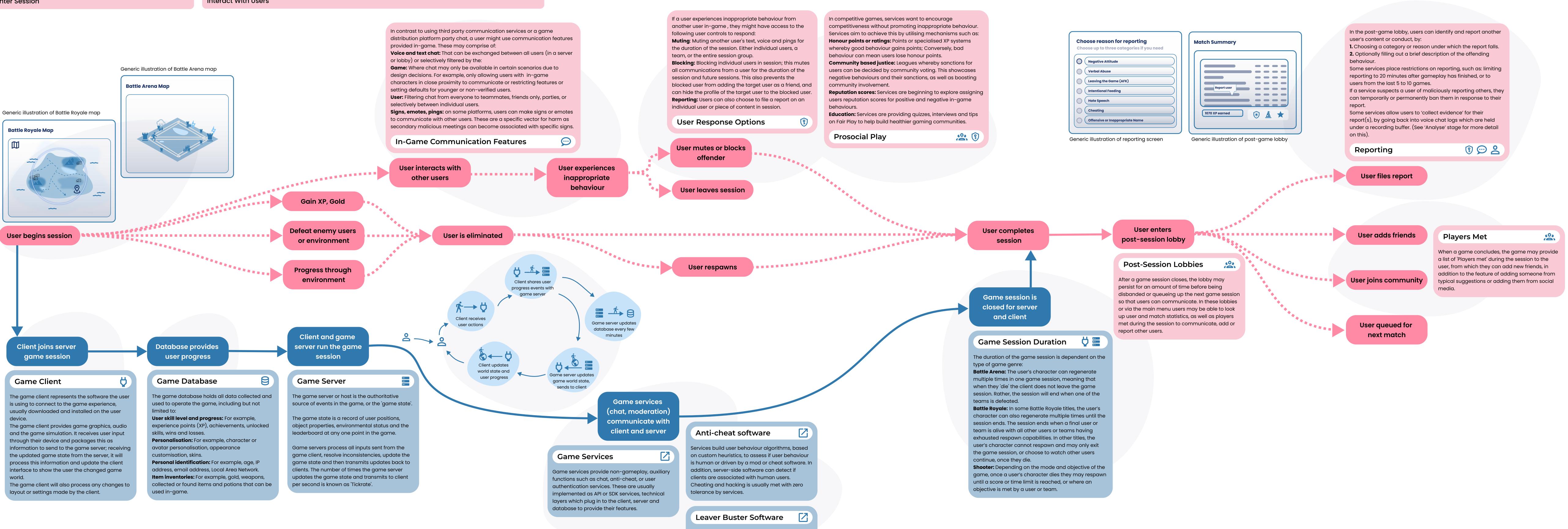




Journey Phase

Open Game Session

Enter Session



The steps and choices users make moving through an interactive service experience.

User Journey

User Journey

Platform Architecture The platform workflows and backend

processes that support the user journey and safety policies.

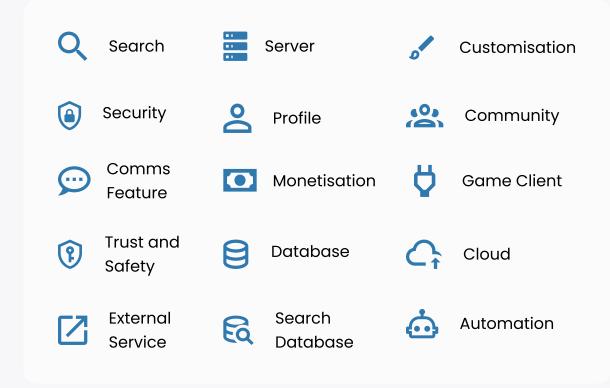


Interact With Users

Services may use automated systems to courage users from repeatedly leaving matches or idling during gameplay. Users might be warned, placed in a low priority queue or have their XP increases from the session prevented. lowever, these systems may have implications or users leaving games for other reasons such s experiencing inappropriate behaviour.

Participate | Build

Build covers Sandbox genre games, whereby the gameplay is much less structured, prescriptive, and linear. This stage covers some of the basic activities a user can explore, as well as the process they might follow if they experience inappropriate behaviour.





Journey Phase

Open Game Session

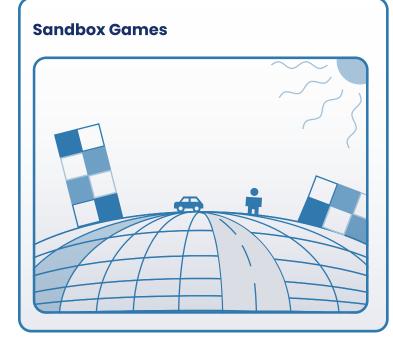
Enter Session

User Journey

The steps and choices users make moving through an interactive service experience.



Generic illustration of Sandbox game screen



User enters server

Platform Architecture

The platform workflows and backend processes that support the user journey and safety policies.

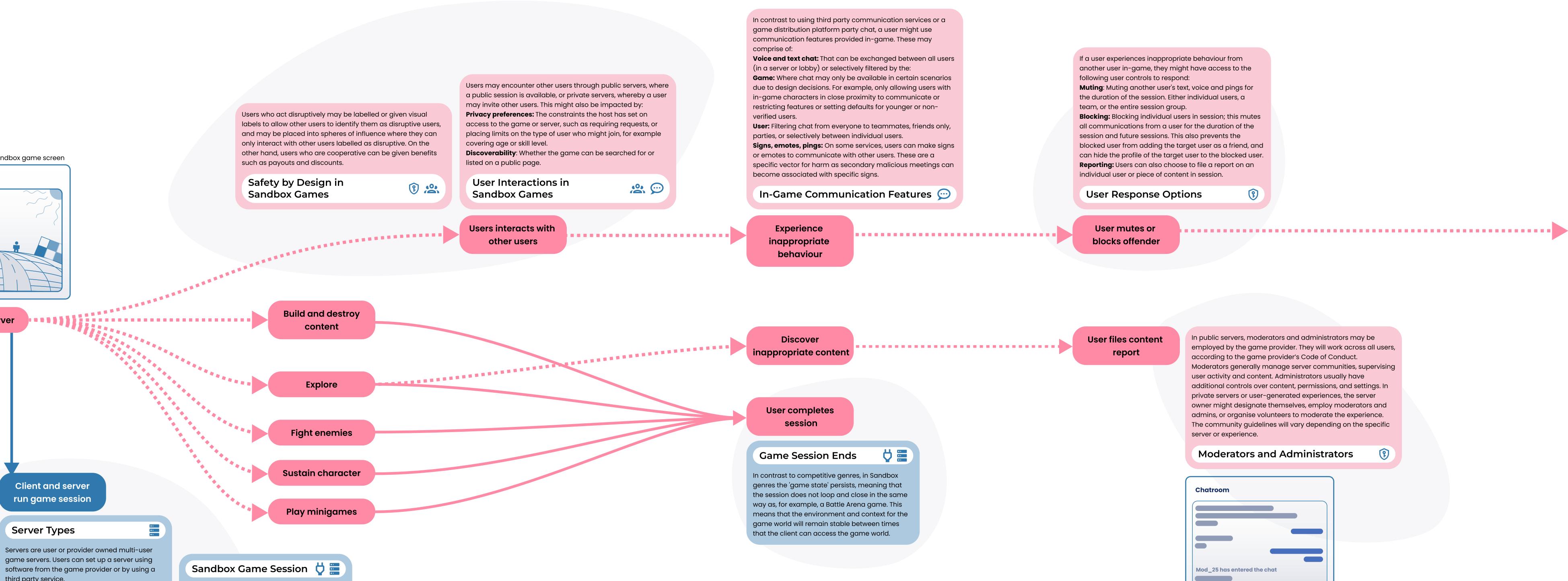
Platform Architecture

Client and server run game session

Server Types

third party service. Public servers: Provider hosted servers open to all

Private servers: Owned and managed by groups, organisations or individuals not affiliated with the game provider. Private server owners control the settings and users allowed in their servers, and can even charge for users to access their server. hese servers can have either particularly strong or weak Trust & Safety measures, depending on the approach of the server owner.



Where in competitive genres, the session is defined by a match or discrete action loop which can last for varying amounts of time, in Sandbox genres the session length is not linked to any discrete loop. Rather the user may join or be assigned to a game session via manual server selection, matchmaking or invitation. The session then persists for as long as the game session is populated by the minimal required number of users

Experience inappropriate behaviour

Generic illustration of chatroom screen

User files user report

Enable | Create

Journey Phase

Create covers functionality wherein interactive services provide users with the ability to create their own experiences, which other users can then take part in.

User Journey

The steps and choices users make moving through an interactive service experience.



Users can usually start to build experiences without any restrictions or agreements. Game providers may set limits on which users can then publish or monetise their experience. These limits may cover factors such as: Age of user: Services may choose to prevent users under a certain age from publishing or monetising their experience. an offensive experience, they may be prevented from publishing experiences in future.

Creator Limits

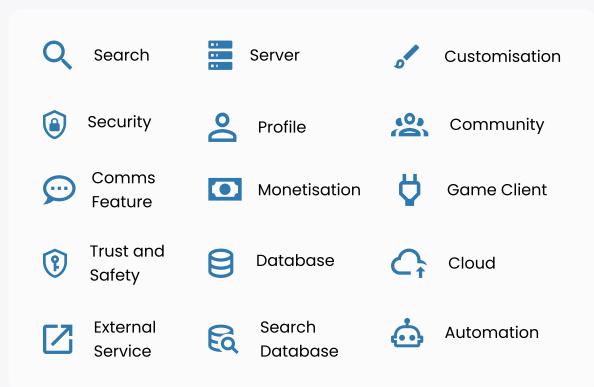
Game providers may also facilitate users creating game experiences for other users. This might happen in one of two

Enabling private servers: On some Sandbox games, if a user creates a private server, they can alter the game settings and environment. They can then choose which users can access this experience.

Providing experience creation studios or modes: Some Sandbox games provide experience creation studios or creative modes where users can create games or experiences These experiences may have the potential to be made public, and even monetised.

Sandbox Experience Creation

User opens studio or creative mode

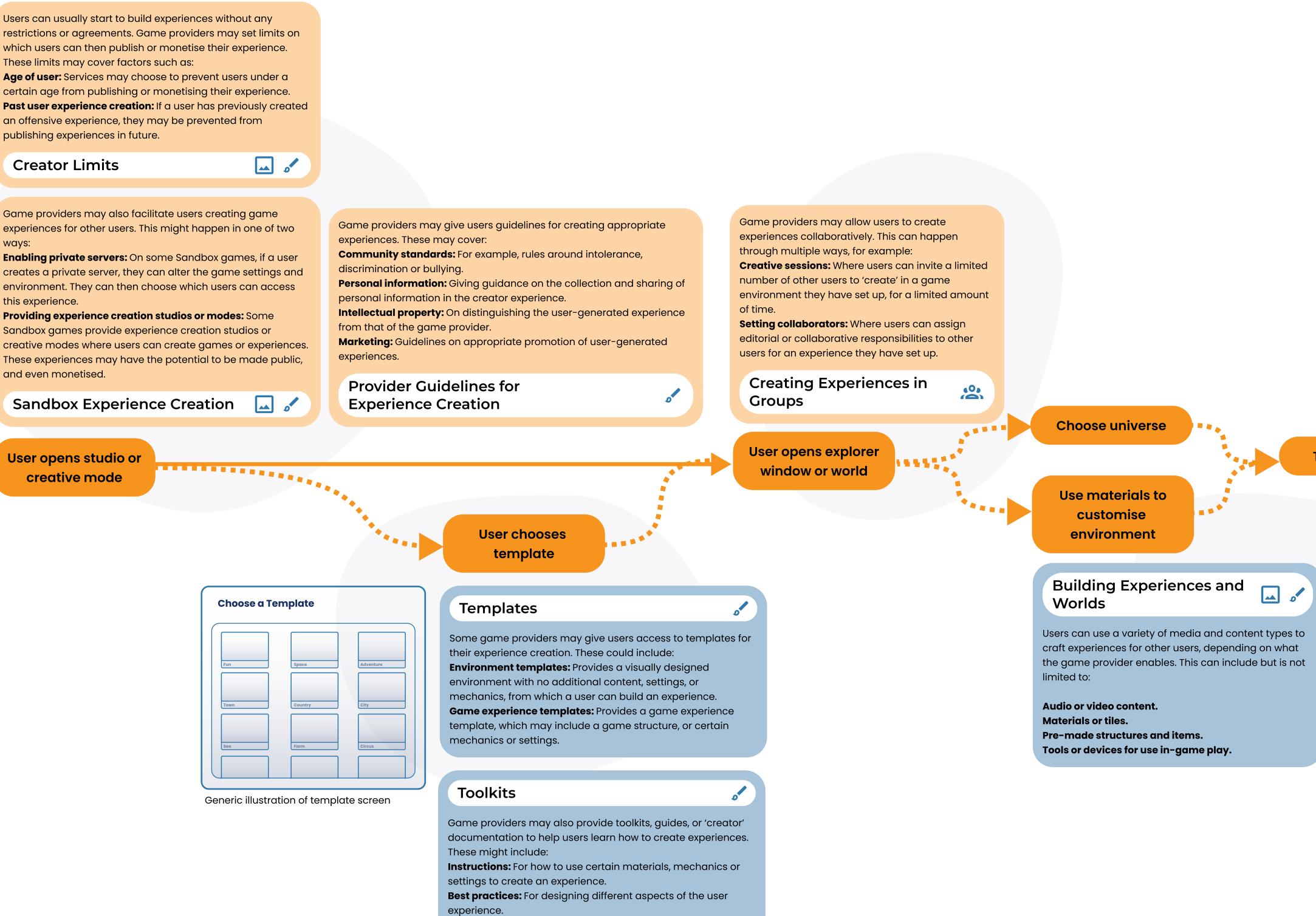


OFCOM PUBLIC



processes that support the user journey and safety policies.





Full tutorials: Taking users end to end through the designing, publication and monetisation process.

User information: For example, on the game's user

emographics.

Generic illustration of age settings Choose age settings 13+ (Suitable for ages 13 and older) 9+ (Suitable for ages 9 and older) All Ages (Suitable for everyone)

> Some game providers may ask users who create experiences to set age restrictions on the content they create, if they would like to publish it for other users to experience. These might consist of: All ages

Adult users only (18+)

Different tiers of child users: Such as 13+ only or 9+ only.

Any age restrictions would usually be set at the point of publication.

Age Restrictions on **Published Content**

User chooses settings

Settings in User-generated Experiences

When a user creates an experience, they can set specific information and settings which affect various aspects of the game. This can include: **Basic:** Name, description, game icon, screenshots and videos, genre, playable

Permissions: Public or private, game owner, collaborators (for play and/or editing). Permissions can also include what permissions different levels of users

Difficulty: Different levels of difficulty for a specific game format.

Non-player characters (NPC): Whether these features are enabled in games where they are relevant.

Mode: Enabling certain modes, such as competitive user vs user combat. Specific feature settings: Whether users can use certain features in-game. Monetisation: Badges, paid access, private servers, developer products.

Places: Create, configure place and version history.

User idle timeout: How many minutes a user may stand idle before being

Maximum tickrate: The maximum number of milliseconds a single tick may take before the server is stopped due to being overloaded.

Localisation: Source language, automatic text capture, use translated content, automatic translation.

Avatar: Presets, avatar type, animation, body parts, clothing.

World: How the user's character moves in the world, such as gravity, jump, walk,

Different providers may ask users to go through different processes in order to either publish or monetise their created experience. These may

Joining a specific developer or creator community or group: These might have specific entry

requirements such as age, and a certain volume of social media presence.

Validating tax status, when monetising an experience

Signing a new set of Terms and Conditions. Purchasing a certain amount of in-game currency or signing up to a premium account.

Publishing a game experience

User changes permissions to public

Test the game

craft experiences for other users, depending on what

Different providers have varying approaches to sharing user-generated experiences with other users. These may include but are not limited to: **Experience codes:** Assigning or asking users to generate a code for their experience, which other users can use to search for that experience. **Discover pages:** Curating user-generated experiences by factors such as popularity, trends, ratings, experience genre, or sponsorships.

Finding User-Generated Experiences

Another user discovers game experience

Moderation of Experiences

If a sufficient number of users report an experience it might be placed 'under review', meaning a team of moderators will review the content and game experience. Moderators can then take action to suspend or delete the game experience, or the creator is given a warning and explanation of how to remedy the issue in the experience.

Automated means of moderating user-generated game experiences may be emerging and we will look to explore these further as they develop.

User reports game experience

Enable | Monetise

The Monetise stage represents the possibilities for service monetisation in gaming. This covers users setting up payment details and virtual wallets, and participating in in-game and outof-game economies, as well as the associated potential inappropriate behaviours that may arise from these activities.



Journey Phase

The steps and choices users make moving through an interactive service experience.



Provide direct payment details

Integrate gaming distributor wallet

Integrate gaming stributor purchases

Integrate mobile payment structure

Q Search Server Customisation Security Profile Lommunity Comms Feature Monetisation 🔂 Game Client Trust and Safety Safety Database Cloud External Search Service Database Automation





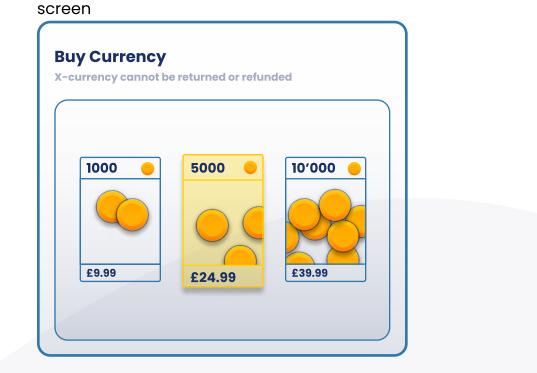


The platform workflows and backend processes that support the user journey and safety policies.



Provide Payment Details

Generic illustration of currency purchase



Game distribution platform wallets: Users can make purchases in game distribution platform stores (such as purchasing virtual currency and items) that are redeemable in-game. In many multi-platform titles these purchases can be synced across devices.

Linked wallets: For some games, the game provider and user can choose to allow game distribution platform account and payment details to make in-game purchases.

In-game wallets: Users can set up wallets in individual games, by providing payment details in the game.

Mobile payment: Users can make in-game purchases by integrating existing payment mechanisms from their mobile. Gift Cards: Users can also use gift cards of set balances in order to make in-game purchases.

Users and services can transact with each other for in-game items and virtual currencies in a variety of ways, for examples User-to-service transactions: Such as premium membership fees and in-game purchases.

User-to-user transactions facilitated by the service: Through in-game trading hubs, or by providing a sign on mechanism in partnership with a third party site so that users can trade items securely.

User-to-user transactions not facilitated by the service: These can take place on a variety of third party sites.

Transaction Types

In-Game Purchases

User sets up wallet

Secondary Economies

Some game commodities, which can be unlocked through gameplay or purchased by users, have gained value that is additional to or separate from in-game functionality. For example, rare skins in some games have built high social capital; meaning they may become worth high volumes of in-game and real-world currency. Some games facilitate the trade of these commodities through in-game trading and purchasing hubs. There are also a broad variety of third party buying, selling and exchanging sites that facilitate commodity transactions outside of the game.

Exchange Mechanisms

Ο

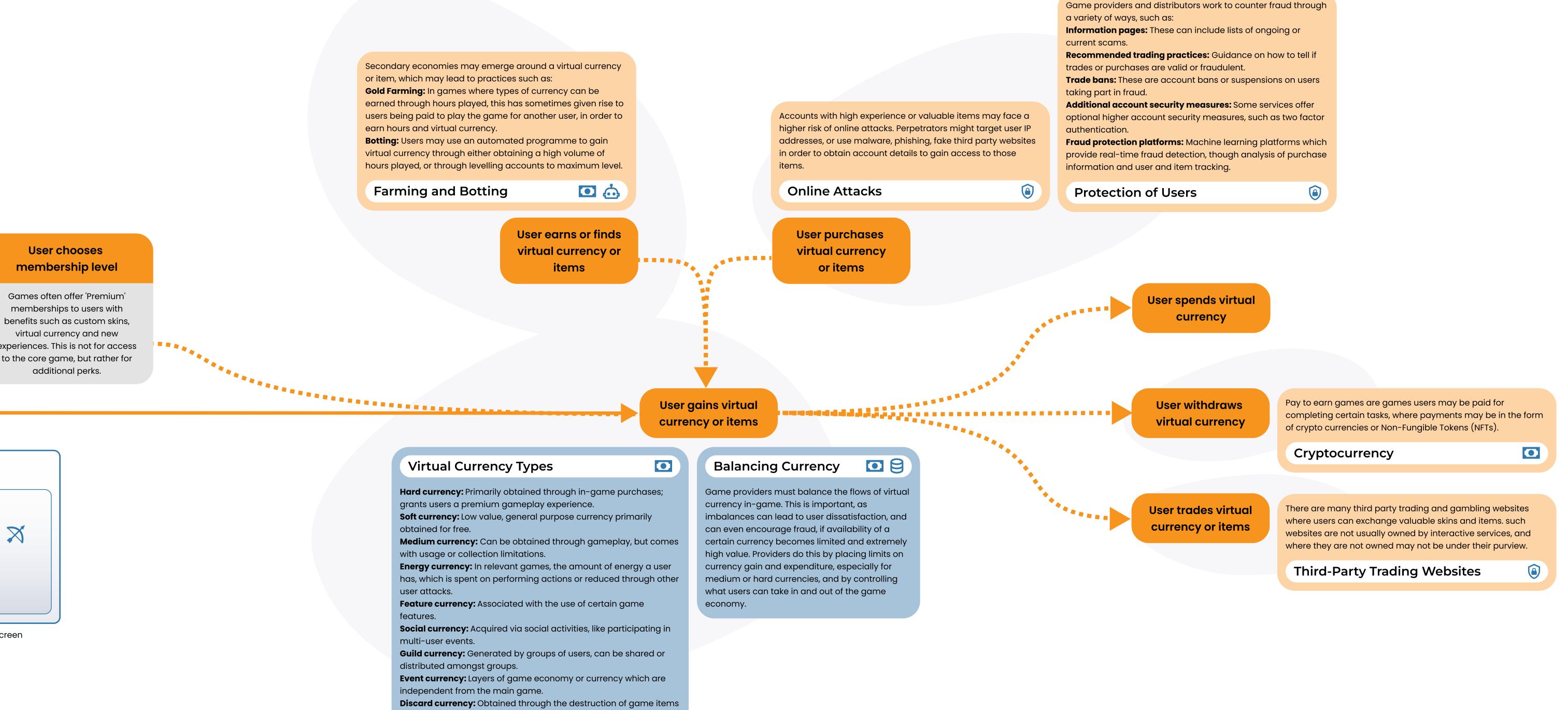
In-game: Some games provide 'Loot Boxes', which can be purchased through gameplay or in exchange for virtual or real currency, and can contain a set of in-game items of unknown quality and quantity.

Out-of-game: Third party sites provide exchanges using skins or items as currency, where users bet on games, tournament or stream outcomes. Users might link their game account or provide their login details to the site, giving the site access to their item inventory, which they then lose control of for the duration of the bet. If they lose the bet, the item(s) are automatically transferred to the winning user(s).

1 Loot box opened! 00 + × × ×

 \mathbf{O}

Generic illustration of Loot Box screen



Informal currency: Elements that users use as a medium of exchange, even though the intended purpose of the game provider for that element was different.

VIP currency: Generated as a byproduct of performing an in-game

Moderate | Analyse

The Analyse stage details the measures put in place by interactive services to collect and analyse content and user behaviour to determine whether it is inappropriate. This includes deciding Trust and Safety measures, collecting data, classification and prioritisation of content and conduct, and improvement of systems.

User Journey

The steps and choices users make moving through an interactive service experience.

User Journey

sanctions.

action is in violation.

Community Standards

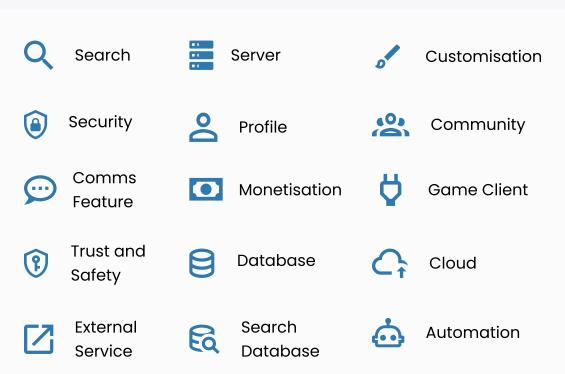
such as voice chat and avatar interactions. Sandbox environments.

Types of Interactions

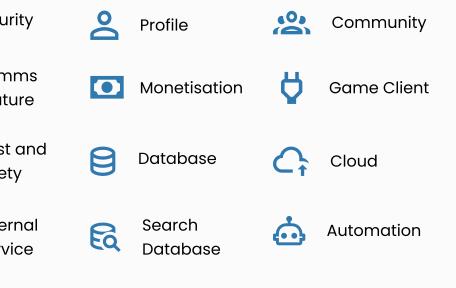
User produces content asynchronously

Service sets riorities for content moderation

formed by: violence databases might be relevant.









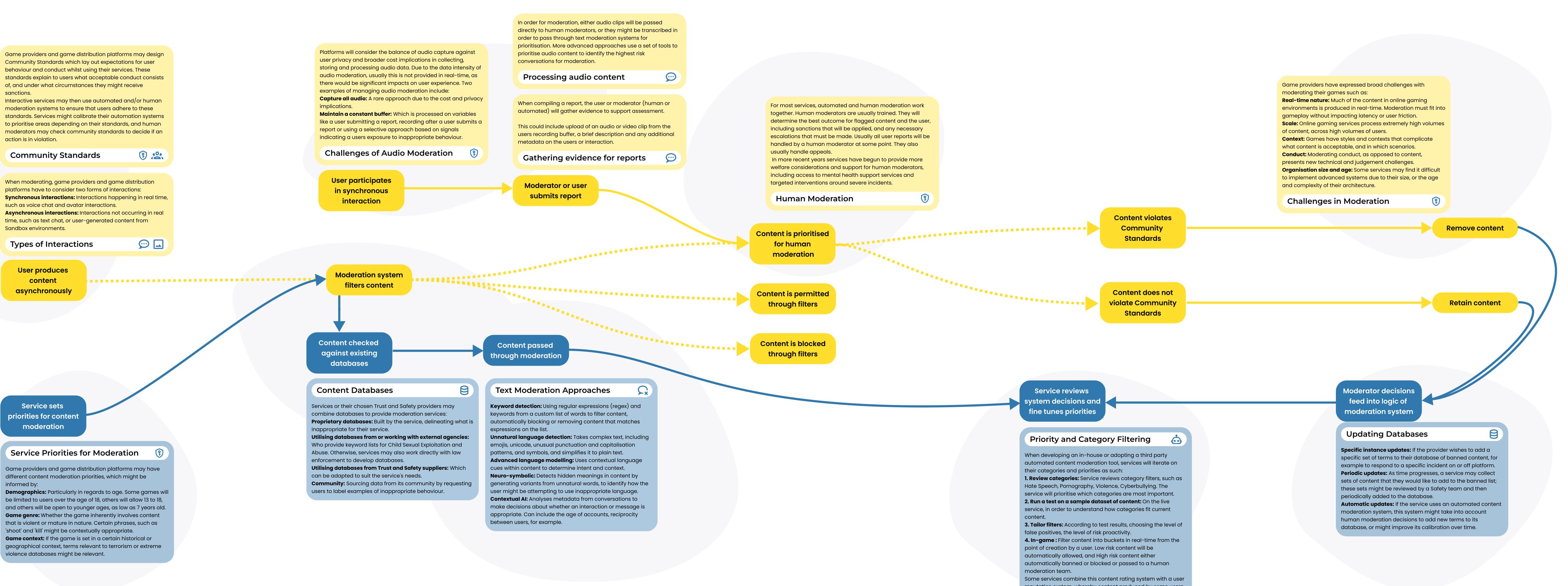
processes that support the user journey and safety policies.

Platform Architecture

Journey Phase

Decide Trust and Safety Approach





reputation system, whereby content produced by some users s automatically sent to moderators, until they become trusted

Moderate | Respond

Journey Phase

The Respond stage provides details on how interactive services respond to the finding of inappropriate content or user behaviour. These may include the process of removing content, sanctioning users and alerting third-parties, such as law enforcement, where necessary and appropriate.

User Journey

The steps and choices users make moving through an interactive service experience.



Moderator witnesses conduct

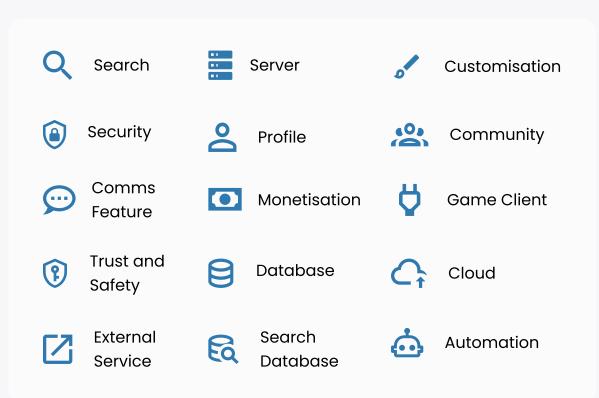
Moderator removes content

Platform Architecture

The platform workflows and backend processes that support the user journey and safety policies.

Platform Architecture

against them.





Alert Law Enforcement

Increasingly gaming services are making partnerships with suicide prevention charities. If suicidal or self harm content is detected, then users might be sent links to one of these support services.

Escalation for Suicide and Self Harm 🔇

Approaches to counter-terrorism vary across the industry. Some companies partner with organisations to co-ordinate response to extremist content, by operating incident response strategies whereby they share resources that services can plug in to existing content moderation services to continually update their counter-terrorism approach.

Escalation for Extremist Content

Content that is identified as Child Sexual Exploitation and Abuse (CSEA) and Material (CSAM) may be immediately removed, either by automated or human moderators. The service will then compile a report containing any inappropriate content, and any metadata about the user they have. Services may have escalatory routes for CSEA and CSAM, where this report may be shared with appropriate parties.

Escalation for Child Abuse Material

Content or conduct shared with external authorities

User is sent a warning

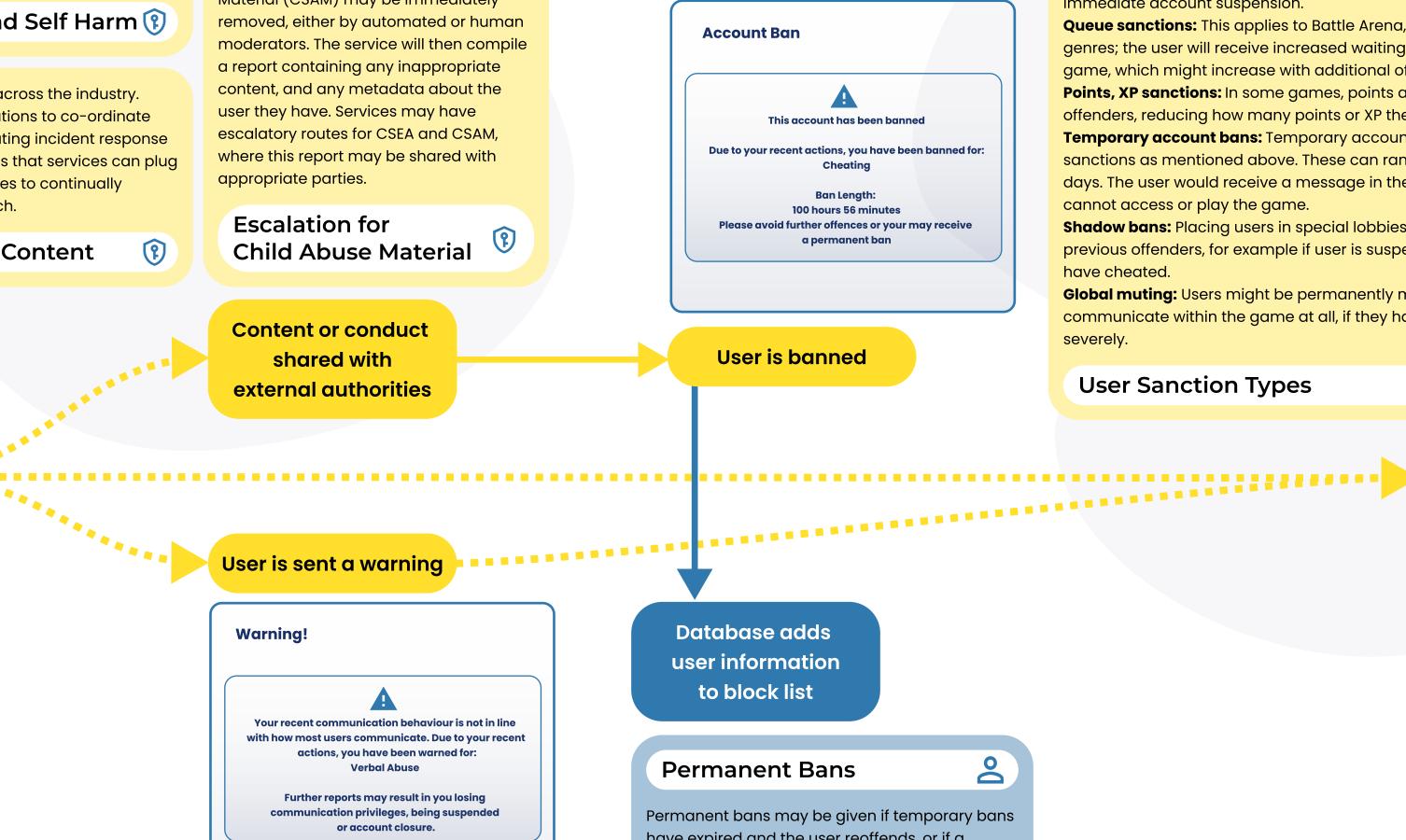
Warning!

Your recent communication behaviour is not in line with how most users communicate. Due to your recent actions, you have been warned for: Verbal Abuse

Further reports may result in you losing communication privileges, being suspended or account closure.

Generic illustration of user warning screen

Generic illustration of account ban screen



ave expired and the user reoffends, or if a serious enough offence has been committed, such as an illegal offence, or cheating. This might ake one of four forms:

ermanent IP bans: Targeting the IP address of the access point of the banned user. Permanent device bans: Targeting an identifier of the access point of the banned user. Console bans: Targeting the console of the user, meaning they cannot access the network on any account using this console.

Account bans: By game providers and game distribution platforms, preventing the use of online features.

User De-Identification

e-identification: Games will usually track user offences over time. However in order to ensure anonymity, users are associated with identification numbers which change on a routine basis. If enough offences happen over time, the user might have sanctions placed

User Reputation Systems

Database provides

user information

Moderator checks

for previous user

offences

An increasing trend in Trust and Safety systems in online gaming is incorporating or prioritising user behaviour into moderation efforts, as a majority of inappropriate content usually originates from a small percentage of

ongevity of account: Restricting user permissions based on how long an account as been held for.

Verification: Restricting user permissions based on whether the user has age or identity verified to the service.

Prior behaviour: If a user has been blocked or reported, or has sent a message which has been blocked, this can contribute to poor personal 'Trust' scores. These systems are also called 'reputation' ratings, and may be shared with users to encourage positive behaviour.

Frequency: Whether the offence is the first infraction from the user, or whether it is a repeat offence. Typically services will apply escalating sanctions. Severity: This might include how many users were affected, and what type of users were affected. This could also include whether the offence has legal implications, such as coming under priority categories such as CSEA, Terrorism and Suicide and Self-Harm.

User Sanctions Factors

Warning message: This would be received before any sanction was imposed, usually on first offence, unless the offence was severe enough to warrant immediate account suspension.

Queue sanctions: This applies to Battle Arena, Battle Royale and Shooter genres; the user will receive increased waiting time in the queue before the game, which might increase with additional offences.

Points, XP sanctions: In some games, points and/or XP sanctions are applied to offenders, reducing how many points or XP they can earn from the session. **Temporary account bans:** Temporary account bans follow less severe sanctions as mentioned above. These can range from 24 hours, to 72 hours, to 7 days. The user would receive a message in the client notifying them they cannot access or play the game.

Shadow bans: Placing users in special lobbies or games with only other previous offenders, for example if user is suspected or has been confirmed to have cheated.

Global muting: Users might be permanently muted, meaning they cannot communicate within the game at all, if they have offended frequently or severely.

User Sanction Types

User is given a

sanction

Platform Sanction Workflows

Services set up workflows around user offences, recidivism, and sanctions. This usually would take the form of a matrix, linking offences, instances (i.e. primary offence, secondary) to actions in the platform workflow around sanctions. These can be both positive and negative. There are different routes for managing these

Human moderation: Some services will only allocate decisions on sanctions to human moderators, who can review a case with all relevant metadata and context. Human and automated: Some services will use a combined approach, where an automated system will hand out warnings and minor sanctions, such as queue delays, and human moderators will hand out severe sanctions like permanent bans.

Some services will outsource response workflows in line with content moderation systems.

Notifying Users

Services may have a variety of approaches to notifying users of their behaviour such as:

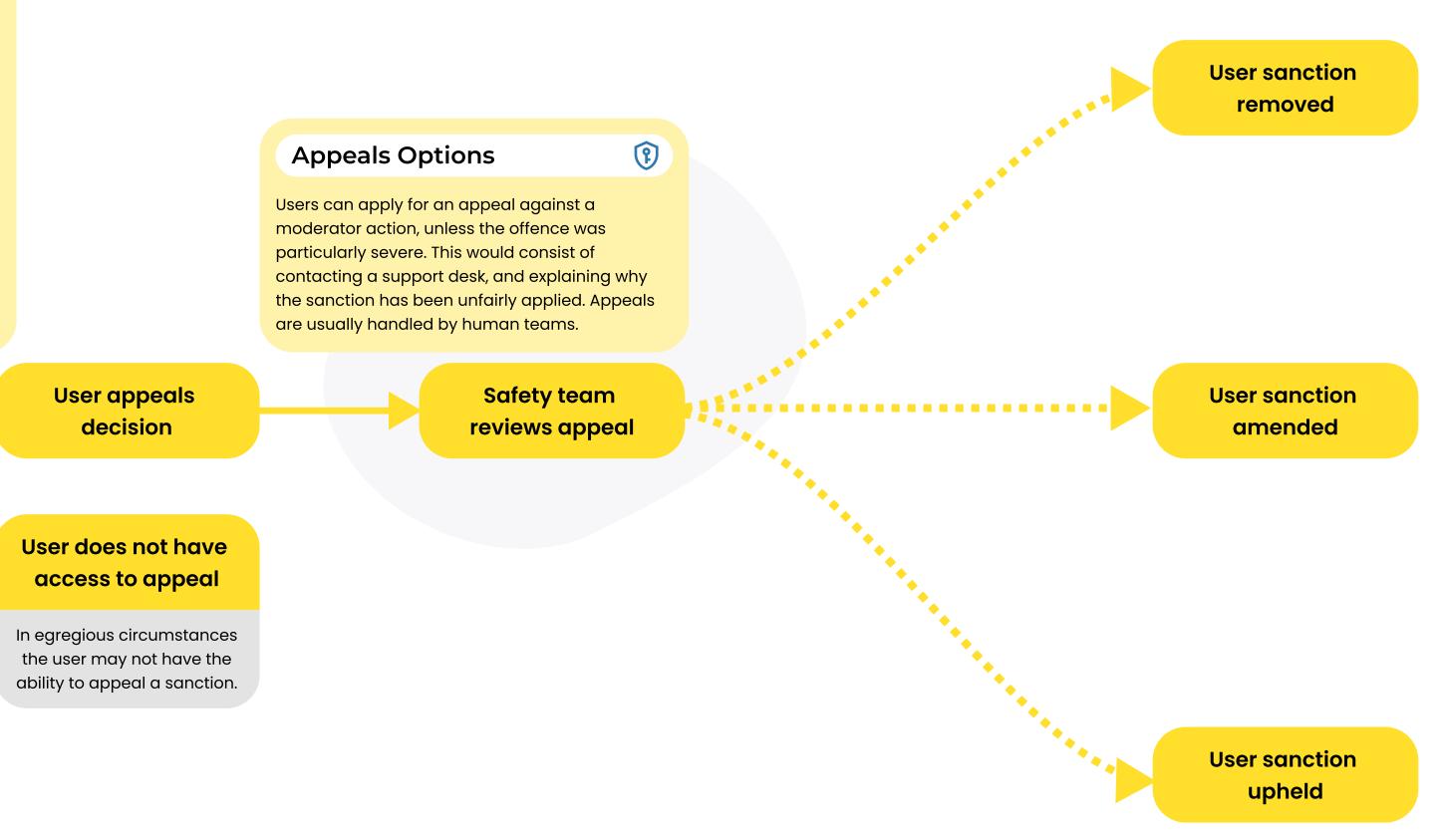
Notification: If a user is offending for the first time, or the offence is not severe, they would usually be notified of a sanction, such as a queue ban. **Explanation:** Some services also offer explanations as to why a user might be receiving a sanction, or sharing with them their offence.

No notification: Services may opt to not notify users because the service does not have a notification system, or the user offence is so severe that services don't want the user to know they are aware of or have recognised the behaviour because they have reported the activity to third parties for further action.

Notification by the service might be affected by the scale of the service, what it might consider to be offending behaviour, and the design of its case management system.

User is notified of sanction





Moderate | Comply

The Comply stage sets out the processes that interactive services undertake to internally audit and report on Trust and Safety. These may include designing Trust and Safety policies, setting priority objectives, monitoring data, escalation and transparency reporting.

User Journey

Journey Phase

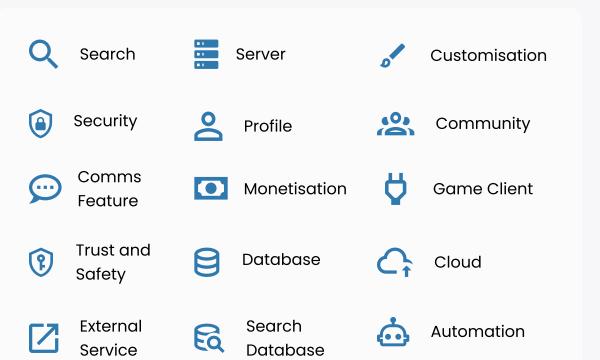
The steps and choices users make moving through an interactive service experience.

User Journey

The platform workflows and backend processes that support the user journey and safety policies.

Platform Architecture Set objectives for Trust and Safety

Platform Architecture







8

Services may choose to partner with external organisations to outsource parts of or all of their Trust and Safety processes, depending

Modernity of service architecture: Some providers operate using legacy systems which are unable to pair with more modern services provided via newer techical layers such as SDK or API. They must then either develop these in-house or work with a third party (such as Safety Tech companies) to create bespoke Trust and Safety features.

Size of service: Where a service has grown quickly, it may not be able to manage the scale of offensive behaviour that may be encountered on it. In such cases, Safety Tech companies might be able to provide the necessary capability more quickly and efficiently.

Types of media on the service: Different Safety Tech providers will have capabilities for the moderation of different media types. This means that services might use one, or a combination of third parties, in order to meet their moderation needs comprehensively.

Relationship with Third Party Trust and Safety

Service designs **Trust and Safety** Policies

Volume of reports: Depending on their Trust and Safety approach and priorities, services will monitor the volume of reports received. Services might also categorise these by type of report. These will also be monitored in specific time periods.

Actions taken against offences: Services might also retain a record of actions taken against offences. This might be categorised into automatic content takedowns and human moderation decisions. When involving user sanctions, services might also record the type of user sanction given out.

Volume of appeals: Services might also record the volume and outcome of appeals processes on their service. This is important for assessing the accuracy and efficacy of Trust and Safety policies. **Escalations:** Services will usually record the number of escalations to external agencies that they take on content or users in a given time period.

Trust and Safety Data Feedback

Service collects data

User-centric data feedback

Services may use user feedback to iterate on their Trust and Safety processes. Some examples might include:

User abandonment: If a user chooses to not continue with their purchase, installation or launch of a game.

User churn: Users leaving a game within a certain period of time. Services might use leaver surveys in order to attain detailed information on user churn.

User satisfaction: Services might ask users for their satisfaction with the game experience at points throughout gameplay, such as in the post game lobby. If a user expresses dissatisfaction, the service might prompt the user to provide further detail on their complaint.

Services will particularly want to understand if certain age verification, assurance or identity solutions create too much user friction, such as credit card detail requests or the use of facial scanning technologies.

Formal reports: Published transparency reports by services focus on a number of key metrics, such as volume of offences and actions taken against such offences. These are usually produced annually, and comparisons are made between different time period such as months, quarters and years.

Informal reporting: Some providers post informal transparency content, such as blogs. Blogs might focus on a specific development or Trust and Safety intervention. These are focused towards users, and explain how new services work, how they impact the user experience, what consequences this will have for user privacy, and any results that have been measured at the point of publication.

Transparency Reporting

Service publishes reports

Services may use data insights on their automated systems to monitor their operation, and to iterate on their Trust and Safety processes. Data points might include:

Accuracy: This can include false positive and false negative reports, such as when the system falsely identifies something as offensive, or misses offensive content. This will consist of a set of metrics describing how often these events occur, and across what types of content.

Speed: This will consist of a set of metrics which might cover: the length of time from an offence happening to a flag by an automated service; the total length of time from incident to a response to the content; or the impact on the speed of messages sent by the use of an automated moderation system.

The service might link these metrics to non-functional service requirements to ensure that the service retains an optimal experience whilst implementing new services.

Automated System Reporting

Service monitors data

Service compiles reports

 \odot

Service uses report to iterate on Trust and Safety policies

Updating Policies

Services may update their Trust and Safety policies and processes over time, potentially considering aspects such as: Taxonomies: Services might update taxonomies of offences or behaviours to incorporate new risks or emerging topics. **Community guidelines:** Services might update or produce new community guidelines to demonstrate and explain responses to new risks or community feedback. Sanctions policy: Services might introduce new sanctions in order to accommodate changing behaviours or patterns on the service.